
Financial Markets and Investments

Claus Munk

cm.fi@cbs.dk
<https://sites.google.com/view/clausmunk/home>

this version: July 2, 2024

Contents

Preface	v
Changes relative to previous version (August 11, 2023)	vii
1 Markets	1
1.1 Market structures	1
1.2 Stocks	2
1.3 Bonds and other debt-related securities	6
1.4 Derivative securities	9
1.5 Alternative asset classes	10
1.6 Main players	13
1.7 Functions	15
1.8 An investment primer	17
1.9 Exercises	20
2 Returns	21
2.1 Returns over a given period of time	21
2.2 The compounding of returns	26
2.3 Annualizing returns	29
2.4 Internal rate of return	30
2.5 Excess returns	32
2.6 Real vs. nominal returns	33
2.7 Returns on levered positions	34
2.8 Returns on short positions	35
2.9 Returns on portfolios	39
2.10 Log-returns	41
2.11 Exercises	44
3 Risk	47
3.1 Random variables and summary statistics	48
3.2 The normal distribution	56
3.3 Multivariate random variables, covariances, and correlations	64
3.4 Computational rules for random variables	68
3.5 Tail risk measures	72
3.6 Return distributions and the investment horizon	77
3.7 Using historical returns	92
3.8 Exercises	101

4 Portfolios	107
4.1 Two-asset portfolio mathematics	107
4.2 Multi-asset portfolio mathematics	122
4.3 Higher-order moments of portfolio returns	135
4.4 Risk reduction through diversification	136
4.5 Special portfolios: arbitrage, replication, and tracking	144
4.6 Exercises	147
5 Bonds	151
5.1 Bond types and characteristics	151
5.2 Bond prices	155
5.3 Bond yields and yield curves	159
5.4 Price relations across bonds	165
5.5 Forward rates	170
5.6 Determinants of the shape of the yield curve	172
5.7 Stylized facts about bond returns and interest rates	174
5.8 Interest rate risk	178
5.9 Immunization	190
5.10 Bonds with risky payments	195
5.11 Bond portfolio management	199
5.12 Concluding remarks	200
5.13 Exercises	201
6 Stocks	207
6.1 The dividend discount model	207
6.2 Prices and other fundamentals than dividends	217
6.3 A decomposition of stock returns	223
6.4 Equity duration	224
6.5 Stylized facts about stock market returns	228
6.6 The cross section of stock returns	242
6.7 Individual stocks	253
6.8 Correlations	255
6.9 Diversification of stock portfolios: an example	259
6.10 Exercises	261
7 One-period portfolio choice	265
7.1 Mean-variance analysis with only risky assets	265
7.2 Mean-variance analysis with both risky assets and a riskfree asset	278
7.3 The optimal portfolio	286
7.4 Discussion and perspectives	290
7.5 Theoretical foundation	298
7.6 Exercises	306
8 Multi-period portfolio choice	315
8.1 Merton's basic model of long-term investments	316
8.2 Alternative preferences	327
8.3 Time-varying investment opportunities	330
8.4 Conclusion	340
8.5 Exercises	341

9 Household portfolio choice	343
9.1 Labor income	344
9.2 Housing	354
9.3 Saving for retirement	362
9.4 Conclusion	364
9.5 Exercises	365
10 The Capital Asset Pricing Model	369
10.1 The basic CAPM	369
10.2 The empirical performance of the CAPM	386
10.3 Alternative versions of the CAPM	394
10.4 The Consumption-based CAPM	400
10.5 Exercises	409
11 Factor models	413
11.1 A general one-factor model	414
11.2 The Single-Index Model	418
11.3 General multi-factor models	424
11.4 The Arbitrage Pricing Theory	428
11.5 The Fama-French models	432
11.6 The factor zoo	438
11.7 Portfolio choice with tradeable factors	446
11.8 Exercises	455
12 Market (in)efficiency and behavioral finance	469
12.1 Efficient markets	469
12.2 Empirical evidence on market efficiency	471
12.3 Investor behavior	475
12.4 Summary	477
12.5 Exercises	477
13 Active portfolio management	481
13.1 The Treynor-Black model	482
13.2 The Black-Litterman model	492
13.3 Performance evaluation	499
13.4 ESG investing	504
13.5 Exercises	516
14 Forwards, futures, and swaps	523
14.1 Forwards	523
14.2 Futures	534
14.3 Swaps	537
14.4 Exercises	544
15 Options	547
15.1 Definition, characteristics, and terminology	547
15.2 Applications	551
15.3 Option markets	556
15.4 Option pricing: general properties	560
15.5 Option pricing: the binomial model	571
15.6 Option pricing: the Black-Scholes model	584
15.7 Option returns	591
15.8 Concluding remarks	594
15.9 Exercises	595

A The lognormal distribution	601
B The Greek alphabet	607
C Short answers to selected exercises	609
Bibliography	619

Preface

Many good textbooks on financial markets and investments already exist with the book by Bodie, Kane, and Marcus (2021) as a popular and excellent example. The book manuscript you are looking at stands out by offering a more rigorous and modern presentation than most existing books. Consequently, the book manuscript relies more on mathematics and statistics than competing presentations, but the necessary quantitative tools are carefully introduced when needed.

For the more mathematical and theoretical parts a theorem-proof style is used. This style serves to ensure that theoretical results are stated in a precise manner and that formal proofs are provided, at least for most of the theoretical results. At the same time the theorem-proof style isolates most of the mathematical derivations in the proofs, which allows for a separation of mathematics from economic intuition and also makes it easy for readers less interested in those proofs to skip them. The economic intuition for the theoretical results is supplied throughout and the practical relevance of the theories is discussed extensively. Examples and exercises illustrate the theoretical results, often with applications of Excel when possible.

As competing books, this book covers classical concepts, methods, and models that remain important in finance theory and practical applications. I have made an effort to include—or at least mention—more recent research whenever relevant. The book covers some topics that are usually not included in similar textbooks, but where substantial progress in research has been made in recent years. To mention a few examples, the book discusses equity duration, life-cycle portfolio decisions with human capital and housing, and the explosion in pricing factors.

I have written this manuscript while being a professor at the Department of Finance at Copenhagen Business School. I gratefully acknowledge contributions of current and former colleagues and students here. I thank Simon Bonde, Thomas Desting Christensen, Linus Christensen, Maximilian Fuchs, Aleksi Keränen, Linda Sandris Larsen, Julian Merz, Peter Raahauge, Julian Terstegge, and Moritz Wedekind for various comments and corrections.

Please, do let me know if you find any typos or other errors in the notes, and if you have any suggestions for improvements I would be happy to hear them. I can be contacted by email at cm.fi@cbs.dk.

Changes relative to previous version (August 11, 2023)

1. Chapter 2. Log-returns now in separate Section 2.11 and not introduced earlier.
Examples 2.1-2.3 updated. Some text revised. Exercise 2.1 updated.
2. Exercises 3.5 and 3.6 updated.
3. Chapter 4. Section 4.4.3 added.
4. New Chapter 9 on household portfolio decisions carved out of former Chapter 8.
Numbers of subsequent chapters changed accordingly. Some exercises from old Chapter 8 moved to new Chapter 9.
5. Chapter 10: examples with beta estimation have been updated.
6. Section 10.4.2: Added explanation of how CCAPM implies CAPM.
7. New Section 13.4 on ESG investing being introduced. Exercise 13.10 added.
8. Data used in some sections of the text and in some exercises (including 2.1, 3.5, 3.6, 11.8, 11.11) has been updated.

CHAPTER 1

Markets

This chapter provides an introduction to financial markets and the investment in financial assets. Section 1.1 briefly explains the organization of trading. Sections 1.2–1.4 describe the main classes of financial assets, namely stocks, bonds, and derivative securities. Some alternative asset classes are listed in Section 1.5. Section 1.6 introduces the main types of investors, and Section 1.7 points out exactly how these investors can benefit from financial markets. Finally, Section 1.8 provides some elementary principles of investing in financial markets.

1.1 Market structures

A financial market is simply a market in which one or more financial assets are traded. Originally, a financial market was a physical location—**an exchange**—where traders met to make agreements to buy or sell certain financial assets with a high degree of transparency in prices and order sizes. Preceded by some semi-formal security markets in Italy, the first known exchange for security trading was established in 1531 in Antwerp, Belgium, where various forms of bonds were traded. Stock exchanges were formed in Amsterdam in 1602, London in 1773, and New York in 1792. Today the trading in most financial markets takes place via an electronic system in which traders can place orders and see how prices evolve. Whether the trading is physical or electronic, an exchange typically regulates the trading in various ways, for example regarding how orders are placed, the order sizes allowed, who can trade, and how deals are settled.

An exchange has an associated clearing house taking care of the clearing and settlement of the transactions made on the exchange. The clearing house is formally the counterparty in any trade on the exchange. A deal between two traders is implemented as two transactions, one between the buyer and the clearing house and the other between the seller and the clearing house. The clearing house guarantees that both traders receive the payments the deal entitles them to, even if the original counterparty defaults and cannot make its promised payment. Hence, the traders on an exchange do generally not have to assess counterparty risk. Typically, the clearing house requires traders who promise some future payments to post collateral in form of margin deposits of a certain size. These deposits are used to cover losses in case of the trader defaulting on his promise.

Some trading takes place in the less organized over-the-counter markets or simply **OTC markets**. (Non-exchange trading venues are also referred to as Alternative Trading Sys-

tems or ATSS.) An OTC market is a computer and telecommunications network linking selected dealers. OTC markets have been much less regulated than exchanges. OTC trades were usually not run through a central clearing house, so the seller and the buyer would enter a bilateral contract in which case an assessment of the counterparty risk becomes relevant. However, in the aftermath of the recent financial crisis, governments and international organizations have pushed for more regulation of the OTC markets. The regulation requires OTC derivative trades to be executed through exchanges or electronic trading platforms and cleared through central counterparties (CCPs). Despite the regulatory push for more exchange-based trading, OTC trading remains prevalent for derivative securities such as futures and options, especially the more exotic species of derivatives. However, OTC markets exist for all main types of securities, including securities that are also traded on organized exchanges.

A significant share of off-exchange security trading takes place in so-called **dark pools**. A dark pool is a private exchange or forum in which investors can trade certain financial securities with other members of the forum. Many dark pools are owned and operated by large financial institutions, others are independent, and some are even owned by companies that also own traditional exchanges. When trading on a traditional exchange, an investor generally has to post how many units of the security she wants to buy or sell and at which price. This may reveal information to other investors who may subsequently place orders that will lead the price to move adversely before the original investor's order has been executed. Dark pools are less transparent and orders are not revealed to other potential investors. This may be advantageous in a given situation, especially for large investors wanting to trade a larger block of securities. On the other hand, the lack of transparency leads to prices being less informative about the true value of the security, so that unwary or unsophisticated investors may trade at unfair prices. Dark pools were originally reserved for large institutional investors, but many dark pools now attract other investors, including retail investors, by allowing smaller trades with faster processing and lower transaction costs than at traditional exchanges.

1.2 Stocks

1.2.1 What is a stock?

A stock is a security issued by a corporation. A **common stock** entitles its holder to a share in the ownership of the company, which includes both cash flow rights and control rights. Cash flow rights mean that the stockholder has a claim on a share of the dividends and other payouts of the company. Control rights mean that the stockholder has the right to vote on matters of corporate policy and in elections for members of the company's board of directors. Typically, a company issues a very large number of common stocks and all the stocks are treated equally in terms of dividends and voting rights.

Some companies further issue a number of **preferred stocks**. When paying dividends, the company must pay the preferred stockholders up to a certain level before it can offer dividends to common stockholders. On the other hand, preferred stocks do not carry voting rights. When referring to stocks, we will generally mean common stock, and also use the terms shares and equity. Likewise, the terms stockholder and shareholder are used interchangeably.

Shareholders are residual claimants to the assets of the company. The creditors of the company have first priority. If the company is to be liquidated, the proceeds from the sale of the company's assets are first to cover as much as possible of the creditors' claims. The stockholders will only get something if the assets are worth more than the creditors'

claims. On the other hand, shareholders have limited liability, which means that they can never lose more than their original investment. If the company is being liquidated and it turns out the assets are not valuable enough to cover the claims of the creditors, the shareholders are not forced or even expected to pay the difference out of their own pockets. They are not personally liable for the obligations made by the company.

Companies can issue stocks through an initial public offering (IPO) where the new stocks are sold following some sort of auction process. The investors pay by cash that enters the company, so a stock issuance is a way for companies to raise money for investment projects. Subsequently, the stocks are typically listed and traded on an exchange, which enables an easy transfer of partial ownership of the company among investors. A listed company can issue additional stocks on the exchange through a seasoned equity offering (SEO).

In some situations, stocks are issued through private arrangements, where the company offers a large equity position to one or a few private investors, again in return for cash. This is referred to as **private equity**. Often these private investors also obtain seats on the board of directors with the purpose of substantially changing the management, operations, or strategy of the company. Sometimes the private equity investor buys a majority share of the equity to obtain full control of the company in order to implement a certain turnaround. The investor hopes that after a period of 5-10 years the company is much more valuable, and then the private equity investor may sell off his equity position maybe through an initial public offering of stocks in the company and cash in a substantial profit. Private equity investments are often made by specialized institutional investors who, in addition to the cash invested, also have the human resources and skills necessary to improve the company's performance.

1.2.2 Stock markets

Stock markets are huge in terms of the number of listed companies, the number of shares traded, and the market capitalization, i.e., the market value of the listed shares. Based on 2020 data published by the World Bank, the U.S. stock markets are by far the largest in the world with a total market capitalization of 40.7 trillion USD, followed by China (12.2 trillion USD), Japan, Hong Kong, the United Kingdom, Canada, India Saudi Arabia, France, Germany, and the Republic of Korea.¹ When measured relative to annual GDP (2020 data), the stock market capitalization is largest in Hong Kong (1,769%), followed by Iran (599%), Saudi Arabia, South Africa, Switzerland, Singapore, the United States, and Canada. In terms of the value of the stocks traded during the year relative to country GDP (2020 data), Hong Kong also tops the rank with 886%. According to SIFMA, the Securities Industry and Financial Markets Association, the average daily trading volume in 2021 was 119.7 billion USD at the NASDAQ stock exchange, 119.6 billion at the Intercontinental Exchanges (dominated by the New York Stock Exchange, NYSE), 83.4 billion at the CBOE exchanges (called BZX, BYX, EDGA, EDGX), 29.9 billion at other exchanges, and 212.2 billion off exchanges.² As of 2020, the stocks of 4,266 domestic companies were listed in the U.S., second only to India with 5,215 companies, and followed by China, Canada, Japan, Spain, Hong Kong, and the Republic of Korea.

According to statistics published by the World Federation of Exchanges, the leading stock exchanges in terms of market capitalization are the New York Stock Exchange, NYSE, with a market cap of 27.7 trillion USD (December 2021), followed by NASDAQ

¹See <http://wdi.worldbank.org/table/5.4>. The rank for the U.K. is based on 2010 data, the latest available data for the U.K.

²See the statistics at <http://www.sifma.org/research/statistics.aspx>. The name NASDAQ is short for National Association of Securities Dealers Automated Quotations.

(24.6 trillion USD), the Shanghai Stock Exchange (8.2 trillion USD), the Euronext exchange covering Amsterdam, Brussels, Dublin, Lisbon, Milan, Oslo, and Paris (7.3 trillion USD), the Japanese Exchange Group (6.5 trillion USD, dominated by the Tokyo Stock Exchange), the Shenzhen Stock Exchange (6.2 trillion USD), the Hong Kong Exchanges (5.4 trillion USD), and the London Stock Exchange (3.8 trillion USD).³

1.2.3 Stock indices

The stock markets around the world offer an abundance of investment opportunities. Keeping track of so many individual stocks is impossible for any investor. To get an indication of the overall performance of the stock market or a certain section of the stock market, investors follow **stock indices**. Some indices are created to reflect the global stock market. For example, as of May 2022 the MSCI All Country World Index includes 2,933 stocks from 23 developed countries and 24 emerging countries, whereas the MSCI World index includes 1,540 stocks from 23 developed countries.⁴ The overall performance of the stocks traded in a certain country, on a certain exchange, or in a certain industry can be measured by a stock market index for that subset of stocks.

For most countries, one or more stock indices attempt to capture the overall stock market in that country. For the U.S., a leading stock index is the Standard & Poors 500 index—or simply the S&P500—which is based on around 500 large stocks listed on either NYSE or NASDAQ. Another popular U.S. stock index is the Dow Jones Industrial Average or simply the Dow Jones, which is based on 30 large listed companies, and thus a much more narrow index than the S&P500. A very broad index is the Russell 3000 index derived from 3,000 listed U.S. companies representing around 97% of the U.S. equity market. Focusing on the smallest approximately 2,000 stocks, the Russell 2000 provides a good benchmark for the performance of small company stocks, in contrast to the Dow Jones and the S&P500. The Wilshire 5000 total market index embraces all stocks actively traded except from stocks with very low prices (penny stocks) and stocks in very small companies.

Some leading stock indices of other major stock markets are the Shanghai Stock Exchange Composite Index from China, the Nikkei 225 index from Japan, the FTSE 100 index (FTSE is often pronounced “footsie” and is short for Financial Times Stock Exchange) from the U.K., the S&P/TSX Composite Index from Canada, the CAC 40 index from France, and the DAX index from Germany. Some indices cover a geographical region. For example, various European stock indices now exist, such as the Euro Stoxx 50 and the Stoxx Europe 600 (again the names reflect the number of stocks included). There are also specialized indices for stocks in certain sectors or stocks with certain characteristics, either within a single country or across countries.

We are not going into details on how a stock index is computed from its constituent stocks. Indices differ with respect to how the constituents are weighted (equally-weighted or value-weighted, maybe only counting the “float”, i.e., the stocks that are available for public trading) and how dividends are accounted for. In fact, most major country-specific stock market indices ignore dividends (Hartzmark and Solomon 2021). Also note that the indices are regularly revised with some stocks leaving and other stocks entering the index according to some stated guidelines.

The indices are not only providing an overview of the performance of the stock markets. They also serve as important benchmarks for investments in the stocks in a given country,

³See <https://statistics.world-exchanges.org/>.

⁴MSCI is short for Morgan Stanley Capital International but, since 2009, MSCI is independent of Morgan Stanley and is listed on the New York Stock Exchange. Detailed information about their indices can be found at <http://www.msci.com>.

region, or industry. By comparing the returns on an investment in a stock to an appropriate index, an investor gets an indication of her performance.

1.2.4 Investing in stocks via funds

As we shall see in subsequent chapters, financial theory recommends to invest in a portfolio of many different stocks as this reduces risk without reducing the expected return. While a stock index keeps track of the value of a certain stock portfolio, you cannot directly invest in a stock index. In principle, you could invest in shares of all the companies by matching the weights they are given when computing the index, but this would be tedious and involve substantial transaction costs. Furthermore, you would need to have a significant amount of cash to purchase just one share in each of the companies, and to match the relative index weights you need to buy many shares in the large companies.

For many indices there is a simpler solution. You can invest in an **exchange-traded fund**, or simply **ETF**, mimicking the index. An exchange-traded fund is an investment fund which owns certain assets, and the ownership of the fund itself is divided into shares that are traded on an exchange, just as the stocks of any individual company. An ETF tracking an index can do so by actually holding the portfolio of stocks that constitutes the index and, of course, rebalancing the portfolio along with changes in the weights of the index. With many investors in the ETF, the fund will have sufficient capital to purchase stocks in all the companies in the correct proportions.

Exchange-traded funds were introduced in the U.S. in 1993, and the ETF market has grown substantially in recent years, especially in the U.S. Many of the ETFs are “sponsored” (constructed, issued, and managed) by a few financial companies, namely BlackRock, State Street Corporation, or Vanguard. The largest ETFs track the S&P 500 index. This includes the so-called Spider (traded under the ticker name SPY) issued by State Street (fund market cap of 378 billion USD as of June 29, 2022), BlackRock’s iShares IVV (301 billion USD), and Vanguard’s VOO (766 billion USD). Other popular ETFs track the CRSP U.S. Total Market Index (VTI by Vanguard), the Dow Jones Industrial Average index (DIA, called diamonds, by State Street), and the NASDAQ 100 (ticker QQQ, called cubes, by Invesco). Other ETFs track indices covering specific industry sectors or stocks with particular characteristics related to the size of the issuing company (large-cap, small-cap), their dividend payments, or the ratio of their current stock price to the book value of equity (value stocks, growth stocks). ETFs tracking select foreign stock markets are also traded in the U.S. The funds must do some trading in the underlying stocks, and some administration has to be taken care of, and the investors in ETFs have to cover these costs. The exact magnitude of the costs varies across ETFs. Some of the highly popular ETFs charge only 0.1% (or even less) of the invested amount per year, whereas more specialized (or less competitive) ETFs charge more, maybe up to 1%.⁵

Many individuals invest in the stock markets via **mutual funds**. A mutual fund is an investment company that issues shares of ownership to the assets the fund holds. These shares are not tradeable among investors, but investors can purchase new shares from the fund or redeem shares to the fund, so that the number of outstanding shares may vary over time. Such funds are called *open-end funds*. The price at which mutual fund shares are issued and redeemed is determined by the net asset value (NAV) of the fund, which is the market value of the fund’s assets minus its liabilities. The price per share is simply the NAV divided by the number of shares outstanding. The shares are redeemable on a

⁵See Lettau and Madhavan (2018) for more information about the construction of ETFs and a discussion of some potential concerns caused by the increasing popularity of ETFs.

daily basis.

Some mutual funds are *passively managed* in the sense that their goal is just to track a certain index without trying to outperform the index. These index funds are thus very similar to ETFs, and like them they should be able to operate at low costs and thus charge only low fees. Other mutual funds are *actively managed* whose defined purpose is to outperform a certain benchmark by actively searching for mispriced assets. As this search process demands the precious time of clever and well-paid analysts, actively managed funds have higher costs and, consequently, they take higher fees (often 1-2% of assets per year) from their customers, who are then betting that those fees are more than outweighed by higher returns on the investments. The mutual fund industry is huge with a large variety of funds operating in different segments of the stock markets.

As a variation on the theme, *closed-end funds* are constructed with a fixed number of issued shares. The shares are not redeemable—they cannot be sold back to the fund company—but they are directly tradeable on an exchange, just like ordinary shares of stock or ETFs. The price of a closed-end fund's shares is thus determined by supply and demand, and often deviates from the NAV with more funds trading below the NAV than above.

The Investment Company Institute (a global association of investment funds) reports that the total value of assets held by mutual funds worldwide was 71.1 trillion USD at the end of 2021. U.S. investment companies held assets worth 34.6 trillion USD, of which 27.0 trillion were owned by mutual funds, 7.2 trillion by exchange-traded funds, and 309 billion by closed-end funds. In total, U.S. investment companies own around 32% of the U.S. stock market. More than 47% of U.S. households own mutual funds with a median value of around \$200,000 per household. While the value of the assets held by U.S. mutual funds has roughly doubled from 2011 to 2021, the 2021 value of ETFs was almost seven times the 2011 value underlining the high growth rate of that part of the investment fund industry.⁶

1.3 Bonds and other debt-related securities

1.3.1 Bond types and issuers

Both governments, companies, and households occasionally seek to borrow money. Governments often finance a budget deficit by issuing bonds in return for cash. The government then commits to deliver a cash flow stream to the owner of the bond in the form of regular interest payments and a repayment of the borrowed amount (the face value of the bond) according to some specified amortization schedule. The maturity date of the bond is when the last payment is due. The bonds are typically listed and traded on an exchange. Most governments issue bonds with a number of different maturities, typically ranging from around one month and up to 30 years.

In the United States, the **government bonds** are issued by the Department of the Treasury and are divided into four types. *Treasury bills* (or T-bills) are issued with a maturity of up to one year, with the most common maturities being one month, three months, six months, and one year. The T-bills have no intermediate interest payments, but only a payment of the face value at the maturity date; bonds with this feature are called zero-coupon bonds. Since T-bills are typically issued at a price below the face value, investors in T-bills still earn a profit. *Treasury notes* (or T-notes) are issued with a maturity between two and ten years. They promise semi-annual interest payments, typically referred to as coupon payments, and they are bullet bonds meaning that the

⁶See <https://www.icifactbook.org/>.

entire face value is paid back at the maturity date. *Treasury bonds* (or T-bonds) are issued with longer maturities, currently up to 30 years, and are also bullet bonds. The above-mentioned bonds are all nominal bonds in the sense that they promise a pre-specified dollar payment stream. But since consumer prices vary, nominal bonds do not guarantee a certain purchasing power. The U.S. Treasury also issues inflation-indexed bonds called Treasury Inflation-Protected Securities (*TIPS*, in short) of different maturities, currently 5, 10, and 30 years. The face value of these bonds is regularly adjusted to reflect changes in the Consumer Price Index. In the United States and some other countries, local governments (municipalities) also issue bonds.

Likewise, companies may issue bonds to finance current operations or new investment projects. Such bonds are called **corporate bonds**. In countries like the United States and the United Kingdom, this is a common way for corporations to obtain loans, whereas companies in other countries tend to rely more heavily on bank loans. Bonds are also issued by various agencies, international institutions, etc. Sometimes the bonds are privately placed at selected investors, but often the bonds are listed and traded on an exchange.

1.3.2 Related securities

Some large corporations regularly issue **commercial papers**, which are short-term (up to nine months, but often only one or a few weeks) contracts similar to zero-coupon bonds. Some commercial papers are asset backed in the sense that the issuer puts up collateral to ensure the promised future payment. The total outstanding face value in the U.S. commercial paper market is around one trillion USD (mid 2020) with around 20% being asset backed, according to the Securities and Exchange Commission. Alternatively, companies may borrow directly from a bank, for example by having a line of credit.

Financial institutions may also issue bonds to finance their operations. Commercial banks take deposits from companies and households with excess capital, and can then lend out the money to capital-demanding entities. Banks sometimes issue **certificates of deposits** (often abbreviated CD), which are contracts formalizing a fixed-period deposit at a pre-determined interest rate. The buyer effectively lends the money to the bank for a fixed period of time, which typically ranges from one month up to several years. Some CDs are tradeable after issuance.

Large financial institutions often engage in repurchase agreements, or simply **repos**. Party A sells a security to party B and commits to buying back the same security at a pre-specified price and point in time. For party A this is a repo, for party B it is called a reverse repo. Effectively, a repo constitutes a secured loan, in which the security serves as the collateral. The security is often a government bond, and the maturity of the repo is often only one to seven days.

An **interbank market** exists in which banks negotiate short-term loan agreements among each other, mainly to manage day-to-day liquidity needs. Most loans in this market are simply overnight loans. The interest rates set in this market are, on the one hand, heavily influenced by the central banks that through various activities can affect the demand and supply for liquidity in the banking sector. On the other hand, the interbank market is a main driver of the interest rates that banks offer and charge both corporate and household customers. In many cases, the interest rate that a bank charges on a corporate or household loan is determined as an interbank interest rate plus a surcharge depending on the credit quality of the borrower. Conversely, the deposit rate offered by a bank can be stipulated as a certain interbank rate less some deduction. The interbank rates also function as benchmarks for some floating rate bonds, adjustable-rate mortgages, and other bond-like securities. The interbank market thus plays an important role in the

transmission of monetary policy from central banks to the economy at large.

A leading U.S. interest rate is the **federal funds rate**, the interest rate at which banks borrow and lend reserve balances overnight, i.e. balances held at the Federal Reserve, the central bank of the United States. The Federal Open Market Committee regularly meets to set a target for the federal funds rate, and the Federal Reserve then tries to implement that target through transactions with commercial banks or by buying or selling government bonds or other securities either in the open market or through repos. Other central banks control short-term interest rates of their respective currencies using similar means.

Another important short-term U.S. interest rate is the **Secured Overnight Financing Rate (SOFR)** which is based on the rates charged on overnight Treasury repos in the interbank market. SOFR was introduced in 2021 as the designated successor of the so-called LIBOR rate that used to be a central reference interest rate. LIBOR is an acronym for London InterBank Offered Rate, and LIBOR rates used to be published for selected maturities ranging from overnight to one year and for loans in the major currencies. The LIBOR rates were based on estimates of borrowing costs submitted by banks and, given the huge markets for LIBOR-benchmarked products, banks could have incentives to submit estimates not reflecting actual borrowing costs. After documented cases of banks manipulating the fixing of LIBOR rates, a transition from LIBOR to transaction-based short-term rates was decided throughout the international financial system. As stated above, SOFR is replacing the overnight LIBOR rate on the US dollar, and SOFR-related rates for other short maturities are also being introduced. In the Euro currency area, the future benchmark is the **Euro Short-Term Rate (€STR)** which reflects unsecured overnight borrowing in Euros between banks. In the U.K., the so-called **Sterling Overnight Index Average (SONIA)** seems to replace role of the overnight LIBOR rate on the Pound Sterling.

Households primarily borrow money in banks or specialized credit institutions. The largest loans to households are typically related to the purchase of an apartment or a house. A **mortgage** is a loan offered to the owner of real estate, where the real estate is used as a collateral for the loan. Traditional mortgages are issued as annuity loans (equal payments) with long maturities, say, 30 years. In some countries mortgages are offered by banks, but in other countries specialized mortgage institutions exist. These institutions offer mortgages to a large number of households (and companies) and finance the mortgages by issuing bonds. These bonds are backed by a certain pool of the mortgages in the sense that the payments flowing to the bond owners depend on the borrowers' actual payments on their mortgages. The actual payments may differ from the promised payments due to the borrower defaulting on the loan or deciding to pay back the loan pre-maturely (maybe to refinance).

1.3.3 Debt markets

The markets for bonds and other debt-related contracts are often split into a **money market** and a **fixed-income market**. The money market consists of the short-term debt instruments maturing in at most one year, which therefore includes Treasury bills, some certificates of deposit, commercial papers, interbank loans, and repos. The fixed-income market consists of the longer-term debt instruments such as Treasury notes and bonds, TIPS, mortgage-backed bonds, municipal bonds, and corporate bonds.

The markets for debt securities are huge as illustrated by the following numbers. According to the Bank of International Settlements, BIS, the total amount outstanding for debt securities issued by U.S.-based entities is 49.1 trillion USD (December 2021) of which 25.8 trillion was issued by the government, 15.7 trillion by financial corporations, and 7.4

trillion by non-financial corporations. China is second in the statistics with a total of 21.8 trillion USD issued of which 8.8 trillion is government debt. Next are Japan, the U.K., France, China, Germany, and Italy.⁷ According to SIFMA, the total amount outstanding in the U.S. fixed income securities is 52.8 trillion USD (forth quarter, 2021) of which 22.6 trillion are Treasury bonds, 12.2 trillion are mortgage-related bonds, 9.9 trillion are corporate bonds, 4.1 trillion are municipal bonds, 1.6 trillion are asset-backed bonds, 1.4 trillion are federal agency securities, and 1.0 trillion are money-market instruments. The average daily trading volume of U.S. Treasury securities is around 600 billion USD.

1.3.4 Bond indices and funds

As for stocks, various bond market indices are also published and used as benchmarks for bond investment performance. However, indices are less relevant for bonds than for stocks. Even within the same country or industry, the stocks of different countries may perform very differently. The performance of bonds is less diverse. The prices of, and thus the rates of return on, the bonds issued by the U.S. government follow each other quite closely. Of course, the returns vary somewhat with the maturity of the bond, but by looking at just a few key maturities investors obtain a good impression of the performance of government bonds with different maturities. Hence, an index averaging all U.S. government bonds does not provide much additional information. The diversity among corporate bonds is bigger as their performance to some extent depend on the profitability of the issuing corporations, so corporate bond indices can be relevant. Similarly, for bond investments across a group of different countries, such as the emerging market countries.

Numerous bond ETFs are traded, some focusing on Treasury bonds, some on corporate bonds, and others on international bonds markets. Such ETFs provide a (typically) cost-efficient way to obtain exposure to various segments of the overall bond market. Moreover, various mutual funds focus on different bond types and markets, and dedicated money-market funds are also attracting many investors.

1.4 Derivative securities

1.4.1 Main types of derivatives

A derivative security or simply a **derivative** is an asset whose dividend(s) and price are derived from the price of another asset, the *underlying* asset, or the value of some other variable such as an interest rate. The four main types of derivatives are forwards, futures, swaps, and options. While a large number of different derivatives are traded in today's financial markets, most of them are variations of these four main types.

A forward contract or simply a **forward** is a binding agreement between two parties to transact a given asset at a pre-specified future point in time and at a pre-specified price. If you know that you have to purchase or have to sell a specific asset at a future date, you can use a forward on that asset to lock in the purchasing or selling price already today. You eliminate the risk of an adverse move in the asset price, but you also give up the chance of a price move in your favor. Forward contracts thus have obvious application for risk management. On the other hand, forwards can also be used for speculation in the future price of the underlying asset.

The profits and losses from an investment in a forward contract are settled at its maturity date. **Futures** contracts are similar to forwards, but with the profits and losses being settled on a daily basis throughout the life of the contract—the so-called marking-

⁷The statistics were found on <http://www.bis.org/statistics/secstats.htm>.

to-market mechanism. This feature allows futures contracts to be traded on organized exchanges, whereas forward contracts are traded over the counter. Futures on stock indices, individual stocks, government bonds, and foreign currency are traded on a large scale on many exchanges around the world. There are also large markets for futures on many different commodities, such as oil, gold, aluminium, electricity, soy bean oil, frozen concentrated orange juice, and live cattle.

A **swap** refers to an agreement to exchange two specified streams of payments. For example, an interest rate swap is the exchange of two streams of interest rate payments on a certain face value, where typically one stream is calculated using a fixed, known interest rate, whereas the other depends on how a certain market interest rate evolves over the life of the contract. Another example is a currency swap where the two payment streams are denominated in different currencies and thus involve an exchange rate.

An **option** is an asset giving the owner the right, but not the obligation, to perform a certain transaction in the future at terms specified today. Typically this transaction is to purchase or sell a given underlying asset at a pre-set price at or before a given future date. Call options give their owner the right to buy the underlying, put options give the right to sell the underlying. Options on stock indices, individual stocks, bonds, interest rates, and exchange rates are traded in many markets. Sometimes the underlying asset itself is a derivative. For example, you can trade options on futures on many commodities.

1.4.2 Markets for derivatives

Some key statistics on the size of derivatives markets are listed in Table 1.1 for exchanges and Table 1.2 for OTC markets. The statistics are published by the Bank for International Settlements (BIS). The derivatives markets are large. The total nominal amount outstanding (the value of the underlying assets) is around 80.1 trillion USD on organized exchanges and 587.5 trillion USD on OTC markets according to the BIS estimates at year-end 2021. The average daily turnover of exchange-traded derivatives is around 7.4 billion USD, substantially smaller than the daily turnover of stocks and government bonds, but still a sizeable amount and the unknown, but surely substantial, OTC derivatives turnover should be added.

Derivatives mostly have maturities less than a year, but longer-maturity contracts also exist, in particular among interest rate derivatives.

The equity derivatives have a fairly even split between options on the one hand and forwards, futures, and swaps on the other hand. For derivatives on interest rates or foreign exchange, forwards, futures, and swaps dominate options.

BIS also provides information on the currencies in which the derivatives are traded. Approximately 66.3% of the exchange-traded derivatives (measured by 2021 turnover) are denominated in U.S. dollars, 17.9% in Euro, 11.1% in Pound Sterling, followed by Australian and Canadian dollars, Japanese Yen, Chinese Renminbi, and Swiss Franc.

1.5 Alternative asset classes

Commodities are by many institutional investors nowadays seen as an asset class. As an investor you can get exposure to commodities through certain financial securities without ever owning the commodities physically. One possibility is to buy or sell commodity futures or even options on such commodity futures. The profits from such investments are purely determined by the evolution in the price of the underlying commodity. You can also buy ETFs tracking the prices of select commodities. An extremely popular ETF is the SPDR Gold Trust (ticker symbol GLD), which has a total value of 62 billion USD

	Futures		Options	
	Outstanding	Turnover	Outstanding	Turnover
All markets	34,129	5,863	45,961	1,523
Interest rate	99.1%	97.6%	99.7%	99.1%
Foreign exchange	0.9%	2.4%	0.3%	0.9%
North America	68.9%	65.5%	65.4%	74.5%
Europe	25.2%	29.9%	32.8%	24.1%
Asia-Pacific	3.5%	2.9%	0.0%	0.6%
Other markets	2.4%	1.7%	1.8%	0.8%

Table 1.1: Derivatives traded on organized exchanges.

Amounts are in billions of U.S. dollars. The amount outstanding is of December 2021, whereas the turnover is the daily average turnover in 2021. Source: Bank of International Settlements, <http://stats.bis.org/statx/srs/table/d1>, retrieved June 30, 2022.

Contracts	Interest rates	Foreign exchange	Equity linked
All, amount outstanding	598,416	475,271	104,249
All, gross market value	12,439	8,612	2,548
<i>Instrument</i>			
Forwards and swaps	91.9%	89.9%	54.5%
Options	8.1%	10.0%	45.5%
<i>Maturity</i>			
Up to 1 year	40.7%	77.7%	63.3%
1-5 years	35.6%	15.4%	31.5%
Over 5 years	23.6%	6.8%	5.2%

Table 1.2: Derivatives traded OTC.

All amounts are in billions of U.S. dollars. The percentages are relative to the amount outstanding. The statistics are from the second half of 2021. Source: Bank of International Settlements, <http://stats.bis.org/statx/srs/table/d5.1>, retrieved June 30, 2022.

as of June 28, 2022.⁸ Another possibility is, of course, to buy shares of companies active in commodity markets or even to set up your own facilities for producing a certain commodity. Some pension funds are, in fact, investing directly in forests and land.⁹

Real estate is an important asset class in terms of its total value. For example, housing wealth constitutes a large share of household assets. In 2019 the value of residential property owned by U.S. households was 32% of total household wealth, and for middle-income households the share is even larger.¹⁰ In 2020, the value of all homes in the United States was estimated by the company Zillow to be 36.2 trillion USD.¹¹ Commercial real estate has to be added to that.

Pension funds and other large institutional investors (and some relatively wealthy individual investors) have been actively investing in real estate for many years. They own various buildings, rent them out to corporations or households, and cash in the rents. In addition, they may hope to resell the real estate later at a profit. Many households own the home they live in, which does not only provide shelter and pleasure but is also a major investment.

Even as a small investor (home owner or renter) you can obtain an investment exposure to real estate markets without physically buying real estate. You can invest in the shares of real estate investment trusts, the so-called REITs. These are investment companies dedicated to investing in more or less specialized real estate or in mortgages. Typically, a REIT is established and managed by a bank or another financial institution, but the shares of the REIT are traded on stock exchanges just as the stocks of other companies. In the United States you can even purchase REIT ETFs, that is, shares of an exchange-traded fund holding a basket of stocks in real estate investment trusts.

Another alternative, available only in some countries, is to invest in derivative securities written on some index of house prices. For example, at the Chicago Mercantile Exchange you can trade futures contracts on the so-called S&P Case-Shiller home price indices for certain U.S. metropolitan areas and a composite nationwide index. When investing in such contracts, your profits depend on the future home prices.

Long-term institutional investors, such as pension funds, are sometimes investing in infrastructure, i.e. roads, bridges, tunnels, etc., often in corporation with the local government. The pension fund finances the construction costs and receives a long-term payment stream in return.

Some investors, in particular rather wealthy individuals (or wannabes), apparently think of art, antiques, wine, stamps, and the like to be serious investment objects. Little (but some) academic research has been conducted on the qualities of such investments, so this is still a rather dark area dominated by newspaper stories of the success of some individual investors. Remember the information bias: successful investors are much more likely to share their experience (and enjoy the admiration of others) than the many less successful investors who prefer to keep quiet. One thing is sure: the markets for such assets are much less liquid, transparent, and organized than traditional financial markets.

⁸Source: <https://etfdb.com/etfs/asset-class/commodity/>

⁹For more on commodity investments, see Ang (2014, p. 364-374).

¹⁰Data is based on the Survey of Consumer Finances, see Tables 8 and 9.1 that can be downloaded from <https://www.federalreserve.gov/econres/scfindex.htm>, and residential property includes both primary residence and other residential property but not equity in non-residential property. See also Campbell (2006, Figures 2 and 3) and Guiso and Sodini (2013).

¹¹Source: <https://www.zillow.com/research/zillow-total-housing-value-2020-28704/>

1.6 Main players

This section summarizes the main players in financial markets and briefly explains how they use the markets.

Firms are primarily demanding capital to finance investments. They can do so by issuing stocks or corporate bonds or both. They can use derivatives to manage mostly short-term risks. Short-term deposits and loans facilitate corporate liquidity management.

Households or individuals are primarily supplying capital. In order to smooth consumption over the life-cycle, young and middle-aged individuals typically save part of their income for financing consumption in retirement. Much of the retirement savings are invested in stocks and bonds. Individuals may also save temporarily to finance large future expenditures such as the down payment on a house or an apartment. Since individuals tend to dislike risk and uncertainty, many individuals maintain some buffer of savings for a rainy day, e.g., to finance some consumption even in times of unemployment or bad health. On the other hand, households are often borrowing money to partially finance larger expenditures such as residential real estate. It is important for households to consider their financial assets and liabilities in connection with the other assets they hold, e.g., the real estate they may own and the human capital they hold in form of their ability to earn labor income.

Many **governments** are primarily demanding capital since they run budget deficits which they finance by issuing bonds and similar securities. However, several countries have set up **Sovereign Wealth Funds** and some of these funds are among the largest investors in the world. An example is the Norwegian Government Pension Fund, which was established in 1990 and until 2006 was known as the Petroleum Fund. The fund is financed by the government's profits from the oil and natural gas sector and thus not really a pension fund (which is financed by contributions of pension savers), but at the discretion of the government the fund can contribute to the future financing of the public pensions paid by the government to the citizens of Norway. At the end of 2021 the fund held assets worth around 1.4 trillion USD. The capital was invested in equity (72.0%),¹² fixed income (25.4%), real estate (2.5%), and renewable energy infrastructure (0.1%). Other large sovereign wealth funds are based in China and several Arab countries.

The **central bank** of a country controls the country's supply of money, which is of key importance for the interest rates and the inflation rate of the country as well as the exchange rates between the currency of the country and other currencies. The central bank can increase the money supply by offering commercial banks to borrow money at a low interest rate in the central bank. Subsequently, this allows the commercial banks to lend money to corporations and households at fairly low interest rates, which could then stimulate consumption and investment throughout the economy. Conversely, the central bank can decrease the money supply by offering commercial banks to deposit money at a high interest rate in the central bank. The high interest rate then spills over to the loans commercial banks offer consumers and corporations, which tends to lower consumption and investment in the economy. Similarly, central banks can provide or reduce liquidity in the financial market by buying or selling government bonds of various maturities. Central banks can also affect financial markets by setting the reserve requirements for commercial banks that force them to hold deposits of a given size in the central bank. Furthermore, most central banks have the authority to supervise and regulate financial markets in other manners in order to maintain a sound banking system.

After the financial crisis that peaked in 2007-2008, central banks have taken a prime

¹²Source: <https://www.nbim.no/en/the-fund/investments/>, retrieved July 1, 2022.

role in the attempt to stabilize the financial system and to push economic growth. Several key central banks lowered their target interest rates to very low levels, in some cases even to zero or negative levels. They further engaged in so-called quantitative easing whereby the central banks spend huge sums on buying various financial assets from financial institutions, including assets that are substantially riskier than the government bonds central banks traditionally trade.

Financial intermediaries include commercial banks, investment banks, investment companies (including pension funds, mutual funds, and hedge funds), mortgage institutions, and brokers. The prime role of some financial intermediaries is to act on behalf of a group of households/individuals. Other intermediaries try to bring lenders and borrowers together and equate supply and demand. Other again just facilitate financial transactions, and finally some intermediaries focus on providing advice to other market participants.

In addition to the mutual funds and ETFs already described, the investment companies include pension funds and hedge funds. Here, **pension funds** or retirement funds invest the savings of their members to provide retirement income. These funds are established or controlled by the government, labor market organizations, or each employer. Many pension funds are very large investors in financial markets. For example, the Japanese Government Pension Fund—the pension fund for Japanese public sector employees—has a capitalization of 1.8 trillion USD (end of 2020). **Hedge funds** are investment companies that are organized as private partnerships and are thus less regulated than other investment companies. They are typically open to wealthy or institutional investors, and investors have to lock up their investments in the fund for some period. Hedge funds are known to follow more “extreme” investment strategies than other funds, strategies that sometimes resemble bets that certain macroeconomic or political events will happen or that some apparent mispricing in financial markets will be corrected. Hedge funds take large management fees from investors, for example, a flat fee of 2% of the invested amount plus 20% of the returns earned.

Foundations and endowments. Foundations in the U.S. manage assets worth around 1 trillion USD with the largest being the Bill & Melinda Gates Foundation with assets of around 50 billion USD. The associated foundation trust manages a stock portfolio worth around 20 billion USD with about half invested in Berkshire Hathaway, the conglomerate holding company led by Warren Buffett.¹³ Several U.S. universities maintain substantial endowment funds held in financial assets with the returns contributing to the financing of operational expenses. At year-end 2021, Harvard University had the largest endowment with assets worth 51.9 billion USD, followed by Yale, Stanford, Princeton, and MIT. In October 2021, 34% of Harvards endowment fund was invested in private equity, 14% in public equity, 33% in hedge funds, 8% in cash and cash-like assets, 5% in real estate, 4% in bonds, and the remaining 2% in other real assets and natural resources.¹⁴

This book focuses on **investments** (supply of capital) and their consequences for market prices of financial assets. Note that all “end users” are individuals. Pension funds, financial and non-financial corporations, and even hedge funds are owned and controlled by some individuals, and should therefore operate in the best interest of these individuals. Hence, it makes sense to pay special attention to how individuals—households if you like—optimally invest in financial assets.

Who owns the financial assets? First, consider the U.S. stock markets. Based on 2019

¹³ See <https://www.gatesfoundation.org/about/foundation-fact-sheet> and <https://hedgefollow.com/funds/Bill+And+Melinda+Gates+Foundation+Trust>, accessed on July 1, 2022.

¹⁴ See <https://www.nacubo.org/Research/2021/Public-NTSE-Tables> and <https://www.hmc.harvard.edu/partners-performance/>, accessed on July 1, 2022.

data, households own 37.7% (but this includes hedge funds and endowment funds), mutual funds own 21.8%, foreign investors 15.1%, ETFs 6.4%, states and local governments 5.5%, private pension companies 5.4%, non-financial corporations 4.2%, life insurance companies 1.2%, and the Federal Reserve and government agencies 0.7%. Turning to U.S. Treasury securities, as of July 1, 2021, the Federal Reserve and government agencies (in particular the Social Security trusts) own 39.9%, foreign investors own 26.3% (with Japan and China as the largest holders), mutual funds 12.1%, depository institutions 5.0%, state and local governments 4.6%, pension funds 4.5%, and insurance companies own 1.5%.¹⁵

1.7 Functions

What does the financial system do? Its prime role is to **bring suppliers of capital and users of capital together**. *Suppliers of capital* currently have excess capital that they want to save or invest for later use. They can do so, for example, by depositing the money in a bank, buying bonds and thus lending money to the issuer, or buying shares of stock in a company. The suppliers include households saving for retirement and firms with profits they do not want to distribute to their owners.

Users of capital have a need for capital to finance current expenditures in excess of current income. This includes households seeking to purchase residential real estate or simply having—hopefully, only temporarily—negative net income. Households typically obtain capital by borrowing from banks, mortgage institutions (who may then issue bonds financing the mortgage), and other credit-supplying companies. Entrepreneurs starting a business often need outside capital to finance buildings, equipment, and employees. Existing firms may require outside capital in order to finance expansions in production capacity. The capital can be raised by issuing corporate bonds, which means that the company is effectively borrowing money from the buyers of the bonds. In some countries, corporate bonds are very rarely issued, and firms borrow directly from banks and other financial institutions. Instead of borrowing, firms can issue shares of stock to investors who, in return for the share price they pay, obtain an ownership stake in the company. Firms may also have short-term liquidity needs, which are covered by loans in banks or the money market. Governments are in many cases also users of capital. A budget deficit is typically financed by the issuance of government bonds.

The above examples highlight that a key function of financial markets is to **transfer resources** through time and across both countries and industries. By saving for retirement, a household transfers resources from their active, income-earning phase of the life cycle to the retirement phase. Without the chance to save, the consumption of a household in any given period would be limited to its income in the period. The financial markets help households smoothing their consumption over life. Likewise, financial markets enable the transfer of capital from countries or industries with excess capital to other countries or industries in need of capital. For example, companies in a highly profitable industry may invest some of the profits in bonds issued by companies in an expanding industry. The capital transfer may also involve an intermediary. One company can deposit excess capital in a bank, which can then lend it out to another company demanding capital.

Financial markets are not only facilitating the transfer of resources, but also ensuring that the capital flows to the most efficient projects. Suppose company A has plenty of capital for investments, but it can only make a 2% rate of return on an additional investment in its own production. On the other hand, company B can obtain a 10% rate of

¹⁵See <https://www.sifma.org/resources/research/who-owns-stocks-in-america-an-update/> and <https://www.thebalance.com/who-owns-the-u-s-national-debt-3306124>, accessed on July 1, 2022.

return on expanding its production capacity, but lacks the capital to initiate the expansion. If company B borrows the required capital from company A at a 5% interest rate, both parties are satisfied, and the more efficient productive investment can be implemented. Alternatively, company A could pay out its excess capital to its owners, who could then invest directly in company B.

The financial system also allows for the transfer of risk between participants and thus facilitates **risk management**. By issuing shares and thus an ownership stake of the company, an entrepreneur obtains capital required for investments, but also shares the risk with the investors. Any future profits and losses are split between the entrepreneur and the investors according to their ownership share. Although the entrepreneur believes she has a great business case, she might not want to take all the risk, even if she had the necessary capital herself.

Long-term investors keep track of their exposure to various forms of risks, such as their exposure to specific industries (through their ownership of stocks and bonds issued by companies in different industries), to macroeconomic factors, to interest rates and foreign exchange rates, etc. They can manage their risks by moving capital from some financial assets to others, for example from stocks to bonds, from one country to another country, or from stocks in one industry to stocks in another industry. Pension funds facing predictable future pension payouts to its members can reduce—maybe completely eliminate—the risk that they cannot meet the promised payments by investing in a well-chosen portfolio of bonds.

Derivative securities are ideal for managing the risk of adverse movements in certain quantities, such as key interest rates, exchange rates, or prices of commodities, stocks, or bonds. An investor exposed to such a risk can reduce or eliminate the risk by taking an appropriate position in a carefully selected derivative security. The investor can effectively transfer resources between different states of the world. After significant growth and innovation in the markets for derivative securities, investor can now manage their exposure to a long list of risk factors through derivatives. Some risks are difficult to “securitize”—for example, risks related to the labor income or health status of a given person—but often insurance companies can then provide protection through tailored insurance contracts.

Of course, financial markets also enable investors to take risk. Investors are generally willing to accept some risks, as long as they are compensated in terms of higher expected returns. Although stock investments are generally riskier than bond investments, most investors participate in the stock market because average returns are larger for stocks than for bonds. Two key themes of this book are exactly how to measure the risk of an investment and by how much such risks are compensated in financial markets.

When selecting their risky investments, some investors do not want to rely only on the better average performance of risky assets relative to riskfree assets. Instead, they invest in specific stocks or other assets they believe will deliver larger-than-average returns. This belief can be due to a fundamental analysis of the asset and issuer, which indicates that the asset is currently under-priced. But, sometimes investors are simply making “bets” based on some gut feeling. Investors take on a lot of risk if they concentrate their investments in a few stocks, if they invest heavily in derivative securities, or if they use leverage to magnify their investments.

Financial markets provide opportunities for investors to **pool resources** by investing in conjunction through funds rather than investing individually. The fund or investment company collects capital from many individual investors and invests the aggregate capital in a certain pool of assets. If an investor with little capital had to invest directly in individual stocks, she could only buy few shares of few companies. By investing the same

amount in a stock market fund, she can obtain (small fractions of) shares in a large number of companies, and such a diversification across stocks is generally recommended as we shall discuss extensively in this book. Furthermore, fund investments reduce transaction costs. The transaction costs of the fund making relatively large trades on behalf of all its members are smaller than the sum of the transaction costs the investors would have to pay if they did their trades individually. In addition, many individual investors appreciate that the fund handles the paperwork, tax reporting, etc., that financial market trading generates.

Financial markets also **provide information** of use when taking certain decisions. The market price of an asset is set to equate demand and supply, and thus reflects an average valuation of the asset among the potential buyers and sellers of the asset. If you need to set a value of an asset for a certain transaction (e.g., a private trade of the asset), the market price of the asset would be a fair benchmark value. Similarly, interest rates set in financial markets are used as fair benchmarks for interest rates in off-market loan agreements.

Firms may use financial market prices for capital budgeting decisions. If a firm plans to enter a new industry, the stock prices of companies already active in that industry may provide useful information for valuing the project. For investment projects involving a commodity, the prices of futures written on that commodity are relevant inputs to the project valuation. For example, if a firm considers drilling for oil at a certain location, it should certainly consider the futures prices for oil.

Financial market prices can also help alleviating potential conflicts of interest. The manager of a firm may have different incentives than its owners (the shareholders). Maybe the manager would like to spend company resources on a private jet or other perks, prefer to build a corporate empire he can rule over instead of running the company at its most profitable size, or maybe he discards risky, but potentially highly profitable, investment projects to reduce the probability of having to report losses in future financial statements. The shareholders may find it difficult (or very expensive) to monitor all the actions of the manager and to judge whether he is making decisions in their interest. This principal-agent problem can be mitigated by tying the compensation of the manager to the market-determined price of the company's stocks, for example by granting the manager company stocks or call options written on the company stocks.

Finally, various quantities and values fixed in financial markets are used by governments, central banks, and financial supervisory authorities. Financial markets are closely linked to the macroeconomy. Significant changes in stock prices or interest rates may reflect that investors have revised their expectations about future macroeconomic conditions, which might induce governments or central banks to intervene in various ways to adjust those expectations.

1.8 An investment primer

What does it mean to invest in a financial asset? How do you do it in practice? Here is a short recipe:

1. Buy financial assets, either by investing on your own through a broker (maybe an online broker) or by investing together with others through an investment company such as a mutual fund.
2. Enjoy the dividends you receive from the assets while you own them.
3. Either keep the assets until their maturity (if the assets have a finite maturity date) or resell the assets before maturity, hopefully with profits.
4. Maybe obtain professional advice or even transfer investment decisions partly to a

portfolio manager.

Figure 1.1 illustrates the process when investing through an intermediary. Whether you make your own investment decisions or leave them to a portfolio manager, the main questions are the same: which assets should you invest in at a given point in time? When should you buy and when should you sell each asset?

Professional investors and portfolio managers often start by making a *strategic asset allocation* decision, which means that they decide on the relative investments in the different major asset classes. The asset classes could for example be domestic stocks, foreign stocks, domestic government bonds, foreign government bonds, corporate bonds, real estate, and commodities. The strategic asset allocation decision is typically based on the historical long-term risk-return characteristics of the different asset classes, but maybe influenced by the investor's own expectations about their future performance. The relative weights of the different asset classes are only infrequently adjusted.

The next step is then, within each asset class, to decide on exactly which assets to invest in. For example, you may have decided to invest 25% of your wealth in U.S. stocks, but which of the thousands of stocks do you want to invest in? This is the *security selection* decision. In particular when it comes to stocks, many investors spend lots of time and effort on trying to find mispriced stocks, i.e., to identify the future winners and losers in the stock market. As we shall discuss in later chapters, this is not an easy task. If it was easy, a lot of investors would simply buy the obviously underpriced stocks, and their prices would immediately increase to a point where the stock is no longer underpriced.

Some investors try to time the markets by moving large amounts between major asset classes, sometimes quite frequently. Such market timing activities are known as *tactical asset allocation*. If you can predict that the overall stock market is likely to take a substantial downturn, you should reduce or completely eliminate your stock holdings and maybe move your money to government bonds or other assets that are not going down with the stock market. However, as we shall discuss in later chapters, it is extremely difficult to predict where any financial market is going.

Based on a sample of large U.S. pension funds, Brinson, Hood, and Beebower (1986) and Brinson, Singer, and Beebower (1991) conclude that the strategic asset allocation decision explains more than 90% of the time series variation in quarterly returns. Consequently, market timing and security selection are far less important for the returns you will realize than the overall allocation across asset classes. Yet most investors seem to spend most of their time on security selection.

A key goal of this book is to determine the best investments for a given investor. Here is a preview of some of the most important findings:

1. Time value of money: \$1 today is worth more than \$1 tomorrow (typically).
2. Risk-return staircase: No pain, no gain.
3. Diversification: Don't put all your eggs in one basket. Avoid risks without extra returns.
4. Golden rule of arbitrage: Buy low, sell high – if possible.
5. Markets are quite efficient in the sense that it is very difficult to detect mispriced assets.
6. Take a life-cycle, “holistic” perspective. When or in what situations do you need the money? What are your other major assets and risks? What is your position in real estate? What about the magnitude and the risk characteristics of your human capital?

The idea of investment and diversification is, in fact, an ancient advice. For example, the following quote is taken from the Old Testament (Ecclesiastes, Chapter 11, verse 1-2):

Cast your bread upon the waters, for after many days you will find it again.

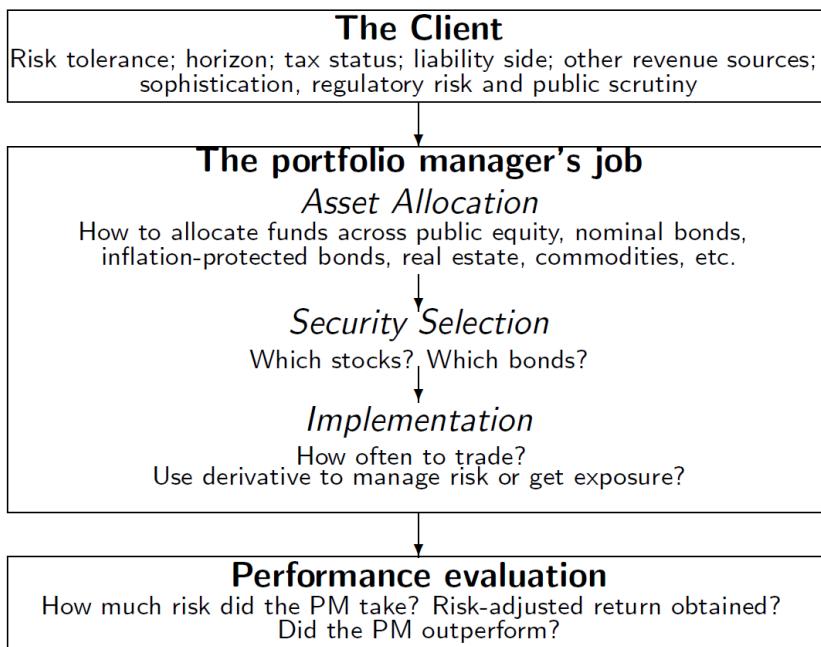


Figure 1.1: The investment process.

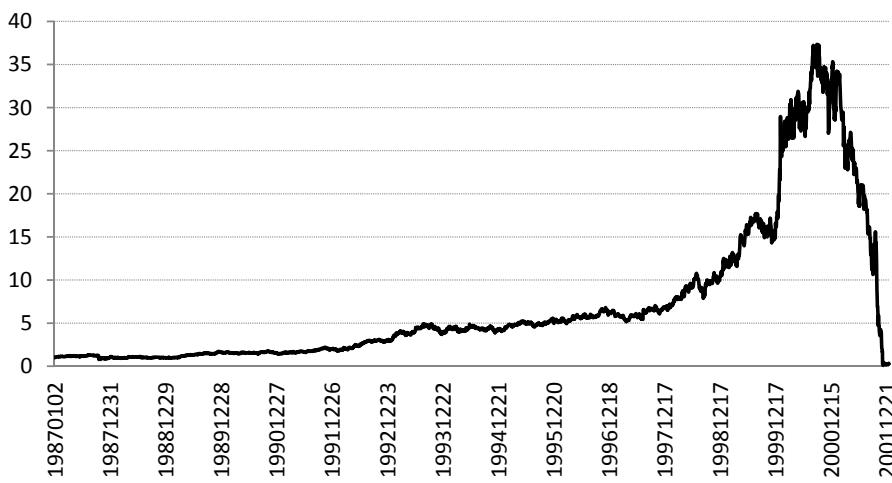


Figure 1.2: The market capitalization of Enron.

The graph shows the market value of all outstanding stocks of Enron Corporation each trading day from January 2, 1987 to January 10, 2002. The market values are normalized to one at January 2, 1987, where the actual market capitalization was around 1.8 bn USD. The data were downloaded on October 4, 2017 from CRSP, the Center for Research in Security Prices at the University of Chicago Booth School of Business.

Give portions to seven, yes to eight, for you do not know what disaster may come upon the land.

The default of the U.S. Energy company Enron in December 2001 is a painfully clear example of the importance of diversification and of looking both at your financial investments and your other assets. In the U.S., many individuals save for retirement in a fund set up—and, to some extent, managed—by the employer. In the case of Enron, 62% of the assets of Enron's employee pension fund were invested in company stock at the end of 2000. For many years, such an investment delivered a nice return, cf. Figure 1.2. But from January 2001 to January 2002, the value of Enron stocks dropped by 99%! Hence, the default of Enron implied that the employees both lost their jobs *and* a large share of their retirement savings. Diversification is one of the key themes of modern investment theory, and we shall return to that theme in several chapters in the book.

1.9 Exercises

No exercises in this chapter so far!

CHAPTER 2

Returns

One of the most basic and important concepts in finance is the return. The return on an investment is a measure of its profitability. The investor in a financial asset typically makes a cash payment when purchasing the asset. When the investor later sells the asset, the difference between the sell price and the purchase price is a profit to the investor (which, of course, can be negative). In addition, the investor might have received dividends from the asset during the holding period, and this cash flow should also be included in the return calculation.

This chapter explains how to calculate returns. Section 2.1 considers the return over a given period. Section 2.2 describes how to calculate returns over multiple periods by compounding the periodic returns. Based on the same idea, Section 2.3 shows how returns are annualized. Section 2.4 explains how to calculate the return in a case where the investment position is changing over the period considered. Section 2.5 introduces the concept of excess returns. Section 2.6 explains the important distinction between nominal and real returns. Sections 2.7 and 2.8 demonstrate how to calculate returns on leveraged investments where part of the invested funds are borrowed, and on short positions where the investor sells an asset instead of purchasing it. Section 2.9 shows how the return on a portfolio of different assets can be calculated from the returns on each of the assets in the portfolio. Finally, the concept of log-returns is introduced in Section 2.10.

2.1 Returns over a given period of time

A return refers to the gains from holding an asset (or a portfolio of assets) over a given time period. The gains can consist of direct payments and a capital gain. Any direct payment is paid to the holder of the asset at some point in time during the period. Think of a dividend payment to the holder of a stock or a coupon payment (i.e., interest payment) to the holder of a bond. In the following, the term “dividend” is used to represent the direct payment, even though the asset might not be a stock. The capital gain is the increase in the price of the asset over the period. The holder of the asset can decide to realize the capital gain by selling the asset at the end of the period. Whether the holder realizes the gain or not, we compute the return over the period in the same way.¹

¹We ignore taxes in the presentation. Sometimes investors have to pay taxes on their financial returns, and in some cases the tax to be paid depends on whether the return is realized or not.

Suppose we want to compute the return on an asset over a time period of length Δt years.² Let t denote the time of the beginning of this period. Hence, the period ends at time $t + \Delta t$. Let P_t and $P_{t+\Delta t}$ denote the prices of the asset at the beginning and the end of the period, respectively. Let $D_{t+\Delta t}$ denote the sum of any cash dividends to the holder of the asset over this period; of course, $D_{t+\Delta t} = 0$ if there are no dividend payments in the period. The **rate of return** over the period is then defined as³

$$r_{t,t+\Delta t} = \frac{D_{t+\Delta t} + P_{t+\Delta t} - P_t}{P_t} = \frac{D_{t+\Delta t} + P_{t+\Delta t}}{P_t} - 1. \quad (2.1)$$

Note that the return is not annualized; we will discuss how to do that below. We can easily rewrite the rate of return as

$$r_{t,t+\Delta t} = \frac{D_{t+\Delta t}}{P_t} + \frac{P_{t+\Delta t} - P_t}{P_t}, \quad (2.2)$$

where the first ratio shows the dividends in percent of the price at the beginning of the period—often referred to as the *dividend yield*—and the second ratio is the capital gain in percent. Also observe that (2.1) implies that

$$P_t = (1 + r_{t,t+\Delta t})^{-1} (D_{t+\Delta t} + P_{t+\Delta t}), \quad (2.3)$$

so that discounting the end-of-period value $D_{t+\Delta t} + P_{t+\Delta t}$ using the rate of return as the discount rate results in the beginning-of-period value P_t .

We shall also work with the **gross return** which is defined as the value of the asset at the end of the period, including any dividends received during the period, divided by the price at the beginning of the period:

$$R_{t,t+\Delta t} = \frac{D_{t+\Delta t} + P_{t+\Delta t}}{P_t}. \quad (2.4)$$

The gross return shows how much you could get at the end of the period per dollar invested in the asset at the beginning of the period. Note that the rate of return and the gross return are related via

$$R_{t,t+\Delta t} = 1 + r_{t,t+\Delta t}. \quad (2.5)$$

A word on notation: When it is clear for which period a return is computed, the time subscripts are often omitted so that, for example, we write r instead of $r_{t,t+\Delta t}$.

It is a firm principle in finance not to directly compare payments at different dates. Most investors would prefer getting a dollar today rather than getting a dollar in a year from now. In order to compare payments at different dates we should discount them back or forward to the same date using appropriate discount rates. Applying this principle to the definition (2.1) of the rate of return, it is clear that the return computation is most appropriate if the dividends are paid late in the period between time t and $t + \Delta t$, ideally just before time $t + \Delta t$. If this is not the case, we could think about how to discount the dividends forward to time $t + \Delta t$, but it is not at all clear what the appropriate discount rate is since it should reflect the risk of the payment.

A more robust procedure is to assume that when a dividend is received, it is immediately reinvested by buying additional units of the same asset. Say a dividend payment allows

² Δ is the upper-case version of the Greek letter “delta”. As most other quantitative textbooks in finance, this book applies various Greek letters. You can find a table with the Greek alphabet in Appendix B.

³The rate of return is sometimes called the *percentage return* or the *net rate of return*.

you to buy x additional shares of the asset per share you had already. Then your position is effectively magnified by the *position multiplier* $1 + x$.

Example 2.1

The stocks of Microsoft Corporation are listed on the Nasdaq stock exchange with the ticker symbol MSFT. Suppose we want to calculate the return on Microsoft stocks in February 2023. According to <http://yahoo.finance.com>, the closing price on February 28 was \$249.42 per share and the closing price on January 31 was \$247.81 per share. The capital gain was therefore \$1.61 per share or 0.650% of the end-of-January price.

In addition, Microsoft paid a dividend of \$0.68 per share in February. If we disregard the exact payment date of the dividend and plug into (2.1), we obtain a rate of return of

$$r = \frac{\$0.68 + \$249.42 - \$247.81}{\$247.81} \approx 0.00924 = 0.924\%.$$

The dividend yield constitutes $\$0.68/\$247.81 \approx 0.274\%$. The rate of return is indeed the sum of the dividend yield and the capital gain: $0.274\% + 0.650\% = 0.924\%$. In this computation, we assume that the dividend received during the month is held in cash until the end of the month.

What happens if we reinvest the dividend in the same stock when we receive it? Investors owning shares on February 14 were entitled to the dividend payment, whereas investors buying shares on February 15 or later were not. In this case, February 15 is referred to as the *ex-dividend date*. The actual dividend was not paid out until March 8, but *for return calculations the dividend is assumed to be paid on the ex-dividend date*. The closing price on February 14 is a *cum-dividend* price as it represents the price of a claim to all future dividends from Microsoft *including* that on February 14. Starting on the morning of February 15, the listed stock price of Microsoft is an *ex-dividend* price as it represents the price of a claim to all future dividends from Microsoft *excluding* that on February 14.

The closing price on February 14 of \$272.17 includes the dividend so the corresponding ex-dividend price is $\$272.17 - \$0.68 = \$271.49$. With the dividend we can—in theory—purchase $\$0.68/\$271.49 \approx 0.002505$ additional shares of Microsoft stock. This corresponds to a position multiplier of 1.002505. Starting the month with one share and reinvesting the dividend when received, we end the month holding 1.002505 shares with a total value of $1.002505 \times \$249.42 \approx \250.0447 . Hence, the rate of return over the month is

$$r = \frac{\$250.0447 - \$247.81}{\$247.81} \approx 0.00902 = 0.902\%,$$

which is close to the 0.924% calculated ignoring the timing of the dividend. This is generally true when dividends are small, when the period considered is short, and when the price of the stock does not fluctuate wildly around the dividend payment date.

Alternatively, we can calculate the *inverted position multiplier* $1/1.002505 \approx 0.997502$, which is how many shares of stocks you would need to buy at the beginning of the month to end up with one share at the end of the month if you reinvest the dividends when received. Before the ex-dividend date we can adjust the stock price for the later dividend by multiplying the true price with the inverted position multiplier. This gives an adjusted stock price. For example, the adjusted end-of-February price is $0.997502 \times \$247.81 \approx \247.1909 . This is the amount you would need to have invested at the end of January in order to end February with a value identical to a single Microsoft share, i.e., \$249.42, with

no intermediate payments. The rate of return over February can then be computed as

$$r = \frac{\$249.42 - \$247.1909}{\$247.1909} \approx 0.00902 = 0.902\%,$$

identical to the number computed above.

In general, if we let D denote the dividend per share and let P denote the closing price on the last cum-dividend date, the dividend allows you to purchase $D/(P - D)$ additional shares, so the position multiplier is

$$1 + \frac{D}{P - D} = \frac{P - D}{P - D} + \frac{D}{P - D} = \frac{P}{P - D},$$

and the inverted position multiplier is

$$\frac{1}{\frac{P}{P-D}} = \frac{P - D}{P} = 1 - \frac{D}{P}.$$

The above approach to return calculations extends to cases with multiple dividends. This is explained in the next example.

Example 2.2

This example is a continuation of Example 2.1. Suppose we want to compute the return on Microsoft stocks over the entire year 2023. The stock price was \$239.82 at the end of December 2022 and \$376.04 at the end of December 2023, an increase of 56.801%. However, we should include the dividends.

In 2023 Microsoft made four dividend payments as shown in Table 2.1. For each dividend payment, we calculate an associated position multiplier as in Example 2.1.

Next, we can calculate a *compound multiplier* following each dividend payment. The compound multiplier is indexed at 1 at a given starting date, which in our case is taken to be the end of December 2022. The compound multiplier at any later date shows the number of stocks you would have at that date if you started owning one stock and then reinvested all intermediate dividends by purchasing additional units of the stock. At the first dividend payment on February 14, it jumps to 1.002505, as this is the position multiplier corresponding to this dividend. The compound multiplier stays at 1.002505 until May 16, where the next dividend is paid with an associated position multiplier of 1.002186. Immediately after this payment, the compound multiplier is $1.002505 \times 1.002186 = 1.004696$ since this is the number of stocks you will have if you started with one stock in early January and reinvested both the February dividend and the May dividend when received. Continuing these calculations, the compound multiplier at the end of 2023 is 1.008867. The stock price at the end of December 2023 was \$376.04, so the return over the full year can be calculated as

$$\frac{1.008867 \times \$376.04 - \$239.82}{\$239.82} \approx 0.58191 = 58.191\%.$$

Basically we are adjusting upwards later prices for intermediate dividends by multiplying the observed price per share on a given date by the compound multiplier up to that date.

Last day with dividend	Dividend	Closing price cum-dividend	Closing price ex-dividend	Position multiplier	Compound multiplier	Inverted multiplier
Feb 14, 2023	0.68	272.17	271.49	1.002505	1.002505	0.997502
May 16, 2023	0.68	311.74	311.06	1.002186	1.004696	0.995326
Aug 15, 2023	0.68	321.86	321.18	1.002117	1.006823	0.993223
Nov 14, 2023	0.75	370.27	369.52	1.002030	1.008867	0.991211

Table 2.1: Microsoft dividends.

The table lists the dividend payments and associated position multipliers for Microsoft stocks in 2023. The value of the compound multiplier shown is valid immediately after the given dividend payment date, whereas the value of the inverted multiplier shown is valid immediately before the dividend payment date. The data on dividends and closing prices were obtained from <http://finance.yahoo.com>.

An alternative procedure is to adjust past prices downwards. We could end up with one share at the end of December 2023 if we had purchased $1/1.008867 \approx 0.991211$ shares at the end of December 2022, which would have cost $0.991211 \times \$239.82 \approx \237.7122 . More generally, holding the terminal date fixed, we can calculate an *inverted multiplier* for any earlier date as the number of stocks you would have had to purchase at that date in order to end up with one stock at the terminal date, assuming that intermediate dividends are immediately reinvested. Based on the (downwards) adjusted closing price at the end of December 2022, we can compute the return over 2023 as

$$\frac{\$376.04 - \$237.7122}{\$237.7122} \approx 0.58191 = 58.191\%.$$

As you can see, there are several paths to the same result.

Academic research on stock returns often applies data from the CRSP (often pronounced ‘crisp’) database maintained by the Center for Research in Security Prices at the University of Chicago’s Booth School of Business. Researchers and students at many business schools and universities have access to CRSP. If you ask CRSP for daily returns on a given stock, the returns are adjusted for dividends as explained in the above examples. However, if you ask CRSP for monthly returns, the return is calculated under the assumption that the dividend is not reinvested until the last day of the month. This will typically lead to returns that are slightly different from the case where the dividend is reinvested at the ex-dividend date.

Some internet resources offer free access to historical prices and dividends on individual stocks as well as leading stock indices in many countries. One provider is **Yahoo Finance** that offers lots of information on their homepage <http://finance.yahoo.com> on exchange-listed stocks both in the United States and many other countries, as well as information on bonds, foreign exchange, and derivative securities. Among other things, you can retrieve historical prices on individual stocks with a daily, weekly, or monthly frequency over a desired period of time. You have to pay attention to some conventions, though. First, if you ask Yahoo for monthly prices, it will label the months by the first trading day of the month. For example, for a row labeled Feb 1, 2023 by Yahoo, the opening price is that of the first trading day in that month, which is February 1. The closing price in the row is the closing price on the *last* trading day in the period starting on the

date shown in the row. Since you asked for monthly prices, this would be the last trading day of February 2023, i.e., February 28. Second, Yahoo provides adjusted closing prices which, in principle, are calculated as in the above examples using the inverted multiplier with the terminal date being equal to the day you look up the prices on Yahoo. However, at the time of writing this text (April 15, 2024), there seems to be a mismatch. If you ask Yahoo for monthly prices, the adjusted closing price shown for months with dividend payments are apparently wrong. Consider, for example, the price data shown for Microsoft when accessing Yahoo on April 15, 2024. The adjusted closing price shown for January 2023 is \$245.1786 whether you ask for daily or monthly prices. For the month of February 2023, the adjusted closing price shown is \$246.7715 but, if instead you ask Yahoo for daily prices, the adjusted closing price for the last day of February is \$247.3896. You get a rate of return of 0.650% if you apply the former adjusted closing price and 0.902% if you apply the latter. The 0.902% is the correct return as explained in Example 2.1. This implies that even if you want to calculate monthly returns, you would need to download daily prices to obtain the correct adjusted closing price at the end of a month with a dividend payment. Note that Yahoo has changed its conventions several times, so you need to check the current conventions when downloading data from its homepage. Other data providers may follow different conventions. Also note that past data is sometimes revised which may affect return calculations.

In the examples we have assumed that dividends can be reinvested by purchasing fractions of shares. This is not easy in real life. Of course, while we assumed an initial position of one share, many institutional or large individual investors own thousands of shares of the same stock. Even with initial holdings of 1,000 shares, some rounding of numbers is necessary when reinvesting the dividends. In the example, the dividend payment in February 2023 could buy you 0.002505 additional shares per share owned, which would then mean 2.505 shares in addition to the 1,000 shares owned already. In practice, this would have to be rounded to 2 (with some cash left which could be deposited on your bank account) or 3 (with some additional money invested, maybe withdrawn from your bank account). Such rounding errors should not matter much for the big picture, so these concerns are ignored when computing returns.

2.2 The compounding of returns

Suppose that we know the returns on an asset in each of n consecutive periods of the same length Δt . What is the return over all n periods? Here we have to take the *compounding of returns* into account.

Suppose you invest \$100 in Apple stocks at the beginning of the year. Say, the rate of return on Apple stocks in January is 10%. Then your stocks are worth \$110 at the end of January. Now suppose the rate of return on Apple stocks in February is 20%. Then you earn the 20% return on the \$110, that is both on your original \$100 investment and on the \$10 return you received in January. In total, your stocks are worth $\$110 \times 1.20 = \132 at the end of February. Compared to your initial \$100 investment, this gives a rate of return of 32% over the two months, which is different from the sum of the rates of return over each month, i.e. $10\% + 20\% = 30\%$. The additional 2% return is exactly the 20% return in February on the 10% return in January.

The next theorem shows how rates of returns and gross returns are compounded.

Theorem 2.1

Let $\Delta t > 0$ denote the period length. For each $i = 0, 1, \dots, n - 1$, let $r_{t+i\Delta t, t+(i+1)\Delta t}$ denote the rate of return and $R_{t+i\Delta t, t+(i+1)\Delta t}$ the gross return over the period between time $t + i\Delta t$ and time $t + (i + 1)\Delta t$. Then, over the entire n -period time interval from t to $t + n\Delta t$, the rate of return $r_{t, t+n\Delta t}$ and the gross return $R_{t, t+n\Delta t}$ are given by

$$r_{t, t+n\Delta t} = (1 + r_{t, t+\Delta t}) (1 + r_{t+\Delta t, t+2\Delta t}) \dots (1 + r_{t+(n-1)\Delta t, t+n\Delta t}) - 1, \quad (2.6)$$

$$R_{t, t+n\Delta t} = R_{t, t+\Delta t} \times R_{t+\Delta t, t+2\Delta t} \times \dots \times R_{t+(n-1)\Delta t, t+n\Delta t}. \quad (2.7)$$

Proof

As in the example before the theorem, think of moving your investment forward period by period. Start by investing one dollar at time t . Then you will have $R_{t, t+\Delta t}$ dollars at time $t + \Delta t$. If you invest that in the asset at time $t + \Delta t$, you will have $R_{t, t+\Delta t} \times R_{t+\Delta t, t+2\Delta t}$ at time $t + 2\Delta t$. Continuing this way, you will end up with Eq. (2.7). Then (2.6) follows from (2.7) since $R_{t+i\Delta t, t+(i+1)\Delta t} = 1 + r_{t+i\Delta t, t+(i+1)\Delta t}$.

It is important to realize that to obtain the multi-period returns stated in the above formulas, we have to reinvest the dividends along the way by purchasing additional units of the asset. This works as in the one-period case discussed in Example 2.1. Every dividend payment leads to a multiplication of the number of shares we have by a position multiplier, which equals one plus the ratio of the dividend received on the existing position to the ex-dividend price of the stock.

Example 2.3

In Example 2.2 we calculated the rate of return on Microsoft stocks in 2023. This one-year return can alternatively be calculated by compounding the monthly returns shown in Table 2.2. In each month we can calculate the rate of return as the product of the position multiplier and the closing price of the end of the month divided by the closing price at the end of the preceding month and then subtract one. We get the same result by calculating the increase in the adjusted closing price from the preceding month, and here we can use either an adjusted closing price derived as the product of the actual closing price and the inverted multiplier or an adjusted closing price downloaded from Yahoo Finance. For August 2023, for example, the rate of return is

$$\frac{1.002117 \times 327.76}{335.92} - 1 = \frac{327.0961}{334.5313} - 1 = \frac{326.4924}{333.9138} - 1 \approx -0.02223 = -2.23\%.$$

Next, we can compound the monthly returns to get the full-year return by applying Eq. (2.6):

$$(1 + 0.03332) \times (1 + 0.00902) \times \dots \times (1 - 0.00757) - 1 \approx 0.58191 = 58.191\%,$$

as also found in Example 2.2.

Month	Closing price	Position multiplier	Rate of return	Compound multiplier	Inverted multiplier	Adj close (my calc)	Adj Close (Yahoo)
Dec 2022	239.82			1.000000	0.991211	237.7122	237.2734
Jan 2023	247.81	1.000000	3.332%	1.000000	0.991211	245.6320	245.1786
Feb 2023	249.42	1.002505	0.902%	1.002505	0.993694	247.8471	247.3896
Mar 2023	288.30	1.000000	15.588%	1.002505	0.993694	286.4819	285.9531
Apr 2023	307.26	1.000000	6.576%	1.002505	0.993694	305.3223	304.7587
May 2023	328.39	1.002186	7.111%	1.004696	0.995866	327.0325	326.4288
Jun 2023	340.54	1.000000	3.700%	1.004696	0.995866	339.1322	338.5062
Jul 2023	335.92	1.000000	-1.357%	1.004696	0.995866	334.5313	333.9138
Aug 2023	327.76	1.002117	-2.223%	1.006823	0.997974	327.0961	326.4924
Sep 2023	315.75	1.000000	-3.664%	1.006823	0.997974	315.1104	314.5288
Oct 2023	338.11	1.000000	7.082%	1.006823	0.997974	337.4251	336.8023
Nov 2023	378.91	1.002030	12.295%	1.008867	1.000000	378.9100	378.2106
Dec 2023	376.04	1.000000	-0.757%	1.008867	1.000000	376.0400	375.3459

Table 2.2: Microsoft returns.

The table shows the monthly prices, position multiplier, and returns on Microsoft stocks from the end of December 2022 to the end of December 2023. The position multiplier for a given month shows the number of shares of the stock you hold at the end of the month if you started the month with one share and reinvested any dividends received immediately by purchasing additional shares. The compound multiplier gives the number of shares you hold at the end of the month if you started in January 2023 with one share and reinvested any dividends immediately. The inverted multiplier shows the number of shares you would have had to hold at the end of the month in order to end December 2023 with one share, provided that all dividends are reinvested when received. The second column from the right shows an adjusted closing price calculated as the product of the actual closing price at the end of that month and the inverted multiplier. The right-most column shows adjusted closing prices downloaded from Yahoo Finance on April 15, 2024 (asking for daily data, cf. the discussion about conventions in the text).

With the approach explained above we can easily control for the lump-sum dividends that are more or less regularly paid out for any stock or any bond. If the asset or portfolio in question pays dividends very frequently, it can be reasonable to approximate the dividend payments by a dividend stream that arrives continuously in time. This is relevant, for example, for a large portfolio of stocks if the individual stocks pay dividends at many different dates so that at more or less any date there will be a dividend payment from at least one of the stocks in the portfolio. Also, interest payments on some deposits or loans are added very frequently to the balance of the account, which may be approximated by a continuous-time stream. A continuous-time dividend stream is sometimes easier to work with in various computations.

How do we compute returns with a continuous-time dividend stream? Consider an asset that pays a continuous dividend yield δ_t at any time t . This means that the total dividend payment over a tiny interval $[t, t + \Delta t]$ is approximately $\delta_t \times P_t \times \Delta t$. If we buy one unit of the asset at time t and keep reinvesting the continuous dividends by purchasing extra (fractions of) units of the assets, how many units will we end up with at time $T > t$? The

answer turns out to be⁴

$$A_T = \exp \left\{ \int_t^T \delta_u du \right\}.$$

The gross return over the interval $[t, T]$ is therefore

$$R_{t,T} = \exp \left\{ \int_t^T \delta_u du \right\} \frac{P_T}{P_t}, \quad (2.8)$$

and the corresponding rate of return is

$$r_{t,T} = \exp \left\{ \int_t^T \delta_u du \right\} \frac{P_T}{P_t} - 1 = \frac{\exp \left\{ \int_t^T \delta_u du \right\} P_T - P_t}{P_t}.$$

For simplicity, the dividend yield is often assumed constant over time so that $\delta_u = \delta$ for all u , and then we have $\int_t^T \delta_u du = \int_t^T \delta du = \delta(T - t)$ and thus

$$R_{t,T} = e^{\delta(T-t)} \frac{P_T}{P_t}, \quad r_{t,T} = \frac{e^{\delta(T-t)} P_T - P_t}{P_t}. \quad (2.9)$$

2.3 Annualizing returns

Returns are often expressed on an annual basis as some percent per year. It is difficult to compare two returns if they are computed over periods of different lengths. By converting each of them into an annualized return, we can make a fair comparison. If we observe a certain monthly rate of return, the corresponding annualized rate of return is the rate of return we would get if that monthly rate of return would repeat itself 12 months in a row. Here, the compounding of the returns must be accounted for. The compounded annualized return is also known as the *effective annual return*. Theorem 2.1 implies that a monthly rate of return r_{mon} and a gross return R_{mon} are annualized like this:

$$r_{\text{ann}} = (1 + r_{\text{mon}})^{12} - 1, \quad R_{\text{ann}} = (R_{\text{mon}})^{12}. \quad (2.10)$$

In a similar way, we can annualize returns over any other fraction of a year, say a month, week, or quarter. More generally, the rate of return $r_{t,t+\Delta t}$ over a period of length Δt =

⁴Here is a proof. Divide the interval $[t, T]$ into M bits of length Δt so that dividends are paid and additional units bought at times $t + \Delta t, t + 2\Delta t, \dots, t + M\Delta t = T$. Now we proceed as in the discrete-time case. Let $A_{t+m\Delta t}$ denote the number of units of the asset we have immediately after time $t + m\Delta t$. We start with $A_t = 1$ unit. At time $t + \Delta t$ we receive a dividend of $\delta_t P_{t+\Delta t} \Delta t$ which we spend on buying $\delta_t \Delta t$ extra assets, bringing our holdings up to $A_{t+\Delta t} = 1 + \delta_t \Delta t$. At time $t + 2\Delta t$ we receive a total dividend of $A_{t+\Delta t} \delta_{t+\Delta t} P_{t+2\Delta t} \Delta t$ which will buy us $A_{t+\Delta t} \delta_{t+\Delta t} \Delta t$ extra units. Our total is now $A_{t+2\Delta t} = A_{t+\Delta t} + A_{t+\Delta t} \delta_{t+\Delta t} \Delta t$. Continuing like this we find that, at any time $s = t + m\Delta t$ for some integer $m < M$, our total holdings immediately after time $s + \Delta t$ are given by $A_{s+\Delta t} = A_s + A_s \delta_s \Delta t$ so that

$$\frac{A_{s+\Delta t} - A_s}{\Delta t} = \delta_s A_s.$$

Letting $\Delta t \rightarrow 0$, the left-hand side will approach the derivative A'_s , and we see that A_s must satisfy the differential equation $A'_s = \delta_s A_s$ as well as the initial condition $A_t = 1$. The solution is

$$A_s = \exp \left\{ \int_t^s \delta_u du \right\}.$$

$1/N$ corresponds to a compounded annualized rate of return of

$$r_{\text{ann}} = (1 + r_{t,t+\Delta t})^{1/\Delta t} - 1 = (1 + r_{t,t+\Delta t})^N - 1 \quad (2.11)$$

since a year is divided into $N = 1/\Delta t$ periods.

Sometimes, an annualized rate of return is calculated without compounding. We refer to this a simple annualized rate of return. For example, we have

$$r_{\text{ann}}^{\text{simp}} = 12 \times r_{\text{mon}}. \quad (2.12)$$

The simple annualized return should be seen as a quick-and-dirty approximation of the more appropriate compounded annualized return, but as the following example illustrates the approximation error can be significant.

Example 2.4

In Example 2.1 we found a return of 7.01965% on Microsoft over May 2016 under the assumption that the dividend received during the month was held in cash until the end of the month. This corresponds to the simple annualized rate of return

$$r_{\text{ann}}^{\text{simp}} = 12 \times 7.01965\% \approx 84.23587\%$$

or the more appropriate compounded annualized rate of return

$$r_{\text{ann}} = (1.0701965)^{12} - 1 \approx 1.2571612 = 125.71612\%.$$

This illustrates the potentially significant difference between the two ways of annualizing short-term returns.

2.4 Internal rate of return

In the above calculations of returns, we assumed that dividends were reinvested in the same asset. This is an appropriate procedure for calculating the return of a given asset or a fixed portfolio of assets. Suppose, however, that we want to calculate the return of an investor over some period of time. During that period, the investor may receive some dividends that she chooses not to reinvest and she may choose to invest additional money by purchasing new assets or to withdraw money by selling some of the assets in her portfolio. The actions of the investor may thus lead to various intermediate cash flows, some positive and some negative, during the period of interest. In this case we can measure the return of the investor by the internal rate of return of the cash flow, which in this context is also referred to as the **dollar-weighted rate of return**.

Example 2.5

At the end of June 2015, Anna invests \$4,000 by purchasing 20 stocks in the company Illuminati at \$100 per share and 10 stocks in the company Spectre at \$200 per share. In the third week of each month, Illuminati pays out a dividend of \$1 per share, whereas Spectre does not pay any dividends in the period considered here. At the end of August 2015, Anna purchases an extra 8 Spectre stocks at a price of \$250 per share. At the end

Date	Action	Cash flow
End June 2015	Buy 20 Illuminati @ \$100 and 10 Spectre @ \$200	-\$4,000
End July 2015	Receive dividends of $20 \times \$1$	+\$20
End August 2015	Receive dividends of $20 \times \$1$. Buy 8 Spectre @ \$250	-\$1,980
End September 2015	Receive dividends of $20 \times \$1$	+\$20
End October 2015	Receive dividends of $20 \times \$1$. Sell 5 Illuminati @ \$120	+\$620
End November 2015	Receive dividends of $15 \times \$1$	+\$15
End December 2015	Receive dividends of $15 \times \$1$. Portfolio worth $15 \times \$110 + 18 \times \$225 = \$5,700$	\$5,715

Table 2.3: Anna's investments.

The table shows the transactions and cash flow in Example 2.5.

of October 2015, Anna sells 5 shares of Illuminati stocks at a unit price of \$120. At the end of December 2015, Anna wants to calculate the return on her investment. At that point the share prices of Illuminati and Spectre are \$110 and \$225, respectively. Table 2.3 summarizes the cash flow generated by Anna's trading strategy. The internal rate of return per month is then the value of r satisfying the equation

$$0 = -\$4000 + \frac{\$20}{1+r} - \frac{\$1980}{(1+r)^2} + \frac{\$20}{(1+r)^3} + \frac{\$620}{(1+r)^4} + \frac{\$15}{(1+r)^5} + \frac{\$5715}{(1+r)^6},$$

which is true for $r \approx 1.3038\%$. The compounded annualized internal rate of return is therefore $(1.013038)^{12} - 1 \approx 0.1682 = 16.82\%$.

A problem with the internal rate of return is that it may not be uniquely determined. We are looking for solutions to an equation of the form

$$0 = \sum_{t=0}^T \frac{C_t}{(1+r)^t} = C_0 + \frac{C_1}{1+r} + \frac{C_2}{(1+r)^2} + \cdots + \frac{C_T}{(1+r)^T},$$

where C_t is the net payment to the investor in period t . If you multiply through by $(1+r)^T$, you can see that the equation is equivalent to

$$0 = C_0(1+r)^T + C_1(1+r)^{T-1} + C_2(1+r)^{T-2} + \cdots + C_T,$$

so we are looking for roots of a polynomial of order T . According to Descartes' rule of sign the maximum number of roots in such a polynomial is equal to the number of times the cash flow sequence $C_0, C_1, C_2, \dots, C_T$ changes sign. In the example above the cash flow sequence is $-4000, 20, -1980, 20, 620, 15, 5715$. Since there are three changes of sign, there could be as many as three roots and thus three internal rates of return. In fact, there are two. Besides the root $0.013038=1.3038\%$ mentioned in the example, another root is (approximately) -2.00951 . Obviously, this alternative root makes no economic sense.

2.5 Excess returns

In many situations we would like to compare the return on some asset to the return on a specific benchmark asset or index. For example, we might compare the return on some risky asset to the return we could have obtained on a riskfree asset to see by how much we were compensated for taking a risk. As another example, the manager of a mutual fund investing in only U.S. stocks might want to compare the return on the chosen stock portfolio to the return on the S&P500 stock index to get an indication of whether he was able to beat the market or not. The difference between the return on a given asset or portfolio and the return on a specified benchmark is referred to as an *excess* return.

Of course, we can easily calculate the excess return in a given period, say, from time t to $t+1$. If you obtained a rate of return of $r_{t,t+1}$ and the rate of return on the benchmark was $r_{t,t+1}^b$, the excess return is simply

$$r_{t,t+1}^{\text{ex}} = r_{t,t+1} - r_{t,t+1}^b.$$

This is the return you would get from investing \$1 in a long position in your portfolio and simultaneously taking a short position of \$1 in the benchmark asset or portfolio.

Compounding excess returns is less straightforward. By investing in your portfolio over two consecutive periods, you generate a rate of return of

$$(1 + r_{t,t+1}) (1 + r_{t+1,t+2}) - 1 = r_{t,t+1} + r_{t+1,t+2} + r_{t,t+1} r_{t+1,t+2}.$$

Analogously, the rate of return on the benchmark over two periods is

$$(1 + r_{t,t+1}^b) (1 + r_{t+1,t+2}^b) - 1 = r_{t,t+1}^b + r_{t+1,t+2}^b + r_{t,t+1}^b r_{t+1,t+2}^b.$$

The excess return over the two period is the difference, i.e.,

$$\begin{aligned} & (1 + r_{t,t+1}) (1 + r_{t+1,t+2}) - (1 + r_{t,t+1}^b) (1 + r_{t+1,t+2}^b) \\ &= r_{t,t+1} + r_{t+1,t+2} + r_{t,t+1} r_{t+1,t+2} - (r_{t,t+1}^b + r_{t+1,t+2}^b + r_{t,t+1}^b r_{t+1,t+2}^b) \\ &= r_{t,t+1}^{\text{ex}} + r_{t+1,t+2}^{\text{ex}} + r_{t,t+1} r_{t+1,t+2} - r_{t,t+1}^b r_{t+1,t+2}^b. \end{aligned}$$

It might be tempting to compound the excess return directly and thus calculate the two-period excess return as

$$(1 + r_{t,t+1}^{\text{ex}}) (1 + r_{t+1,t+2}^{\text{ex}}) - 1 = r_{t,t+1}^{\text{ex}} + r_{t+1,t+2}^{\text{ex}} + r_{t,t+1}^{\text{ex}} r_{t+1,t+2}^{\text{ex}},$$

but since

$$r_{t,t+1}^{\text{ex}} r_{t+1,t+2}^{\text{ex}} = (r_{t,t+1} - r_{t,t+1}^b) (r_{t+1,t+2} - r_{t+1,t+2}^b)$$

is generally different from $r_{t,t+1} r_{t+1,t+2} - r_{t,t+1}^b r_{t+1,t+2}^b$, this way of compounding excess returns is incorrect. For example, if in both periods your return is 10% and the benchmark return is 2%, the correctly computed two-period excess return is

$$(1.10)^2 - (1.02)^2 = 0.1696 = 16.96\%,$$

whereas the erroneous calculation would lead to

$$(1.08)^2 - 1 = 0.1664 = 16.64\%.$$

The error is modest, however, as it will be in most cases, except if the periodic returns are large or you compound over many periods.

2.6 Real vs. nominal returns

Economists distinguish between nominal returns and real returns. A nominal return is a measure of the monetary gains, whereas a real return is a measure of the gains in terms of purchasing power. A 10% nominal rate of return on an investment over a given period might seem nice, but if consumer prices increased by 20% in the same period, our purchasing power was reduced so the investment was actually bad for us. If we invested \$1 at the beginning of the period, we would have \$1.10 at the end of the period. However, the consumer goods we could buy for \$1 at the beginning now cost \$1.20. In terms of purchasing power the gross return is

$$\frac{\$1.10}{\$1.20} \approx 0.9167.$$

This is the *real gross return* over the period. The corresponding *real rate of return* is

$$0.9167 - 1 = -0.0833 = -8.33\%.$$

Our purchasing power was reduced by 8.33% in this case.

More generally, suppose that between time t and time $t + \Delta t$ the nominal gross return equals $R_{t,t+\Delta t}$ and the nominal rate of return equals $r_{t,t+\Delta t} = R_{t,t+\Delta t} - 1$. These returns can be computed from the nominal prices and the nominal dividend as shown in Eqs. (2.4) and (2.1), respectively. Suppose the inflation rate over the same period is $I_{t,t+\Delta t}$. The inflation rate is the percentage increase in the consumer price index (CPI), i.e.,

$$I_{t,t+\Delta t} = \frac{\text{CPI}_{t+\Delta t} - \text{CPI}_t}{\text{CPI}_t}. \quad (2.13)$$

Then the real gross return over the period is

$$R_{t,t+\Delta t}^{\text{real}} = \frac{R_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}} \quad (2.14)$$

and the real rate of return is

$$r_{t,t+\Delta t}^{\text{real}} = \frac{r_{t,t+\Delta t} - I_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}}, \quad (2.15)$$

which follows from the calculation

$$\begin{aligned} r_{t,t+\Delta t}^{\text{real}} &= R_{t,t+\Delta t}^{\text{real}} - 1 = \frac{R_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}} - 1 = \frac{R_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}} - \frac{1 + I_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}} \\ &= \frac{R_{t,t+\Delta t} - (1 + I_{t,t+\Delta t})}{1 + I_{t,t+\Delta t}} = \frac{R_{t,t+\Delta t} - 1 - I_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}} = \frac{r_{t,t+\Delta t} - I_{t,t+\Delta t}}{1 + I_{t,t+\Delta t}}. \end{aligned}$$

If the inflation rate $I_{t,t+\Delta t}$ is close to zero, the approximation

$$r_{t,t+\Delta t}^{\text{real}} \approx r_{t,t+\Delta t} - I_{t,t+\Delta t} \quad (2.16)$$

is quite precise, i.e., the real rate of return is approximately equal to the nominal rate of

return minus the inflation rate.

Example 2.6

In 2023, the nominal rate of return on the S&P 500 index was 26.06%, including dividends. The inflation rate in the United States in 2023 was 3.12%. Hence, the real rate of return on the index was

$$\frac{0.2606 - 0.0312}{1.0312} \approx 0.2225 = 22.25\%,$$

while the approximation gives $0.2606 - 0.0312 = 0.2294 = 22.94\%$.

The CPI reflects the price of a particular basket of goods with specific weights on the different goods included. These weights represent an average consumer's mix of goods. If your relative consumption of different goods deviates substantially from this average, the inflation rate relevant for you might differ from the official inflation rate derived from the CPI. Therefore, the true real return of an asset to a given investor can depend on the investor's preferences for different consumption goods. Nevertheless, the real returns typically reported are based on the CPI.

2.7 Returns on levered positions

Sometimes an investor wants to invest more money in a stock than she has on hand. Then she can borrow additional funds to increase her investment. She is *gearing* or *leveraging up* her investment. When the loan and subsequent investment are made via a broker, this is referred to as *buying on margin*. The broker requires an interest rate on the loan.

To compute the rate of return on a levered position, we first introduce some notation. Suppose the original investment takes place at time $t = 0$ and we want to compute the return until some later date denoted $t = 1$. Let E denote the equity (net worth) of the investor, L the loan from the broker to the investor, P the unit price of the stock, and V the total value of the stocks invested in. Each of these variables can be measured at time 0 and at time 1 as will be indicated by a subscript. Furthermore, let D denote the dividend payment per stock over the period and r_{loan} the interest rate on the loan.

The initial investment of the investor at time 0 is E_0 . Adding a loan of L_0 , the total investment is $V_0 = E_0 + L_0$. The number of stocks purchased is thus V_0/P_0 . The percentage margin is E_0/V_0 , which is the fraction of the total investment covered by the investor's own funds.

At the end of the period, the value of the total stock position is

$$V_1 = \frac{V_0}{P_0}(P_1 + D) = \frac{E_0 + L_0}{P_0}(P_1 + D) = (E_0 + L_0)(1 + r_{\text{stock}}),$$

where

$$r_{\text{stock}} = \frac{P_1 + D}{P_0} - 1$$

is the rate of return on the stock. The investor has to pay back $L_1 = L_0(1 + r_{\text{loan}})$ to the

broker, leaving

$$\begin{aligned} E_1 &= V_1 - L_1 = (E_0 + L_0)(1 + r_{\text{stock}}) - L_0(1 + r_{\text{loan}}) \\ &= E_0(1 + r_{\text{stock}}) + L_0(r_{\text{stock}} - r_{\text{loan}}) \end{aligned}$$

to the investor. Relative to the initial investment of E_0 , the investor's rate of return is

$$r = \frac{E_1}{E_0} - 1 = r_{\text{stock}} + \frac{L_0}{E_0}(r_{\text{stock}} - r_{\text{loan}}). \quad (2.17)$$

Had the investor just purchased stock without leveraging up, she would get a rate of return equal to r_{stock} . The last term in the above equation is the additional rate of return due to leverage. In line with intuition, this is positive if the stock return turns out to be higher than the interest rate paid on the loan. So if you are convinced that the stock return is going to exceed the interest rate, leveraging up is profitable. On the other hand, the additional return from leveraging is negative if the realized stock return is below the interest rate. Both high and low stock returns are amplified because of the leverage. A levered position is therefore more risky than an unlevered position, as we will show more formally in Section 3.4.

The ratio L_0/E_0 is called the *leverage ratio*. If you want to invest $V_0 = \$100,000$, but only have $E_0 = \$20,000$, you need to borrow $L_0 = \$80,000$. This is a leverage ratio of $L_0/E_0 = 4$. The higher the leverage ratio, the higher the risk, as the following example illustrates.

Example 2.7

Suppose you consider leveraging up an investment in a given stock, and the borrowing rate is $r_{\text{loan}} = 0.04 = 4\%$. Table 2.4 shows the rates of return on levered positions with leverage ratios 1 and 5 for various rates of return on the stock. We see that returns are getting more extreme as the leverage ratio increases. Figure 2.1 provides a graphical illustration of how returns on a leveraged position depends on the returns on the stock with each line representing a given leverage ratio. Note that the lines cross where the stock return equals the borrowing rate since the return on any leveraged position is then equal to the stock return.

Companies have varying degrees of leverage as measured by their debts relative to their equity (here, the market values of debt and equity, not the face value or book value, are relevant). Analogously to the above considerations, the rate of a return of a stock issued by a company is, other things equal, more risky the higher the leverage of the company. An investor who would like to gear her investment in the stock market, but is not able to obtain a loan to do so, can indirectly obtain a similar risk profile by investing her funds in stocks of highly leveraged companies. Such stocks offer *embedded leverage* to investors.

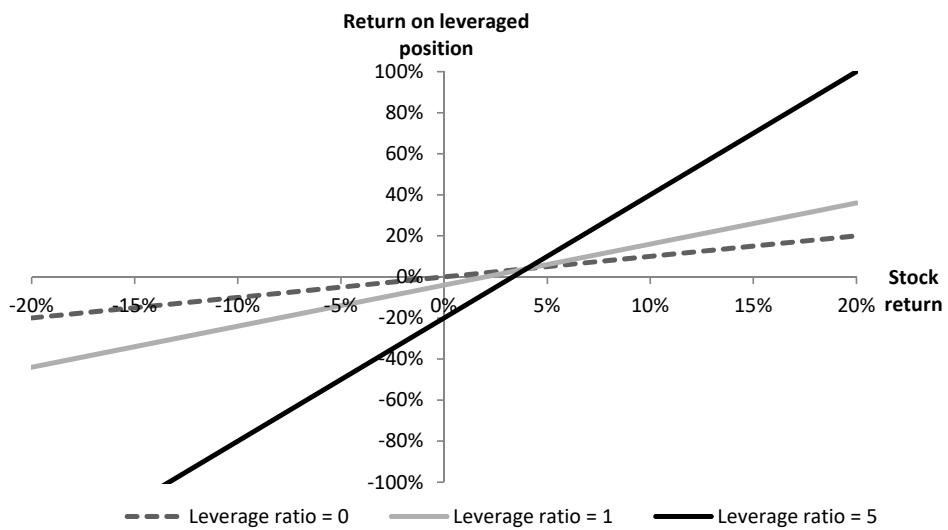
2.8 Returns on short positions

Suppose you are convinced that the stock price of the company XYZ will drop significantly over the next three months. What can you do to profit from this belief? Of course, if you already own some shares of XYZ stocks, you would then typically sell them immediately. But sometimes you can even sell XYZ stocks that you do not own! This is referred to as *selling short* or just *shorting* the stock. First, the short-seller borrows the

Stock return r_{stock}	Return on levered position	
	$L_0/E_0 = 1$	$L_0/E_0 = 5$
-20%	-44%	-140%
-16%	-36%	-116%
-12%	-28%	-92%
-8%	-20%	-68%
-4%	-12%	-44%
0%	-4%	-20%
4%	4%	4%
8%	12%	28%
12%	20%	52%
16%	28%	76%
20%	36%	100%

Table 2.4: Leverage and returns.

The table illustrates the effect of leverage on the rate of return on a stock investment. See Example 2.7.

**Figure 2.1: Leverage and returns.**

The figure illustrates the effect of leverage on the rate of return on a stock investment. See Example 2.7.

stock, then sells it, and at a later date he “covers the short position” by buying back the stock and returning it to the lender. The lender is missing out on any dividend payments from the stock over the period, so the short-seller has to compensate the lender by paying him an amount of money equal to the dividends. The lender can typically demand to get his stock back at short notice, in which case the short-seller has to buy the stock at the current price or borrow the stock from someone else. To distinguish it from a short sale, the more conventional investment of purchasing an asset and holding on to it is sometimes referred to taking or having a *long position* in the asset or simply *going long the asset*.

How do you compute the return on a short position? Assume in the following that you sell short a stock from time 0 to time 1. Let P_0 and P_1 denote the market price of the stock at the two dates and let D denote the dividend paid by the stock in the period between these dates. If you take a long position in the stock, you have to pay P_0 at time 0 but you receive D and then P_1 from selling the stock at time 1, so your rate of return would be

$$r_{\text{long}} = \frac{D + P_1 - P_0}{P_0},$$

where we ignore any timing issue with the dividend.

If, instead, you short sell the stock at time 0, you receive P_0 at time 0 but have to pay D and then P_1 when you close the short position at time 1. Your dollar return is therefore $P_0 - P_1 - D$. What about the rate of return? Normally, we compute the rate of return as the dollar return divided by the initial investment. In the case of a short sale, you can say that there is no initial investment as you receive money upfront. However, when entering the short sale you take on a liability: you promise to deliver back the stock in the future. The present market value of this liability is the current price of the stock, that is P_0 . Hence, the rate of return on a short sale is usually calculated relative to the initial price of the stock:

$$r_{\text{short}} = \frac{P_0 - P_1 - D}{P_0}. \quad (2.18)$$

Then the rate of return on a short sale is simply the negative of the rate of return on a long position:

$$r_{\text{short}} = -r_{\text{long}}. \quad (2.19)$$

When reinvesting any dividends received in the stock, one unit of the stock at time 0 would grow to $A_{0,1} \geq 1$ units at time 1 as explained in Section 2.1. Therefore, a person lending out one unit at time 0 would expect to get $A_{0,1}$ units back at time 1. Hence, the short seller receives P_0 at time 0, but must pay $A_{0,1}P_1$ at time 1 to satisfy the lender. This implies that the rate of return for the short seller is

$$r_{\text{short}} = \frac{P_0 - A_{0,1}P_1}{P_0}.$$

The rate of return on a long position would be $r_{\text{long}} = (A_{0,1}P_1 - P_0)/P_0$, so again we see that the conclusion (2.19) holds.

The best that can happen for a short seller is that the stock price goes to zero and no dividend is paid. Then the rate of return would be 100% which is therefore the maximum rate of return on a short sale. On the other hand, if the price prediction of the short seller turns out wrong and the price of the stock increases, the rate of return will be negative. Since there is no theoretical upper limit on P_1 , there is no lower bound on the rate of return of the short seller. If the price of the stock more than doubles, the rate of return of the short seller will be even lower than -100%. This is another indication that short

selling is a risky strategy. To limit the loss of the short position, many short sellers place a stop order with their broker when they initiate the short sale. Then the stock is bought back automatically if the price reaches a certain level above the current price.

Example 2.8

Suppose you sell short a share of stock in the company XYZ at a price of \$20. One month later you buy it back at \$15. During the month, XYZ paid a dividend of \$1 per share. What is your rate of return?

Out of the \$5 profit on the trades, you pay \$1 in dividend compensation to the lender of the stock, so your net profit is \$4. As explained above, the profit should be seen in relation to the initial stock price of \$20, so the rate of return is 20%. More formally, we can substitute the numbers into Eq. (2.18) to get

$$r_{\text{short}} = \frac{\$20 - \$15 - \$1}{\$20} = \frac{\$4}{\$20} = 0.2 = 20\%.$$

In practice, the short sale of a stock can be implemented by setting up a margin account with a broker who can then lend you the shares you want to short. The broker takes the shares from his own inventory or borrows them from the accounts of some of the broker's other clients or from another broker. Note that the proceeds from the short sale must stay on the account. In addition the broker requires the short seller to deposit margin (cash or assets that are easy to sell, e.g., government bonds) to insure the broker against a loss if the stock price goes up. The balance or equity of the margin account is computed at any time as the difference between the deposit (including the proceeds from the short sale) and the market value of the shorted stocks at that time. If the stock price decreases, as hoped, the equity will increase. But if the stock price increases, the equity will decrease. Typically the broker has a minimum or maintenance margin requirement that has to be fulfilled at all times in the sense that the equity must be at least a certain percent of the market value of the shorted stocks. If this minimum is not met, the broker issues a margin call prompting the client either to deposit additional cash or assets or to reduce or even close the position. The investor typically receives a lower interest rate on the margin account than the prevailing market interest rate, and individual investors might earn no interest at all. This missed interest constitutes a cost to shorting stocks.

For any stock the *short interest* at a given time is the number of shares currently sold short, typically calculated as a fraction of the total number of publicly tradable shares. Leading stock exchanges often report the short interest of the listed stocks. In principle, the short interest can exceed the number of shares outstanding since each stock can be shorted more than once. If you short a stock, you borrow it and sell it to a new owner. Another short seller can borrow the same stock from that new owner and sell it again. Suppose the short interest in a stock is a considerable fraction of the number of outstanding shares of the stock. If the stock price has recently increased, maybe due to good news about the company, a large number of disappointed short sellers might simultaneously want to buy back the stocks to reduce their losses or have to buy back the stocks due to margin calls they cannot meet. Such a *short squeeze* pushes the price of the stock even further up, thus increasing the losses that the short sellers must realize.

Some see short selling as destabilizing in the sense that shorting can exert a downward price pressure on a stock whose price may already have dropped recently. At various times

in history, typically during financial crises, market regulators have restricted or banned short selling. Others claim that the possibility of short selling is beneficial by making prices more informative. Short selling provides a way for pessimistic investors to influence prices. A large short interest of a stock indicates that many investors believe the stock is currently overvalued. This may lead both current stockholders and potential buyers to reconsider their assessment of the value. Substantial short selling of a stock can sometimes be an early warning of mismanagement or even illegal practices in the company issuing the stock. Various academic studies conclude that short-sale bans generally reduce market liquidity and slow down the price formation process, see, e.g., Beber and Pagano (2013).

2.9 Returns on portfolios

The combination of holdings of different assets at a given point in time is called a *portfolio*. Suppose first that you form a buy-and-hold portfolio of two assets: asset 1 and asset 2. The rates of return (or percentage returns) of the assets over the period considered are denoted by r_1 and r_2 , respectively. When you buy the portfolio, let w denote the fraction of your total investment which is invested in asset 1. Hence a fraction of $1 - w$ of wealth is invested in asset 2. Here a negative portfolio weight of an asset corresponds to a short position in that asset. Now you hold on to the assets to the end of the period without intermediate rebalancing of the portfolio. Then, one can show that rate of return on the portfolio is

$$r_p = wr_1 + (1 - w)r_2. \quad (2.20)$$

This follows as a special case of the theorem below. In words, the rate of return on a portfolio is a value-weighted average of the rates of return on the assets in the portfolio. There is a similar relation for the gross return on the portfolio:

$$R_p = wR_1 + (1 - w)R_2. \quad (2.21)$$

These relations are true both if you use known, historical returns and if you use possible future returns. In the latter case, this implies that the relations (2.20) and (2.21) hold no matter what the returns of the two assets turn out to be.

Example 2.9

Suppose you invest in $N_1 = 20$ shares of stock 1 at a unit price of $P_{1t} = \$50$ and in $N_2 = 30$ shares of stock 2 at a unit price of $P_{2t} = \$100$. The total investment is $V_t = 20 \times \$50 + 30 \times \$100 = \$4000$. The portfolio weight of stock 1 is $w = 20 \times \$50 / \$4000 = 0.25$ or 25% and, consequently, the portfolio weight of stock 2 is $1 - w = 0.75$ or 75%. Assume for simplicity that none of the stocks pay dividends in the following period.

If the price of stock 1 ends up at $P_{1,t+1} = \$60$ and the price of stock 2 at $P_{2,t+1} = \$110$, then the rate of return is 20% on stock 1 and 10% on stock 2. The weighted average rate of return is thus

$$0.25 \times 20\% + 0.75 \times 10\% = 12.5\%.$$

The value of the portfolio at the end of the period is

$$V_{t+1} = 20 \times \$60 + 30 \times \$110 = \$4500,$$

up \$500 or 12.5% from the initial value in agreement with the above calculation.

Alternatively, suppose the price of stock 1 ends up at $P_{1,t+1} = \$40$ and the price of stock 2 at $P_{2,t+1} = \$120$. Then the rate of return is -20% on stock 1 and $+20\%$ on stock 2. The weighted average rate of return is thus

$$0.25 \times (-20\%) + 0.75 \times 20\% = 10\%.$$

The value of the portfolio at the end of the period is $V_{t+1} = 20 \times \$40 + 30 \times \$120 = \$4400$, up $\$400$ or 10% from the initial value. Again, this agrees with the weighted average calculated above.

The next theorem extends the above conclusions to portfolios of N assets enumerated from 1 to N . We are interested in the rate of return r_p on the portfolio of a given period of time. For return calculations, the relevant characteristic of the portfolio is the portfolio weights in the different assets. For each $i = 1, \dots, N$, we let π_i denote the portfolio weight of asset i , which we define as the fraction of the entire portfolio's value which is invested in asset i . By definition, the portfolio weights sum to one:

$$\pi_1 + \pi_2 + \dots + \pi_N = 1.$$

We will explore the properties of portfolios further in subsequent chapters and also set up models for finding the optimal portfolio for any given investor.

Theorem 2.2

For each $i = 1, \dots, N$, let r_i denote the rate of return and R_i the gross return on asset i , and let π_i denote the portfolio weight of asset i . Then the portfolio's rate of return r_p and gross return R_p are

$$r_p = \pi_1 r_1 + \pi_2 r_2 + \dots + \pi_N r_N = \sum_{i=1}^N \pi_i r_i, \quad (2.22)$$

$$R_p = \pi_1 R_1 + \pi_2 R_2 + \dots + \pi_N R_N = \sum_{i=1}^N \pi_i R_i. \quad (2.23)$$

Proof

To simplify the notation, we provide a proof only for the case $N = 2$, i.e. for portfolios of two assets. Suppose the return is computed from time t to time $t + 1$. For each asset $i = 1, 2$, let P_{it} denote the unit price at time t and $P_{i,t+1}$ the unit price at time $t + 1$. Similarly, let $D_{i,t+1}$ denote the dividend of asset i per unit between t and $t + 1$. If, at time t , you invest in N_1 units of asset 1 and N_2 units of asset 2, the value of the portfolio is

$$V_t = N_1 P_{1t} + N_2 P_{2t}.$$

The weights of assets 1 and 2 are

$$w = \frac{N_1 P_{1t}}{V_t}, \quad 1 - w = \frac{V_t}{V_t} - \frac{N_1 P_{1t}}{V_t} = \frac{V_t - N_1 P_{1t}}{V_t} = \frac{N_2 P_{2t}}{V_t}.$$

At time $t + 1$ the value of the portfolio including dividends is

$$\begin{aligned} V_{t+1} &= N_1(P_{1,t+1} + D_{1,t+1}) + N_2(P_{2,t+1} + D_{2,t+1}) \\ &= N_1 P_{1t} \frac{P_{1,t+1} + D_{1,t+1}}{P_{1t}} + N_2 P_{2t} \frac{P_{2,t+1} + D_{2,t+1}}{P_{2t}} \\ &= N_1 P_{1t} R_{1,t+1} + N_2 P_{2t} R_{2,t+1}, \end{aligned}$$

where $R_{i,t+1}$ denotes the gross return on asset i as defined in Eq. (2.4). The gross return on the portfolio over the period is now computed as

$$\begin{aligned} R_{t+1} &= \frac{V_{t+1}}{V_t} = \frac{N_1 P_{1t} R_{1,t+1} + N_2 P_{2t} R_{2,t+1}}{V_t} \\ &= \frac{N_1 P_{1t}}{V_t} R_{1,t+1} + \frac{N_2 P_{2t}}{V_t} R_{2,t+1} = w R_{1,t+1} + (1 - w) R_{2,t+1} \end{aligned}$$

and, consequently, the rate of return of the portfolio is

$$\begin{aligned} r_{t+1} &= R_{t+1} - 1 = w R_{1,t+1} + (1 - w) R_{2,t+1} - w - (1 - w) \\ &= w(R_{1,t+1} - 1) + (1 - w)(R_{2,t+1} - 1) = w r_{1,t+1} + (1 - w) r_{2,t+1}, \end{aligned}$$

which proves (2.20).

We stress that the above conclusions hold for buy-and-hold portfolios, i.e. the portfolio is chosen at the beginning of the period and kept unchanged throughout the period over which the returns are measured. The portfolio weights to be used in the portfolio return formulas are determined from the asset values at the beginning of the period. During the period the relative prices of the assets in the portfolio will likely change, which would imply that the portfolio weights change. In some cases, investors trade during the period to keep portfolio weights constant over time. This requires selling some units of the assets whose price increased a lot and purchasing additional units of the assets whose price dropped. Then the above formulas for portfolio returns are not valid.

2.10 Log-returns

In addition to rates of return and gross returns, we shall also occasionally work with log-returns. The **log-return** $r_{t,t+\Delta t}^{\log}$ over a period between t and $t + \Delta t$ is defined as

$$r_{t,t+\Delta t}^{\log} = \ln(1 + r_{t,t+\Delta t}) = \ln R_{t,t+\Delta t}, \quad (2.24)$$

where $r_{t,t+\Delta t}$ is the rate of return and $R_{t,t+\Delta t}$ the gross return over the period. Conversely, we have

$$R_{t,t+\Delta t} = e^{r_{t,t+\Delta t}^{\log}}, \quad r_{t,t+\Delta t} = e^{r_{t,t+\Delta t}^{\log}} - 1.$$

Using the expression (2.4) for the gross return, we get that

$$P_t = e^{-r_{t,t+\Delta t}^{\log}} (D_{t+\Delta t} + P_{t+\Delta t}), \quad (2.25)$$

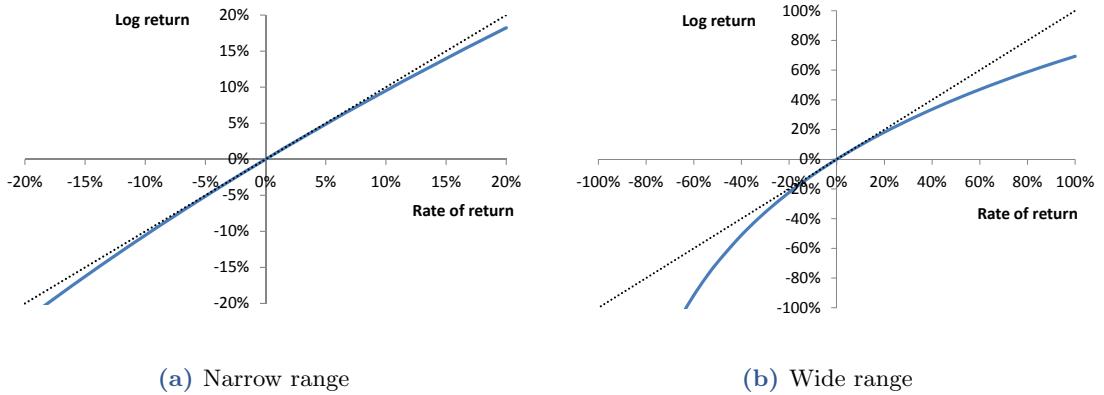


Figure 2.2: Log or not.

Figure 2.2: Log or not.

The graphs show the relation between the rate of return (horizontal axis) and the log-return (vertical axis). The dotted line is the 45 degree line. The log-return is smaller than the rate of return, but the difference is small for rates of return close to zero.

showing that log-returns correspond to exponential or continuous discounting as is also discussed below. As we shall see, log-returns have some nice properties that simplify various computations.

For a rate of return of zero, the corresponding log-return is also zero, since $\ln(1) = 0$. For any non-zero rate of return, the log-return is smaller than the rate of return. For small returns there is not much difference between the rate of return and the corresponding log-return. This is confirmed by the left panel of Figure 2.2. The 7.01965% rate of return in Example 2.1 corresponds to a log-return of $\ln(1.0701965) \approx 0.06784231$ or roughly 6.78%. The right panel shows that the difference is substantial for returns that are larger in absolute value. Also note that while the rate of return is bounded from below by -100% (assuming limited liability, the future prices and dividends are non-negative), the log-return is not bounded from below.

One can interpret the log-return r^{\log} corresponding to a given annual rate of return r as the simple annualized rate of return that would lead to an effective annual rate of return of r if returns are compounded infinitely frequently, i.e., continuously in time. To see this, first note that if the compounding happens N times per year, then the simple annualized rate of return $r_{(N)}^{\text{simp}}$ corresponding to the effective annual rate of return r is given by

$$\left(1 + \frac{r_{(N)}^{\text{simp}}}{N}\right)^N = 1 + r.$$

It is a mathematical fact that

$$\left(1 + \frac{x}{N}\right)^N \rightarrow e^x \quad \text{for } N \rightarrow \infty. \quad (2.26)$$

Hence, we see that with continuous compounding the simple annualized rate of return $r_{(\infty)}^{\text{simp}}$ satisfies

$$e^{r(\infty)^{\text{simp}}} = 1 + r$$

and is therefore given by

$$r_{(\infty)}^{\text{simp}} = \ln(1 + r),$$

which is exactly the log-return.

Theorem 2.1 showed how to compound rates of return and gross returns. The next theorem has the corresponding result for log-returns.

Theorem 2.3

Let $\Delta t > 0$ denote the period length. For each $i = 0, 1, \dots, n - 1$, let $r_{t+i\Delta t, t+(i+1)\Delta t}^{\log}$ denote the log-return over the period between time $t + i\Delta t$ and time $t + (i+1)\Delta t$. Then, over the entire n -period time interval from t to $t + n\Delta t$, the log-return $r_{t, t+n\Delta t}^{\log}$ is given by

$$r_{t, t+n\Delta t}^{\log} = r_{t, t+\Delta t}^{\log} + r_{t+\Delta t, t+2\Delta t}^{\log} + \dots + r_{t+(n-1)\Delta t, t+n\Delta t}^{\log}. \quad (2.27)$$

Proof

The result follows from (2.7) by taking these steps:

$$\begin{aligned} r_{t, t+n\Delta t}^{\log} &= \ln R_{t, t+n\Delta t} = \ln(R_{t, t+\Delta t} \times R_{t+\Delta t, t+2\Delta t} \times \dots \times R_{t+(n-1)\Delta t, t+n\Delta t}) \\ &= \ln R_{t, t+\Delta t} + \ln R_{t+\Delta t, t+2\Delta t} + \dots + \ln R_{t+(n-1)\Delta t, t+n\Delta t} \\ &= r_{t, t+\Delta t}^{\log} + r_{t+\Delta t, t+2\Delta t}^{\log} + \dots + r_{t+(n-1)\Delta t, t+n\Delta t}^{\log} \end{aligned}$$

where we have applied the rule $\ln(xy) = \ln x + \ln y$.

The theorem shows that the log-return over n periods, taking compounding into account, is simply the sum of the log-returns in each of the periods. This is a simpler relation than the corresponding relation (2.6) for rates of return, which is one advantage of working with log-returns. Similarly, annualization is easy, e.g., monthly log-returns are annualized simply by multiplying by 12:

$$r_{\text{ann}}^{\log} = 12 \times r_{\text{mon}}^{\log}. \quad (2.28)$$

Hence, aggregating log-returns over time is easy.

Another advantage of working with log-returns is that they fit better together with the normal distribution than rates of returns do. To model the uncertainty about a future return, we will often associate a probability distribution with that return. Here, the normal distribution is typically the first choice as this distribution is well known and tractable. The normal distribution assigns positive probabilities to any real number from $-\infty$ to $+\infty$. The log-return can be any real number and may therefore be normally distributed. Moreover, if the log-return in each period is normally distributed, then the log-return over multiple periods is also normally distributed. This follows from Eq. (2.27) and the fact that the sum of normally distributed random variables is also a normally distributed random variable. In contrast, the rate of return on an asset or an unlevered portfolio cannot be smaller than -100% since you cannot lose more than you invested, so it does not really make sense to assume that the rate of return is normally distributed. And if you would be willing to assume that the rate of return each period was normally distributed, then the rate of return over multiple periods would not be normally distributed due to the compounding rule (2.6) and the fact that a product of normally distributed random variables is not normally distributed. We discuss the normal distribution and its properties

in detail in Section 3.2 and the application of the normal distribution to investments in several subsequent chapters.

A disadvantage of using log-returns is that standard rules for returns on buy-and-hold portfolios do not apply. We saw in Section 2.9 that the rate of return on a buy-and-hold portfolio is a simple weighted average of the rates of return on the assets in the portfolio. Similarly for gross returns. But the same relation is not true for log-returns. To see this, consider a portfolio of two assets so that the gross return is $R_p = wR_1 + (1 - w)R_2$. Recall that the log-return on the portfolio is $r_p^{\log} = \ln R_p$ by definition. Since $\ln(wR_1 + (1 - w)R_2) \neq w \ln R_1 + (1 - w) \ln R_2$, we get that

$$r_p^{\log} \neq wr_1^{\log} + (1 - w)r_2^{\log}.$$

More generally, the log-return of the portfolio is *not* a weighted average of the log-returns on the assets in the portfolio:

$$r_p^{\log} \neq \pi_1 r_1^{\log} + \pi_2 r_2^{\log} + \cdots + \pi_N r_N^{\log}.$$

The exact relation between the log-return on the portfolio and the log-returns on the assets in the portfolio is more complicated:

$$r_p^{\log} = \ln \left(\pi_1 e^{r_1^{\log}} + \pi_2 e^{r_2^{\log}} + \cdots + \pi_N e^{r_N^{\log}} \right).$$

In particular, if we assume that the log-returns on the individual assets are jointly normally distributed, then the log-return on the portfolio is not going to be normally distributed.

We emphasize that these considerations involve returns on buy-and-hold portfolios, i.e., portfolios that are not rebalanced. In practice, many investors regularly rebalance their portfolios. As we will discuss in more detail in Section 8.1, it turns out that if individual assets' log-returns are normally distributed, and the portfolio is continuously rebalanced to keep portfolio weights constant, then the log-return on the portfolio is also normally distributed.

2.11 Exercises

Exercise 2.1. Consider the stocks of Ford Motor Company, which are listed on the New York Stock Exchange. Find the data on the internet (e.g., <http://finance.yahoo.com>) to answer the following questions:

- (a) What was the rate of return on an investment in Ford stocks in the year 2023?
- (b) What was the rate of return on Ford stocks in each of the months in 2023? How do you take dividend payments into account when calculating the returns?

Exercise 2.2. The price of stocks in Hypothetics Inc. is currently \$50 per share. By investing \$500 of your own money and borrowing the rest at an annual interest rate of 5%, you purchase 40 shares of the stock.

- (a) What is your leverage ratio?
- (b) Suppose you hold on to the shares for a one-year period, and that Hypothetics does not pay dividends during the year. What is the rate of return on your leveraged position if the stock price one year from now is \$20, \$30, \$40, \$50, \$60, \$70, \$80, \$90, or \$100, respectively?

Exercise 2.3. Suppose you hold the following portfolio of three stocks:

Stock name	Unit price	Number of stocks held
Alfalfa Industrial	\$25	80
Beat Roots Instruments	\$50	60
Citrust Investments	\$100	50

- (a) Calculate the portfolio weight of each stock.
- (b) Suppose that over the next month the stock prices change to \$15 (Alfalfa), \$40 (Beat Roots), and \$108 (Citrust). None of the companies are paying dividends. What is rate of return on each of the stocks? What is the rate of return on the portfolio? What are the portfolio weights now?
- (c) Suppose that after the price changes, you want to rebalance your portfolio so that the weights go back to their initial values. How do you achieve that? Can you do it without investing additional money or extracting any money from your investment portfolio?

Exercise 2.4. 52 weeks ago you invested \$400,000 in stocks of the company TGIM (Thank God It's Monday, Inc.). Every Monday in those 52 weeks, the stock price increased by 5%. Every Friday in those 52 weeks, the stock price decreased by 5%. On all other days, the stock price did not change. The company did not pay any dividends during the year. What is your rate of return over the 52-week period? Answer the same question if you replace the 5% by 1% and by 10%.

CHAPTER 3

Risk

When contemplating an investment in an asset, you might have a good idea about its future price and dividends, but in general you cannot know them for sure. Therefore, the return you obtain is uncertain or, in other words, risky until the end of the investment period. In mathematical terms, an uncertain quantity like a return is represented by a *random variable* characterizing the set of possible realizations and the associated probabilities. To understand and quantify the risk of investments, some knowledge of random variables is useful. Section 3.1 introduces random variables and probability distributions, as well as the key concepts of expectations, variances, and standard deviations. When making investment decisions, investors typically compare the return they expect to obtain with the risk of the investment, and the most basic risk measures are the variance and the standard deviation of the return. However, the variance or standard deviation do not always give a complete picture of the risk of an investment, and we also introduce the skewness and kurtosis which are two additional quantities describing the shape of a probability distribution. We also initiate the discussion of how we can quantify the tradeoff between expected return and risk, a topic that we shall review in later chapters.

The most frequently applied probability distribution in financial economics is the normal distribution which is described in Section 3.2. Many financial models assume that either the rate of return or the log-return on relevant assets are normally distributed, so it is important to know what such an assumption means and implies.

Investors look at many investment alternatives and often hold a portfolio of many assets. Since the return on each asset is a random variable, investors must deal with multiple random variables and how they are related. For example, the covariances and correlations between the returns on the different assets are important for the return on the portfolio. Section 3.3 introduces these concepts.

Section 3.4 lists and explains a number of highly useful computational rules for random variables such as how to compute the expectation and the variance of a sum of random variables. These rules are particularly useful for handling portfolios since the portfolio return is a weighted sum of the returns of the assets in the portfolio.

Some investors are particularly concerned about the risk of very low (highly negative) returns, i.e., the left tail of the probability distribution for the return. Section 3.5 defines and explains the so-called value at risk and expected shortfall, which are two frequently used measures of tail risk.

Investors decide not only which assets to invest in but also for how long they hold onto the assets. Hence, they are interested in how return and risk measures vary with the investment horizon. This is closely related to how the return in consecutive periods are related to each other. Does a large return in one period tend to be followed by a large return next period? If you assume that periodic rates of return or periodic log-returns are normally distributed, what can you say about returns over many periods? Section 3.6 provides a number of useful results and illustrations.

Investors care about returns in the future. If the future resembles the past, they can learn about the probability distribution of future returns by looking at patterns in past realized returns. For example, they may estimate the expected future return as the average realized return over a number of periods in the past. Section 3.7 explains how to form empirical return distributions and how to estimate key quantities of a probability distribution from a number of observed outcomes.

3.1 Random variables and summary statistics

We distinguish between discrete and continuous random variables. A *discrete random variable* has a finite set of possible realizations each with an associated probability. A *continuous random variable* has infinitely many possible realizations such as all non-negative real numbers or all values in some closed interval. We will describe each type of random variable in the following subsections.

Whenever there are many possible realizations, it may be difficult to communicate and evaluate the entire probability distribution. And also difficult to compare two different probability distributions which is what you have to do when choosing between two risky investments. Therefore, some key summary statistics are typically computed. The main summary statistics are the expected value (or the expectation), the variance, and the standard deviation. We explain these concepts below both for discrete and continuous random variables.

3.1.1 Discrete random variables

Let X be a discrete random variable that has S possible outcomes or realizations, which we denote by x_1, x_2, \dots, x_S . In this case, we sometimes say that there S states and that the realization in state s is x_s . At some future point in time, we know exactly which state is realized and thus which one of the possible outcomes that is the realized value of X , but right now we can only associate probabilities to the different possible outcomes. For each $s = 1, 2, \dots, S$, let p_s denote the probability that the realized value of X is going to be x_s , i.e.

$$p_s = \text{Prob}(X = x_s).$$

The *cumulative probability distribution function* F_X is defined as follows: For each x , $F_X(x)$ is the probability that the realized value of X is equal to x or lower than x :

$$F_X(x) = \text{Prob}(X \leq x) = \sum_{s \text{ with } X_s \leq x} p_s. \quad (3.1)$$

Note that the sum is over the states for which the realized value is smaller than or equal to x .

By applying any mathematical function to a random variable, you create a new random variable. For example, if X is the random variable that takes on the value 2 with a probability of 40% and the value 5 with a probability of 60%, then $Y = 3X + 5$ is the

random variable that takes on the value $3 \times 2 + 5 = 11$ with a probability of 40% and the value $3 \times 5 + 5 = 20$ with a probability of 60%. You can also add, subtract, multiply, or divide a random variable by another random variable.

The *expectation* or *expected value* of X is defined as

$$\mathbb{E}[X] = \sum_{s=1}^S p_s X_s, \quad (3.2)$$

i.e. the probability-weighted average of the possible outcomes. We can also define the expectation of a function $g(X)$:

$$\mathbb{E}[g(X)] = \sum_{s=1}^S p_s g(X_s). \quad (3.3)$$

The *variance* of X is defined as

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{s=1}^S p_s (x_s - \mathbb{E}[X])^2, \quad (3.4)$$

which is the sum of squared deviations from the expected value, weighted by the probabilities. The variance can also be written as

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad (3.5)$$

i.e. the expectation of the square less the square of the expectation.¹ The *standard deviation* of X is simply the square root of the variance:

$$\text{Std}[X] = \sqrt{\text{Var}[X]}. \quad (3.6)$$

If the random variable represents the rate of return on some asset or portfolio, the standard deviation is often referred to as the *volatility*. Both the variance and standard deviation are by definition non-negative, and they equal zero only if there is just a single possible realization of the random variable, meaning that there is no uncertainty. Obviously, there is a one-to-one relation between the variance and the standard deviation, so they contain the same information.

Example 3.1

Suppose you are interested in the rate of return r on the stocks of company XYZ over the next year. After intense analysis you firmly believe that there are five possible realizations of the return depending on the growth rate of the general economy. The possible values r_s (expressed in decimal numbers, not in percent) and their probabilities p_s are shown in columns 3 and 4 of Table 3.1. Column 5 shows the cumulative probabilities; returns

¹This follows from the calculation

$$\begin{aligned} \sum_{s=1}^S p_s (x_s - \mathbb{E}[X])^2 &= \sum_{s=1}^S p_s \left(x_s^2 + (\mathbb{E}[X])^2 - 2x_s \mathbb{E}[X] \right) = \sum_{s=1}^S p_s x_s^2 + (\mathbb{E}[X])^2 \sum_{s=1}^S p_s - 2 \mathbb{E}[X] \sum_{s=1}^S p_s X_s \\ &= \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2(\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

are ordered from lowest to highest, so just add up probabilities in the rows above. The probability distribution is illustrated in Figure 3.1.

Column 6 of the table shows the product $p_s r_s$ for each s , and the sum of 0.050 or 5.0% is therefore the expected return. Note that if you do this type of calculation in Excel, you can calculate the expectation by using the function **SUMPRODUCT** with two arguments, one being the range of cells containing the probabilities and the other being the range of cells containing the possible rates of return. In that case you do not need a column with the numbers $p_s r_s$.

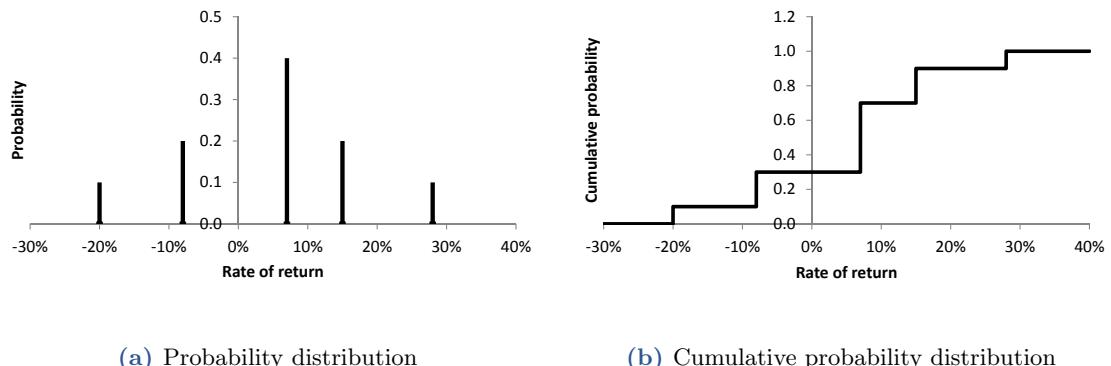
The two last columns are for the computation of the variance following Eq. (3.4). Again, by using **SUMPRODUCT** in Excel, you do not need the rightmost column. The variance turns out to be 0.01708. Had we written the rates of return in column 3 in percentage terms, e.g., -20 instead of -0.20, the variance calculation would result in 170.8 which has the unit of squared-percent, written as $(\%)^2$, since the variance formula involves squares of percentage differences.

Finally, from Eq. (3.6) the standard deviation is simply $\sqrt{0.01708} \approx 0.1307$ or 13.07%. The latter also equals the square root of $170.8(\%)^2$; taking the square root of $(\%)^2$ will give you %.

State s	Label	Rate of ret., r_s	Prob. p_s	Cum. prob.	$p_s r_s$	$r_s - E[r]$	$(r_s - E[r])^2$	$p_s(r_s - E[r])^2$
1	Lousy	-0.20	0.10	0.10	-0.020	-0.25	0.0625	0.00625
2	Bad	-0.08	0.20	0.30	-0.016	-0.13	0.0169	0.00338
3	Normal	0.07	0.40	0.70	0.028	0.02	0.0004	0.00016
4	Good	0.15	0.20	0.90	0.030	0.10	0.0100	0.00200
5	Great	0.28	0.10	1.00	0.028	0.23	0.0529	0.00529
Sum			1.00		0.050			0.01708

Table 3.1: Returns and probabilities.

Returns, probabilities, and relevant calculations. See Example 3.1.



(a) Probability distribution

(b) Cumulative probability distribution

Figure 3.1: Probability distributions.

The probability distribution and the cumulative probability distribution in Example 3.1.

3.1.2 Continuous random variables

Above we assumed a finite number of possible realizations of the uncertain quantity. In the example there were only five possible realizations of the return on the stock. However, for most real-life financial quantities it is unrealistic to assume only a low number of possible realizations.² For example, we observe stock returns of many, many different magnitudes and that should be reflected by our assumptions about possible future returns.

A continuous random variable can take on infinitely many possible values, which in most cases are all values in some interval, as for example the set of all non-negative real numbers $[0, \infty)$ where the symbol ∞ means infinity. A continuous random variable is in many respects easier to handle than a discrete random variable that can take on a large number of possible values. Therefore, if an uncertain quantity can take on a large number of possible values and these values are close to each other, it is often approximated by a continuous random variable. For example, stock exchanges typically list prices in multiples of 0.01 units of the local currency, and since stock prices are generally bounded from below by zero and unbounded from above, an uncertain future stock price is best represented by a continuous random variable on $[0, \infty)$.

The probability distribution of a continuous random variable is represented by its probability density function (sometimes abbreviated pdf) and its cumulative distribution function (sometimes abbreviated cdf). Let X denote a continuous random variable and let $f_X(x)$ denote the associated probability density function. This means that for any numbers a and b with $a < b$, the probability that the realized value of X turns out to be between a and b is given by

$$\text{Prob}(a < X < b) = \int_a^b f_X(x) dx. \quad (3.7)$$

To better understand what $f_X(y)$ for a given number y means, let Δ be a small positive number and use Eq. (3.7) with $a = y - (\Delta/2)$ and $b = y + (\Delta/2)$ to get

$$\text{Prob}\left(y - \frac{\Delta}{2} < X < y + \frac{\Delta}{2}\right) = \int_{y - \frac{\Delta}{2}}^{y + \frac{\Delta}{2}} f_X(x) dx \approx f_X(y) \times \Delta.$$

The approximation follows from the fact that the integral equals the area below the graph of the density function and, for a small Δ , this area is well approximated by a rectangle with height $f_X(y)$ and width Δ . The probability density function evaluated at any value y therefore captures the probability that the random variable ends up in a small neighborhood of y . Obviously, a probability density function has to be non-negative.

With infinitely many possible values, the probability of the random variable being equal to any specific value is, in fact, zero. Therefore, when computing the probability that X is in some interval, it does not matter whether we include the boundaries or not:

$$\text{Prob}(a < X < b) = \text{Prob}(a \leq X < b) = \text{Prob}(a < X \leq b) = \text{Prob}(a \leq X \leq b).$$

The cumulative distribution function associated with X shows, for every number x , the probability that X is smaller than or equal to x :

$$F_X(x) = \text{Prob}(X \leq x) = \text{Prob}(-\infty < X \leq x) = \int_{-\infty}^x f_X(y) dy. \quad (3.8)$$

²Some quantities are naturally limited to a low or moderate number of possible values. An example is the credit ratings that rating agencies assign to various debt securities.

This has the same meaning as the discrete-time analogue in Eq. (3.1). With a continuous random variable, it does not matter whether we include the equality in “ $X \leq x$ ” or not. If we know that X cannot be smaller than some number x_{\min} , which means that $F_X(x_{\min}) = \int_{-\infty}^{x_{\min}} f_X(y) dy = 0$, then we can split the integral in Eq. (3.8) to find that

$$F_X(x) = \int_{-\infty}^{x_{\min}} f_X(y) dy + \int_{x_{\min}}^x f_X(y) dy = \int_{x_{\min}}^x f_X(y) dy.$$

In words: when you cumulate probabilities from below, you can start from the lowest possible value.

Eq. (3.8) shows that the cumulative distribution function can be derived from the probability density function by appropriate integration. Conversely, by differentiating in (3.8) with respect to x , we obtain

$$F'_X(x) = \frac{d}{dx} \left(\int_{-\infty}^x f_X(y) dy \right) = f_X(x), \quad (3.9)$$

which shows that the probability density function can be obtained from the cumulative distribution function by differentiation.

Note that we have

$$\text{Prob}(a < X < b) = \text{Prob}(X < b) - \text{Prob}(X \leq a) = F_X(b) - F_X(a), \quad (3.10)$$

which is also true for discrete random variables. Furthermore, note that

$$F_X(\infty) = \int_{-\infty}^{\infty} f_X(x) dx = 1, \quad (3.11)$$

since the probability that X takes on *some* real value is 1.

We define the k 'th *percentile* or the $k\%$ percentile as the value of x for which

$$\text{Prob}(X \leq x) = k\% \Leftrightarrow F_X(x) = k\% \Leftrightarrow x = F_X^{-1}(k\%), \quad (3.12)$$

where you can replace $k\%$ by $k/100 = 0.01 \times k$, and where F_X^{-1} is the inverse of the cumulative distribution function.

The *expectation* of X is defined as

$$\text{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx, \quad (3.13)$$

analogous to the discrete summation in Eq. (3.2). Likewise, we define the expectation of a function $g(X)$ as

$$\text{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

In particular, the *variance* is then defined as

$$\text{Var}[X] = \text{E}[(X - \text{E}[X])^2] = \int_{-\infty}^{\infty} (x - \text{E}[X])^2 f_X(x) dx. \quad (3.14)$$

The variance rule

$$\text{Var}[X] = \text{E}[X^2] - (\text{E}[X])^2 \quad (3.15)$$

also applies for continuous random variables, and the *standard deviation* of X is again

defined as the square root of the variance,

$$\text{Std}[X] = \sqrt{\text{Var}[X]}, \quad (3.16)$$

just as for discrete random variables.

Another frequently used summary statistic for a probability distribution is the *median*. For a continuous random variable X , the median equals the value m for which

$$\text{Prob}(X \leq m) = \text{Prob}(X \geq m) = \frac{1}{2},$$

i.e., the median separates the upper half and the lower half of the distribution. The median is identical to the 50% percentile. We let $\text{Med}[X]$ denote the median of X . If the probability distribution is symmetric around some value M in the sense that the density function satisfies $f_X(M-y) = f_X(M+y)$ for all $y \geq 0$, then the median and the mean are both equal to M . This is for example the case for the normal distribution studied below.

Any non-negative function that integrates up to one is a valid probability density function, but some probability density functions are much more frequently used than others. For the most frequently applied probability density functions, values of the corresponding cumulative distribution function can be computed using built-in functions in most spreadsheets and statistical software packages and can also be read off standard probability tables often found in appendices to textbooks in statistics. Moreover, simple closed-form expressions for means and variances exist in these cases. Therefore, in most applications, you do not really need to compute integrals like those in Eqs. (3.8), (3.13), and (3.14). Section 3.2 describes the normal distribution which is the most commonly used probability distribution in finance, as in many other disciplines.

3.1.3 Higher-order moments

The n 'th moment of the random variable X is $E[X^n]$, while the n 'th central moment is $E[(X - E[X])^n]$. In particular, the expectation is the first moment and the variance is the second central moment. But in finance the third and fourth moments also receive some attention.

The *skew* or *skewness* of a random variable X is defined as the standardized third central moment:

$$\text{Skew}[X] = \frac{E[(X - E[X])^3]}{(\text{Std}[X])^3}. \quad (3.17)$$

For any symmetric distribution (such as the normal distribution), the skew is zero. If the probability density function leans to the left so that more than half of the probability mass (and therefore the expectation) is above the mode (the outcome with the maximum probability), the skew is positive. Conversely, if the probability density function leans to the right, the skew is negative. Figure 3.2 shows an example of a positively skewed and a negatively skewed distribution in comparison with a non-skewed distribution. The three distributions have identical means and identical variances.

The *kurtosis* of a random variable is closely related to the fourth standardized central moment, $E[(X - E[X])^4]/(\text{Std}[X])^4$. For a normal distribution this quantity equals 3. Since we often compare with a normal distribution, the kurtosis is defined as the deviation

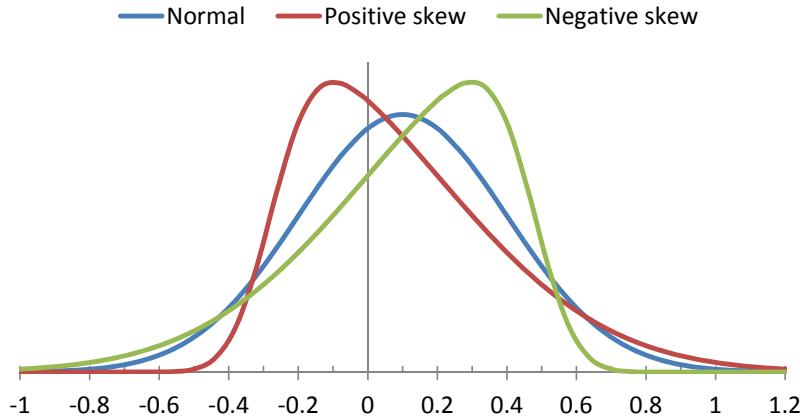


Figure 3.2: Skewed distributions.

The three curves show probability density functions. The blue curve represents a normal—and thus non-skewed—distribution. The red curve illustrates a positively skewed ($\text{Skew} = 0.85$) and the green curve a negatively skewed distribution ($\text{Skew} = -0.85$). The distributions have the same mean of 0.1 and the same standard deviation of 0.3. The two non-normal distributions are of the distribution type that is sometimes referred to as the “skewed normal distribution.”

of the fourth standardized central moment from that of a normal distribution:

$$\text{Kurt}[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{(\text{Std}[X])^4} - 3. \quad (3.18)$$

Sometimes this is referred to as the *excess kurtosis*. A distribution with a positive kurtosis is said to be *leptokurtic*. The graph of such a distribution has *fatter tails* than the normal distribution. That the tails are fat simply means there is a higher than normal probability of large positive and large negative returns realizations. A distribution with a negative kurtosis is said to be *platykurtic* and has slimmer tails than the normal distribution. Figure 3.3 depicts an example of a distribution with a positive kurtosis and a distribution with a negative kurtosis, together with the zero-kurtosis normal distribution. The right panel clearly shows the differences in the left tail. Since the distributions shown are symmetric, the right tail is just the mirror image of the left tail.

3.1.4 The risk-return tradeoff

The expected rate of return $\mathbb{E}[r]$ is a key quantity in assessing the attractiveness of an investment. For risky investments, the expected excess rate of return is often used, i.e., the expected rate of return less the riskfree rate of return r_f over the same period. The expected excess rate of return $\mathbb{E}[r] - r_f$ is sometimes called the *risk premium*. The risk of an investment is often measured by the standard deviation of the distribution of the rate of return, $\text{Std}[r]$. Investors prefer large expected returns and low risk, but as we shall see in later chapters large expected returns are typically associated with a large risk. Hence, it is relevant to measure the risk-return tradeoff, i.e., the expected return or risk premium

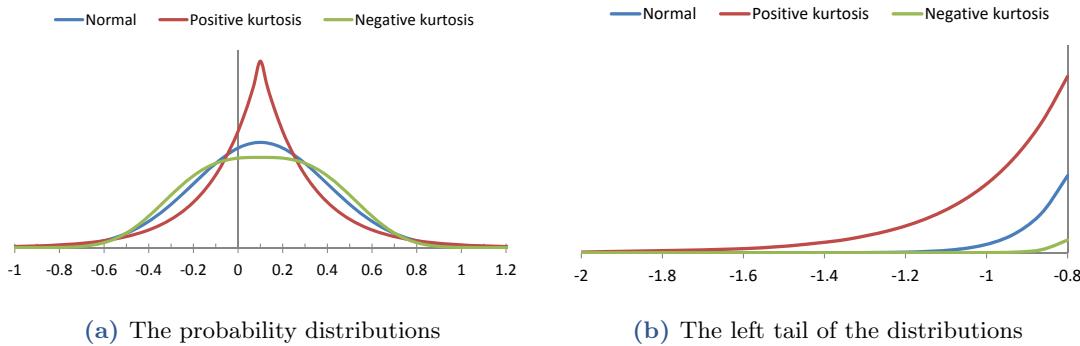


Figure 3.3: Distributions with non-zero kurtosis.

The three curves show probability density functions. The blue curve represents a normal—and thus zero-kurtosis—distribution. The red curve illustrates a distribution with positive kurtosis ($\text{Kurt} = 3$) and the green curve a distribution with negative kurtosis ($\text{Kurt} = -0.58$). The distributions have the same mean of 0.1 and the same standard deviation of 0.3. The two non-normal distributions are of the distribution type that is sometimes referred to as the “generalized normal distribution.”

relative to the risk taken.³

A popular way to quantify the risk-return tradeoff is the so-called *Sharpe ratio* defined as the ratio between the risk premium and the standard deviation,

$$\text{SR} = \frac{\text{E}[r] - r_f}{\text{Std}[r]}. \quad (3.19)$$

The ratio shows the reward (in terms of extra expected return) per unit of risk. The Sharpe ratio is named after William F. Sharpe who contributed immensely to modern investment theory and was awarded the 1990 Nobel Prize in Economics.

While the standard deviation is a useful quantification of the dispersion of a probability distribution, it might not necessarily be the best risk measure. First, the standard deviation includes both positive and negative deviations from the expected value, but you can argue that only the negative deviations—or downside risk—should be included in a risk measure. Hence, instead of the standard deviation some analysts and investors apply the **lower partial standard deviation**, which is computed just as the standard deviation but only using the realizations below either the expected value or below the riskfree return over the same period. The so-called **Sortino ratio**, named after Frank A. Sortino who advocated the use of downside risk in investment decisions, is similar to the Sharpe ratio, but the lower partial standard deviation is used in the denominator instead of the standard deviation.

Moreover, while the standard deviation (or the lower partial standard deviation) might be a reasonable measure of the riskiness of an investor’s total portfolio of investments, it is not obvious that it is relevant for individual assets. When assessing the risk of a particular asset, investors are probably more interested in the contribution of each asset to the overall risk of the portfolio. We revisit this issue and introduce other risk-reward measures in subsequent chapters.

³It is really a tradeoff between risk and *expected* return, but “risk-return tradeoff” sounds better than “risk-expected return tradeoff.”

3.2 The normal distribution

The normal distribution is the most frequently applied distribution for continuous random variables. Several of the models presented in later chapters assume that either the rate of return or the log-return (over some specified period) is normally distributed. The normal distribution is fully characterized by just two parameters—the mean and the variance—from which the full probability density function and all other relevant moments follow.

3.2.1 Definition and main properties

A random variable X is said to be normally distributed (or “follow a normal distribution”) with mean μ and variance σ^2 , if X is a continuous random variable with the probability density function f_X given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}, \quad (3.20)$$

where $\pi \approx 3.14159$. In this case, we write $X \sim N(\mu, \sigma^2)$, but note that some people write $X \sim N(\mu, \sigma)$ instead. There is no simpler expression for the cumulative distribution function than the general integral-type expression in (3.8). Note that a normally distributed random variable can take on any value from $-\infty$ to ∞ .

In accordance with the terminology used for the parameters μ and σ^2 , one can show

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} xf_X(x) dx = \mu, \\ \text{Var}[X] &= \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x) dx = \sigma^2. \end{aligned}$$

Obviously, we then have $\text{Std}[X] = \sigma$. Note that the full distribution is characterized by the expectation and the variance or, equivalently, the expectation and the standard deviation. Both the skewness and kurtosis defined in Eqs. (3.17) and (3.18) are zero:

$$\text{Skew}[X] = \text{Kurt}[X] = 0.$$

The normal distribution is sometimes referred to as a *Gaussian distribution* in honor of the German scientist Carl Friedrich Gauss who suggested the normal distribution in 1809.

Panel (a) of Figure 3.4 depicts three probability density functions for the normal distribution, all with mean zero, but with different σ -parameters. The larger the σ , the more spread out is the distribution, in line with the interpretation of the variance. Note that the density function can have values larger than one. Also note that the probability density function is bell-shaped and in particular symmetric around the mean so that $f_X(\mu + y) = f_X(\mu - y)$ as can also be shown directly from (3.20). Due to the symmetry, the mean and the median are identical so that the probability of an outcome below the mean is exactly 50% and, consequently, so is the probability of an outcome above the mean. In particular, we have

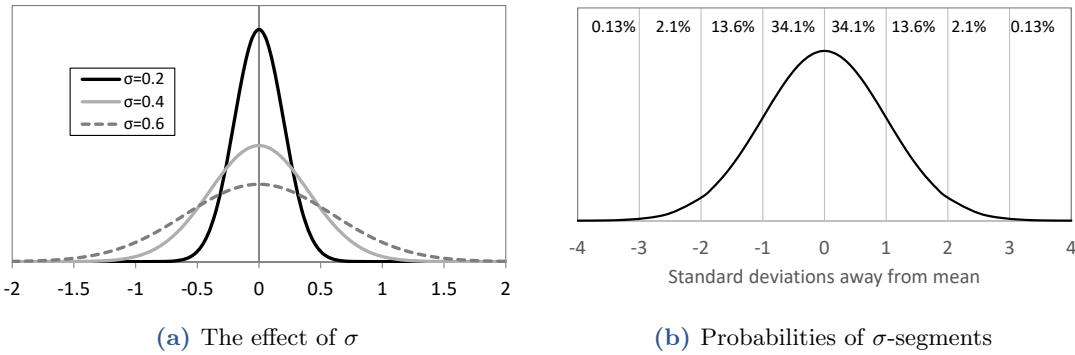
$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \text{Med}[X] = \mu.$$

If $\mu = 0$ and $\sigma^2 = 1$, the random variable X is said to be *standard normally distributed*,

0.1%	0.5%	1%	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%	99.5%	99.9%
-3.090	-2.576	-2.326	-1.960	-1.645	-1.282	-0.674	0.000	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Table 3.2: Selected percentiles of the standard normal distribution.

The table shows percentiles of the standard normal distribution that are frequently used.

**Figure 3.4:** The probability density function of a normal distribution.

In Panel (a), each curve illustrates the probability density function for a normal distribution with mean $\mu = 0$ and the σ -parameter shown in the figure legend. In Panel (b), the curve shows the probability density function of a normal distribution as a function of the number of standard deviations away from the mean. The percentage written in each vertical box shows the probability that the outcome ends up in that segment.

which we write $X \sim N(0, 1)$. The probability density function is then simply

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \quad (3.21)$$

We use $N(x)$ to denote the cumulative distribution function of a standard normal distribution so that

$$N(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \quad (3.22)$$

As for any continuous random variable, we have $\text{Prob}(X \geq x) = 1 - \text{Prob}(X \leq x) = 1 - N(x)$, but since the standard normal distribution is symmetric around zero we also have $\text{Prob}(X \geq x) = \text{Prob}(X \leq -x) = N(-x)$, so we can conclude that

$$\text{Prob}(X \geq x) = N(-x), \quad N(-x) = 1 - N(x). \quad (3.23)$$

The $k\%$ percentile of the standard normal distribution is the value x that satisfies

$$N(x) = k\% \Leftrightarrow x = N^{-1}(k\%). \quad (3.24)$$

Table 3.2 shows selected percentiles from the standard normal distribution. Note that since the distribution is symmetric around zero, the $k\%$ percentile equals minus the $(100 - k)\%$ percentile. For example, the 2.5% percentile is -1.960 and the 97.5% percentile is 1.960 . These numbers also imply that there is a 95% probability that the realized outcome is between -1.960 and $+1.960$. Likewise, there is a 99% probability that the outcome is between -2.576 and $+2.576$.

Excel function	Purpose	Math expression
<code>NORM.S.DIST($x; 0$)</code>	Probability density func. for $N(0, 1)$	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
<code>NORM.S.DIST($x; 1$)</code>	Cumulative distribution func. for $N(0, 1)$	$N(x)$
<code>NORM.S.INV(p)</code>	Percentile from $N(0, 1)$	$N^{-1}(p)$
<code>NORM.S.INV(RAND())</code>	Random draw from $N(0, 1)$	
<code>NORM.DIST($x; \mu; \sigma^2; 0$)</code>	Probability density func. for $N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$
<code>NORM.DIST($x; \mu; \sigma^2; 1$)</code>	Cumulative distribution func. for $N(\mu, \sigma^2)$	$N\left(\frac{x - \mu}{\sigma}\right)$
<code>NORM.INV($p; \mu; \sigma^2$)</code>	Percentile from $N(\mu, \sigma^2)$	$\mu + \sigma \times N^{-1}(p)$
<code>NORM.INV(RAND(); $\mu; \sigma^2$)</code>	Random draw from $N(\mu, \sigma^2)$	

Table 3.3: Excel functions for the normal distribution.

The table lists Excel functions that are useful for making relevant calculations for normally distributed random variables. The Excel function `RAND()` delivers a random number between 0 and 1. More precise methods exist for making random draws from a normal distribution.

It can be shown that

$$X \sim N(\mu, \sigma^2) \Leftrightarrow \frac{X - \mu}{\sigma} \sim N(0, 1). \quad (3.25)$$

Since $X \leq x$ if and only if $\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}$, we have

$$X \sim N(\mu, \sigma^2) \Rightarrow \text{Prob}(X \leq x) = \text{Prob}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = N\left(\frac{x - \mu}{\sigma}\right). \quad (3.26)$$

When $X \sim N(\mu, \sigma^2)$, the $k\%$ percentile is given by

$$x = \mu + \sigma \times N^{-1}(k\%). \quad (3.27)$$

The percentiles for the standard normal distribution shown in Table 3.2 can therefore easily be translated into percentiles for a non-standard normal distribution.

Panel (b) in Figure 3.4 shows the probabilities that the outcome ends up in different intervals defined by the distance from the mean μ in terms of multiples of the standard deviation σ . For example, there is only a probability of 0.13% that the outcome is more than three standard deviations smaller than the mean and, by symmetry, the same probability that the outcome is more than three standard deviations larger than the mean. Hence, only with a probability of around 0.27% (differs from $2 \times 0.13\%$ only since all numbers are rounded to the number of digits shown), the outcome is more than three standard deviations away from the mean. The probability is 68.3% that the outcome is within one standard deviation from the mean and 95.4% within two standard deviations from the mean.

Table 3.3 lists built-in Excel functions that are useful when working with normally distributed random variables. Similar functions are available in other computational software. The next example illustrates some applications of these functions for calculating various probabilities involving the return on an asset.

Example 3.2

Suppose the rate of return on Apple stocks over the next year, r_A , is normally distributed with mean 0.10 and variance 0.25. Calculate the following probabilities:

(a) $\text{Prob}(r_A < 0) = ?$

In Excel use `NORM.DIST(0;0.1;0.5;1)`, which gives 0.4207 or 42.07%. Here 0.5 is the standard deviation, calculated as the square root of the variance, $\sqrt{0.25} = 0.5$.

(b) $\text{Prob}(r_A > 0.3) = ?$

Note that $\text{Prob}(r_A > 0.3) = 1 - \text{Prob}(r_A < 0.3)$. We can calculate $\text{Prob}(r_A < 0.3)$ in Excel as `NORM.DIST(0.3;0.1;0.5;1)`. We thus get $\text{Prob}(r_A > 0.3) = 0.3446 = 34.46\%$.

(c) $\text{Prob}(0 < r_A < 0.3) = ?$

Recall that $\text{Prob}(0 < r_A < 0.3) = \text{Prob}(r_A < 0.3) - \text{Prob}(r_A < 0)$ and both of the probabilities on the right-hand side of the equality sign can be calculated in Excel using `NORM.DIST`, cf. the above questions. We get $\text{Prob}(0 < r_A < 0.3) = 0.2347 = 23.47\%$.

(d) $\text{Prob}((r_A)^2 > 0.04) = ?$

First observe that $(r_A)^2 > 0.04$ if either $r_A > 0.2$ or $r_A < -0.2$. Hence, $\text{Prob}((r_A)^2 > 0.04) = \text{Prob}(r_A > 0.2) + \text{Prob}(r_A < -0.2)$. The latter probability can be calculated directly using `NORM.DIST`, which gives $\text{Prob}(r_A < -0.2) = 0.2743$. We have to rewrite the first probability as $\text{Prob}(r_A > 0.2) = 1 - \text{Prob}(r_A < 0.2) = 0.4207$. In sum, we get $\text{Prob}((r_A)^2 > 0.04) = 0.4207 + 0.2743 = 0.6950 = 69.50\%$.

(e) $\text{Prob}(r_A^{\log} < 0.3) = ?$

Recall that $r_A^{\log} = \ln(1 + r_A)$ so

$$r_A^{\log} < 0.3 \Leftrightarrow \ln(1 + r_A) < 0.3 \Leftrightarrow r_A < e^{0.3} - 1 \approx 0.34986.$$

Therefore, $\text{Prob}(r_A^{\log} < 0.3) = \text{Prob}(r_A < e^{0.3} - 1)$, which can be calculated by `NORM.DIST`, giving 0.6914 or 69.14%.

In some finance models, the rates of return of stocks are assumed to be normally distributed. This is certainly not the worst assumption you can make since a frequency plot of the historically observed returns on a given stock typically resembles the bell-shaped density of a normal distribution. However, the fit is not perfect, as will be discussed in Section 3.7 below. Stock returns of individual firms (in particular, small firms) typically display a positive skew, whereas returns of stock indices and broad stock portfolios display a negative skew, cf. Albuquerque (2012). Stock returns are often said to have fat tails, i.e., be leptokurtic. However, this is not true for all stocks, and the kurtosis—as the skew—may vary over time. More information on empirical distributions of stock returns is presented in Section 6.5.

The rates of return on an investment with limited liability cannot be below -1 (i.e., -100%), as you cannot lose more than you invest. Since the normal distribution assigns a positive probability to that event, it is clear that rates of return cannot be perfectly described by a normal distribution. However, for reasonable values of μ and σ^2 , this probability is very, very low so you might be willing to accept such an inconsistency. The assumption that the rate of return r is normally distributed also has the unpleasant implication that the log return $r^{\log} = \ln(1+r)$ is not a well-defined random variable exactly

because of the positive probability of r being smaller than -1 in which case $\ln(1 + r)$ is undefined.

On the other hand, log-returns are not bounded from below by -1 or any other number, cf. the discussion following Eq. (2.24). Hence, the before-mentioned problems do not arise if you assume the log-return r^{\log} is normally distributed. Under such an assumption, the rate of return $r = \exp(r^{\log}) - 1$ stays above -1 and is a well-defined random variable. The assumption of normally distributed log-returns is also frequently made in finance. The next subsection explores which implications such an assumption has for the rate of return.

3.2.2 Normally distributed log-returns

A continuous random variable X is said to follow a *lognormal distribution* (or to be lognormally distributed) with parameters m and s^2 if the natural logarithm $\ln X$ follows a normal distribution with mean m and variance s^2 . As defined in (2.24), the log-return r^{\log} over a given period is linked to the rate of return r and the gross return R by

$$r^{\log} = \ln R = \ln(1 + r).$$

Since $r^{\log} = \ln R$, an assumption that the log-return r^{\log} is normally distributed with mean m and variance s^2 is equivalent to an assumption that the gross return R is lognormally distributed with parameters m and s^2 . Then $r = R - 1$ follows a so-called *shifted lognormal distribution*. The lognormal distribution is a well-studied distribution with known and tractable expressions for probabilities and key moments, of which many are stated and derived in Appendix A together with some implications that are particularly useful in the Black-Scholes option pricing model that we study in Chapter 15. The results on the lognormal distribution lead to similar expression for the shifted lognormal distribution. Hence, if we assume that the log-return r^{\log} is normally distributed, we can characterize the distribution that the rate of return $r = \exp(r^{\log}) - 1$ follows. The next theorem collects some important results.

Theorem 3.1

Assume $r^{\log} \sim N(m, s^2)$. Then, for any $x > -1$, the probability density function and the cumulative distribution functions for the rate of return r are

$$f_r(x) = \frac{1}{(1+x)\sqrt{2\pi s^2}} \exp\left\{-\frac{(\ln(1+x)-m)^2}{2s^2}\right\}, \quad (3.28)$$

$$F_r(x) = N\left(\frac{\ln(1+x)-m}{s}\right), \quad (3.29)$$

and $f_r(x) = F_r(x) = 0$ for $x \leq -1$. The percentiles in the distribution of the rate of return are given by

$$p = \text{Prob}(r < x_p) \Leftrightarrow x_p = \exp\left\{sN^{-1}(p) + m\right\} - 1. \quad (3.30)$$

The key moments for the rate of return r are

$$\mathbb{E}[r] = e^{m+\frac{1}{2}s^2} - 1, \quad (3.31)$$

$$\text{Med}[r] = e^m - 1, \quad (3.32)$$

$$\text{Var}[r] = e^{2m+s^2} (e^{s^2} - 1), \quad (3.33)$$

$$\text{Skew}[r] = (e^{s^2} + 2) \sqrt{e^{s^2} - 1}, \quad (3.34)$$

$$\text{Kurt}[r] = e^{4s^2} + 2e^{3s^2} + 3e^{2s^2} - 6. \quad (3.35)$$

Proof

Note that $r \leq x$ if and only if $r^{\log} \leq \ln(1+x)$. Hence, the cumulative distribution function for r is given by

$$F_r(x) = \text{Prob}(r \leq x) = \text{Prob}\left(r^{\log} \leq \ln(1+x)\right) = N\left(\frac{\ln(1+x) - m}{s}\right),$$

where the last equality follows from (3.26) and the assumption that $r^{\log} \sim N(m, s^2)$. According to (3.9), the probability density function f_r follows from differentiation of F_r :

$$\begin{aligned} f_r(x) &= F'_r(x) = \frac{1}{(1+x)s} N'\left(\frac{\ln(1+x) - m}{s}\right) \\ &= \frac{1}{(1+x)s} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\ln(1+x) - m)^2}{2s^2}\right\}, \end{aligned}$$

which gives (3.28). The percentiles follows from

$$\begin{aligned} p &= \text{Prob}(r < x_p) = F_r(x_p) = N\left(\frac{\ln(1+x_p) - m}{s}\right) \\ \Leftrightarrow N^{-1}(p) &= \frac{\ln(1+x_p) - m}{s} \\ \Leftrightarrow \ln(1+x_p) &= sN^{-1}(p) + m \\ \Leftrightarrow x_p &= \exp\left\{sN^{-1}(p) + m\right\} - 1. \end{aligned}$$

The relevant moments of the lognormal distribution—and thus of the gross return R —are derived in Appendix A. From these expressions, we can easily calculate the corresponding moments of $r = R - 1$. For example, the mean is $\mathbb{E}[r] = \mathbb{E}[R] - 1$. The variance is $\text{Var}[r] = \text{Var}[R - 1] = \text{Var}[R]$ and similarly for the skewness and kurtosis.

The skewness and kurtosis are both positive. The median is smaller than the mean, which is consistent with the positive skewness. The maximum value of the probability density function is reached for $x = \exp(m - s^2)$, the so-called *mode* of the distribution. The next example illustrates some calculations of relevant moments and probabilities when

the log-return on an asset is normally distributed.

Example 3.3

Suppose the log return on Facebook stocks over the next year, r_{FB}^{\log} , is normally distributed with mean 0.04 and variance 0.16. Calculate the following quantities:

(a) $E[r_{FB}] = ?$

The assumption is that $r_{FB}^{\log} = \ln(1 + r_{FB})$ is normally distributed with mean $m = 0.04$ and standard deviation $s = \sqrt{0.16} = 0.4$. From Eq. (3.31) it follows that $E[r_{FB}] = e^{m+\frac{1}{2}s^2} - 1 = 0.1275$ or 12.75%.

(b) $\text{Std}[r_{FB}] = ?$

From Eq. (3.33) it follows that the variance of the rate of return is $\text{Var}[r_{FB}] = e^{2m+s^2} (e^{s^2} - 1) = 0.2206$. The standard deviation is therefore $\text{Std}[r_{FB}] = \sqrt{0.2206} \approx 0.4697$ or 46.97%.

(c) $\text{Prob}(r_{FB}^{\log} < 0) = ?$

Since the log-return is normally distributed, this probability is calculated directly using `NORM.DIST(0;0.04;0.4;1)`, which gives 0.4602 or 46.02%.

(d) $\text{Prob}(r_{FB} < 0) = ?$

Note that

$$r_{FB} < 0 \Leftrightarrow 1 + r_{FB} < 1 \Leftrightarrow r_{FB}^{\log} = \ln(1 + r_{FB}) < \ln(1) = 0,$$

so $\text{Prob}(r_{FB} < 0) = \text{Prob}(r_{FB}^{\log} < 0)$. Hence, we get the same answer as in the previous question: 0.4602 or 46.02%.

(e) Find x so that $\text{Prob}(r_{FB} < x) = 0.05$.

We are looking for the 5% percentile in the distribution of r_{FB} . According to Eq. (3.30), this is given by

$$x = e^{sN^{-1}(0.05)+m} - 1 = e^{0.4 \times (-1.645) + 0.04} - 1 \approx -0.4610,$$

where we have taken $N^{-1}(0.05) = -1.645$ from Table 3.2. The value of -1.645 is rounded to the number of digits shown. If you insert the unrounded value of $N^{-1}(0.05)$, e.g. calculated by `NORM.S.INV(0.05)` in Excel, directly in the above formula, you get $x = -0.4609$.

Alternatively, you can work it out in the following way: first note that $r_{FB} < x$ if and only if $\ln(1 + r_{FB}) < \ln(1 + x)$ so $\text{Prob}(r < x) = \text{Prob}(\ln(1 + r) < \ln(1 + x)) = 0.05$. Since $\ln(1 + r)$ is normally distributed we can find $\ln(1 + x)$ as `NORM.INV(0.05;m;s)` which gives -0.6179 . Finally, $\ln(1 + x) = -0.6179$ implies $x = e^{-0.6179} - 1 = -0.4609$.

The expected rate of return can be approximated as

$$E[r] = e^{m+\frac{1}{2}s^2} - 1 \approx m + \frac{1}{2}s^2,$$

where the approximation is good for values of $m + \frac{1}{2}s^2$ near zero. Note that the expected

rate of return exceeds the expected log-return,

$$\mathbb{E}[r] \geq \mathbb{E}[r^{\log}],$$

since $e^{m+\frac{1}{2}s^2} - 1 \geq e^m - 1 \geq m$. If $m < 0$ and $m + \frac{1}{2}s^2 > 0$, the expected rate of return is positive, but the median rate of return $e^m - 1$ is negative so that the probability of a negative rate of return is more than 50%. Also note that $e^{s^2} - 1 > s^2$ so if $2m + s^2 > 0$ and thus $e^{2m+s^2} > 1$, we have

$$\text{Var}[r] \geq \text{Var}[r^{\log}].$$

Suppose you want the rate of return r to have an expectation of $\mathbb{E}[r] = \mu$ and a variance of $\text{Var}[r] = \sigma^2$. Of course, this is satisfied if you assume that the rate of return is normally distributed with mean μ and variance σ^2 . Alternatively, you can assume that the log-return $r^{\log} = \ln(1 + r)$ is normally distributed with parameters m and s^2 given by

$$m = 2 \ln(1 + \mu) - \frac{1}{2} \ln \left([1 + \mu]^2 + \sigma^2 \right), \quad s^2 = \ln \left(1 + \frac{\sigma^2}{(1 + \mu)^2} \right). \quad (3.36)$$

In this case the rate of return follows the shifted lognormal distribution characterized in Theorem 3.1 and, with m and s^2 given by (3.36), it can be verified that $\mathbb{E}[r] = \mu$ and $\text{Var}[r] = \sigma^2$.

Figure 3.5 compares the rate of return distribution in the two cases. The expected rate of return is fixed at $\mu = 0.1$ or 10%, which is reasonable for some stocks or a stock index over a one-year investment horizon. In the left panel, the standard deviation of the return is $\sigma = 0.2$ or 20%, which is relatively low and may correspond to a stock index or a broad portfolio of stocks. In the right panel, the standard deviation is 50%, which is relatively high and may correspond to an individual stock. (We present empirical estimates of the mean and variance of returns in Section 3.7 and in later chapters.) With the low standard deviation in the left panel, the rate of return distributions are very similar under the two distributional assumptions. However, with the high standard deviation in the right panel, the two distributions differ more. For the case with normally distributed log-returns, we can clearly see the positive skewness and the limited left tail that correctly assigns a zero probability of realizing a rate of return smaller than -1 , i.e. smaller than -100% . By contrast, with a normally distributed rate of return, this event is assigned a positive probability: less than 0.000002% in the left panel, but around 1.39% in the right panel.

3.2.3 Are real-life returns normal?

How can we assess whether it is reasonable to assume that the rate of return or the log-return on some asset over a future period of some length is normally distributed? We can often find a time series of past returns on the same asset over non-overlapping periods of that same length. Assuming that all these past returns were drawn from the same distribution as the future return we care about, we can use the past returns to get an impression of that distribution. We can form a histogram of the observed returns to depict the empirical distribution and, for example, assess how well it resembles a normal distribution.

Figure 3.6 shows a histogram of the monthly rates of return (left panel) and monthly log-returns (right panel) on the S&P 500 stock index over the 924 months from January 1946 to 2022. The returns include dividends and are adjusted for stock splits, etc. The data was obtained from the CRSP database maintained by the Center for Research in Security

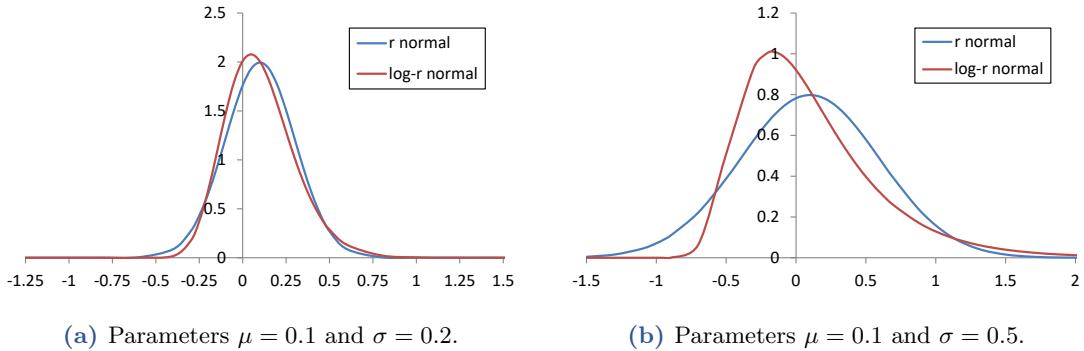


Figure 3.5: Comparing rate of return distributions.

Each curve shows a probability density function for the rate of return. The blue curves are for the case where the rate of return itself is normally distributed with mean μ and variance σ^2 . The red curves are for the case where the log-return is normally distributed with mean m and variance s^2 , where m and s^2 are given by (3.36). In the left panel, $\mu = 0.1$ and $\sigma = 0.2$ so that $m \approx 0.07905$ and $s \approx 0.18034$. In the right panel, $\mu = 0.1$ and $\sigma = 0.5$ so that $m \approx 0.00140$ and $s \approx 0.43338$.

Prices at the University of Chicago's Booth School of Business. The blue bars show the observed frequency in different intervals or *bins*. For example, in the left panel, the tallest bar shows that in 202 months the rate of return was between 0 and 0.02 (i.e. 2%). The orange curve shows the expected frequency for each bin if the returns were coming from a normal distribution with same mean and variance as the observations. Overall, the figure suggests that both the monthly rates of return and the monthly log-returns fit the normal distribution quite well. Section 3.7 discuss the use of historical returns more thoroughly.

3.3 Multivariate random variables, covariances, and correlations

When investing in multiple financial assets, it is important to what extent their returns move together. By investing in two assets whose prices tend to move in opposite directions, we can diversify (reduce) risk as we shall see in subsequent chapters.

Given two random variables X_1 and X_2 , the pair or “vector” (X_1, X_2) is said to be a two-dimensional or a bivariate random variable. For example, X_1 could represent the rate of return on one stock and X_2 the rate of return on another stock. Similarly we can define random variables of any higher integer dimension.

For a two-dimensional random variable (X_1, X_2) , the *joint cumulative distribution function* is the function $F_{(X_1, X_2)}$ defined by

$$F_{(X_1, X_2)}(x_1, x_2) = \text{Prob}(X_1 \leq x_1, X_2 \leq x_2). \quad (3.37)$$

This is to be understood as the probability that both $X_1 \leq x_1$ and $X_2 \leq x_2$ are satisfied. If both X_1 and X_2 are discrete random variables, the vector random variable (X_1, X_2) is also discrete, and the joint probability distribution is characterized by probabilities $\text{Prob}(X_1 = x_1, X_2 = x_2)$. If X_1 and X_2 are continuous, so is the pair (X_1, X_2) and formally

$$F_{(X_1, X_2)}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{(X_1, X_2)}(y_1, y_2) dy_1 dy_2,$$

where $f_{(X_1, X_2)}$ is then called the joint probability density function of (X_1, X_2) .

Given the joint distribution of (X_1, X_2) , we can find the distributions of X_1 and X_2 ,

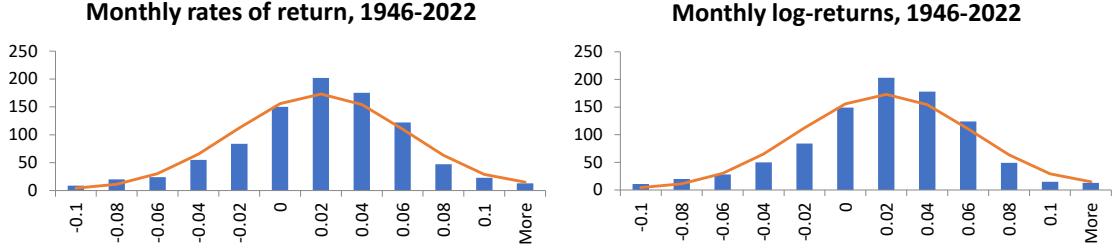


Figure 3.6: Distribution of observed S&P 500 returns.

The bars show the frequency in different bins of observed monthly returns on the S&P 500 stock index between January 1946 and December 2022. Data were downloaded from CRSP via WRDS on September 19, 2023. The orange curve shows for each bin the expected number of observations if the returns were drawn from a normal distribution with a mean and variance identical to the sample mean and variance of the observed returns. The left panel shows rates of return and the right panel log-returns.

the so-called *marginal distributions*. For example, if (X_1, X_2) is continuous with joint probability density function $f_{(X_1, X_2)}$, we can find the marginal probability density function of X_1 by integrating over all possible values of X_2 , that is

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_{(X_1, X_2)}(x_1, x_2) dx_2.$$

Two random variables X_1 and X_2 are said to be *independent* if

$$\text{Prob}(X_1 \in B_1, X_2 \in B_2) = \text{Prob}(X_1 \in B_1) \text{ Prob}(X_2 \in B_2)$$

for all sets $B_1, B_2 \subseteq \mathbb{R}$ or, equivalently, if

$$F_{(X_1, X_2)}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2)$$

for all pairs of possible values (x_1, x_2) . Intuitively this means that the probability that X_1 takes on any specific value is unaffected by which value X_2 assumes.

We can extend the definition of the expected value to functions of multi-dimensional random variables. If (X_1, X_2) is a two-dimensional continuous random variable and $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, the expected value of $g(X_1, X_2)$ is defined as

$$\mathbb{E}[g(X_1, X_2)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x_1, x_2) f_{(X_1, X_2)}(x_1, x_2) dx_1 dx_2.$$

Analogously, if (X_1, X_2) is discrete, the expectation is given by a double sum.

We define the *covariance* between X_1 and X_2 by

$$\text{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2], \quad (3.38)$$

where the last equality can be shown mathematically so that we have two alternative ways of computing the covariance as will be illustrated in the example below. Note that, in particular,

$$\text{Cov}[X_1, X_1] = \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 = \text{Var}[X_1].$$

If X_2 equals a constant a , we get $\text{Cov}[X_1, a] = 0$. Also observe that

$$\text{Cov}[X_1, X_2] = \text{Cov}[X_2, X_1]. \quad (3.39)$$

If X_1 and X_2 are discrete random variables, the covariance can be calculated by summation:

$$\begin{aligned} \text{Cov}[X_1, X_2] &= \sum_{s=1}^S p_s (X_{1s} - \text{E}[X_1])(X_{2s} - \text{E}[X_2]) \\ &= \left(\sum_{s=1}^S p_s X_{1s} X_{2s} \right) - \left(\sum_{s=1}^S p_s X_{1s} \right) \left(\sum_{s=1}^S p_s X_{2s} \right). \end{aligned} \quad (3.40)$$

If X_1 and X_2 are continuous random variables, the summation is replaced by integration.

The *correlation* between X_1 and X_2 is

$$\text{Corr}[X_1, X_2] = \frac{\text{Cov}[X_1, X_2]}{\text{Std}[X_1] \times \text{Std}[X_2]}, \quad (3.41)$$

which is a number in the interval $[-1,1]$. The correlation measures the extent to which there is a linear relation between the two random variables. If $X_2 = aX_1 + b$ for some constants a and b , then $\text{Corr}[X_1, X_2] = 1$ if a is positive and $\text{Corr}[X_1, X_2] = -1$ if a is negative, and then X_1 and X_2 are said to be perfectly positively or perfectly negatively correlated, respectively.

If X_1, X_2 are independent random variables, then $\text{E}[f(X_1)g(X_2)] = \text{E}[f(X_1)]\text{E}[g(X_2)]$ for any two functions f and g . In particular,

$$\text{E}[X_1 X_2] = \text{E}[X_1]\text{E}[X_2],$$

which implies that

$$\text{Cov}[X_1, X_2] = 0, \quad \text{Corr}[X_1, X_2] = 0.$$

Hence, independent random variables are uncorrelated. The converse is generally not true. Even if two random variables have a zero correlation, they need not be independent. However, if (X_1, X_2) follows a bivariate normal distribution with $\text{Corr}[X_1, X_2] = 0$, then X_1 and X_2 are independent.

Example 3.4

Example 3.1 considered the stock XYZ having five possible rates of return. Suppose we are also interested in the stock ABC and want to compute the covariance and correlation between the stocks. Table 3.4 presents the necessary information. Again subscript s refers to state number s . Moreover, r_{1s} and r_{2s} denote the rates of return on XYZ and ABC, respectively. For example, in state 2 (labelled ‘bad’ in Example 3.1) XYZ gives a return of -8% and ABC a return of 5% . The return on ABC is fairly large when the return on XYZ is either very low or very high but, on the other hand, low (negative) when the return on XYZ is moderate. Each entry in the row labeled ‘Exp’ (for expectation) is computed using a “sumproduct” of the column above that entry and the column with the probabilities.

The table shows that ABC has an expected rate of return of 0.06 or 6% , a return variance of 0.0259 or $259(\%)^2$, and thus a standard deviation of $\sqrt{0.0259} = 0.161$ or 16.1% .

p_s	r_{1s}	r_{2s}	$r_{1s} \times r_{2s}$	$r_{1s} - E[r_1]$	$r_{2s} - E[r_2]$	$\frac{(r_{1s} - E[r_1])}{\times(r_{2s} - E[r_2])}$	$(r_{1s} - E[r_1])^2$	$(r_{2s} - E[r_2])^2$
0.1	-0.2	0.1	-0.02	-0.25	0.04	-0.01	0.0625	0.0016
0.2	-0.08	0.05	-0.004	-0.13	-0.01	0.0013	0.0169	0.0001
0.4	0.07	-0.1	-0.007	0.02	-0.16	-0.0032	0.0004	0.0256
0.2	0.15	0.2	0.03	0.1	0.14	0.014	0.01	0.0196
0.1	0.28	0.4	0.112	0.23	0.34	0.0782	0.0529	0.1156
Exp	0.05	0.06	0.0116			0.0086	0.01708	0.0259

Table 3.4: Returns and probabilities.

Information relevant for calculating the expectations, variances, and covariance in Example 3.4.

The table also shows that, using the first sum in Eq. (3.40), the covariance between the two rates of return is 0.0086 or 86(%)². Alternatively, using the second sum in Eq. (3.40), first compute the expectation $E[r_1 r_2]$ of the product, which is the 0.0116 appearing in the bottom row, and then subtract the product of the expectations, i.e.,

$$\text{Cov}[r_1, r_2] = E[r_1 r_2] - E[r_1] E[r_2] = 0.0116 - 0.05 \times 0.06 = 0.0086.$$

In any case, the correlation is then computed using Eq. (3.41):

$$\text{Corr}[r_1, r_2] = \frac{0.0086}{0.131 \times 0.161} = 0.409$$

(again, results are calculated without rounding the input numbers). The dots in Figure 3.7 represent the five possible pairs of returns. The diagram also shows the best straight line relating the returns on the two stocks. The positive slope of the line means that the correlation is positive. Since the dots are not located on the line, the correlation is not perfect. Note that the numerical value of the correlation is not the slope of the best straight line, but measures how close the possible realizations (the dots) are to the line.

A special two-dimensional distribution is the *bivariate normal distribution*. The pair (X_1, X_2) is said to be bivariate normally distributed with means μ_1, μ_2 , standard deviations σ_1, σ_2 , and correlation ρ , if the probability density function is given by

$$f_{(X_1, X_2)}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right\} \quad (3.42)$$

for any real numbers x_1 and x_2 . It can be shown that (X_1, X_2) follows a bivariate normal distribution if and only if every linear combination $a_1 X_1 + a_2 X_2$ follows a (one-dimensional or univariate) normal distribution. In particular, if (X_1, X_2) follows a bivariate normal distribution with means μ_1, μ_2 , standard deviations σ_1, σ_2 , and correlation ρ , then $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and $\text{Corr}[X_1, X_2] = \rho$. We can also define higher-dimensional normal distributions, but the corresponding probability density function is then a very complicated expression, at least if you do not use vector and matrix notation.

The assumption that the set (X_1, X_2, \dots, X_N) of random variables follows a multivariate

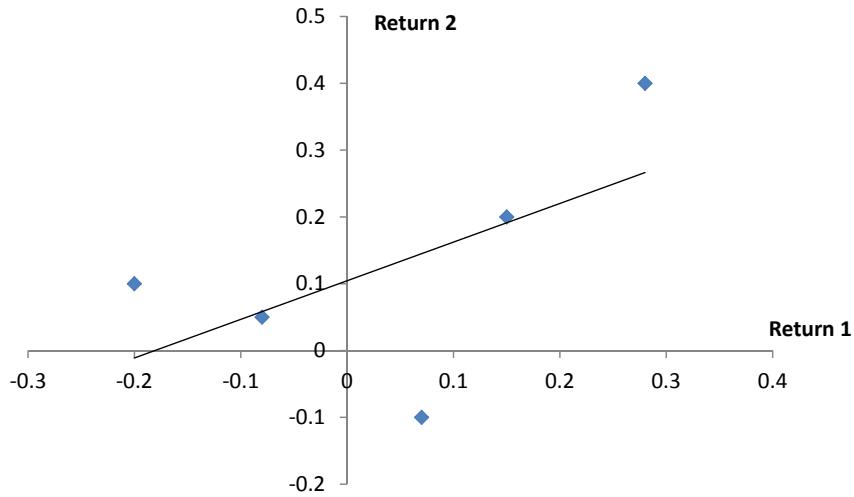


Figure 3.7: Return pairs.

The graph shows the five possible pairs of returns and the best straight line in Example 3.4.

normal distribution is often stated as X_1, X_2, \dots, X_N are *jointly normally distributed*. Note that such an assumption is stronger than just assuming that each of the random variables X_1, X_2, \dots, X_N follows a one-dimensional normal distribution.

3.4 Computational rules for random variables

We are often interested in transformations (functions) of random variables. Recall that Eq. (2.20) states that the rate of return on a buy-and-hold portfolio of two assets is $r_p = wr_1 + (1 - w)r_2$, where r_1 and r_2 are the rates of return on the two assets, and w is the portfolio weight of asset 1. For example, if you invest 60% in stock 1 with a rate of return r_1 and 40% in stock 2 with a rate of return r_2 , then the rate of return on your total position is

$$r_p = 0.6 \times r_1 + 0.4 \times r_2.$$

This is a transformation of the random variables r_1 and r_2 . More generally, the rate of return on an N -asset portfolio is given by $r_p = \sum_{i=1}^N \pi_i r_i$, cf. Eq. (2.22). We are often interested in the probability distribution and key moments of the portfolio return r_p , and how they are related to the distribution and moments of the assets in the portfolio. Sometimes powers, products, ratios, or even more complicated transformations of random variables turn out to be relevant.

3.4.1 Moments

The next theorem states rules for moments of linear transformations of random variables.

Theorem 3.2

Let X_1, X_2, \dots, X_N, Y denote random variables, and let a_1, a_2, \dots, a_N and b_1, b_2, \dots, b_N be real numbers. Then

$$\mathbb{E} \left[\sum_{i=1}^N a_i X_i \right] = \sum_{i=1}^N a_i \mathbb{E}[X_i], \quad (3.43)$$

$$\begin{aligned} \text{Var} \left[\sum_{i=1}^N a_i X_i \right] &= \sum_{i=1}^N \sum_{j=1}^N a_i a_j \text{Cov}[X_i, X_j] \\ &= \sum_{i=1}^N a_i^2 \text{Var}[X_i] + 2 \sum_{i=1}^N \sum_{j=i+1}^N a_i a_j \text{Cov}[X_i, X_j], \end{aligned} \quad (3.44)$$

$$\text{Cov} \left[\sum_{i=1}^N a_i X_i, Y \right] = \sum_{i=1}^N a_i \text{Cov}[X_i, Y], \quad (3.45)$$

$$\text{Cov} \left[\sum_{i=1}^N a_i X_i, \sum_{j=1}^N b_j X_j \right] = \sum_{i=1}^N \sum_{j=1}^N a_i b_j \text{Cov}[X_i, X_j]. \quad (3.46)$$

These rules are clearly useful for portfolio returns, and we explore this in detail in Chapter 4. For skewness and kurtosis

In the special case of two random variables X_1 and X_2 and numbers a_1, a_2, b_1 , and b_2 , we have

$$\mathbb{E}[a_1 X_1 + a_2 X_2] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2], \quad (3.47)$$

$$\text{Var}[a_1 X_1 + a_2 X_2] = a_1^2 \text{Var}[X_1] + a_2^2 \text{Var}[X_2] + 2a_1 a_2 \text{Cov}[X_1, X_2], \quad (3.48)$$

$$\begin{aligned} \text{Cov}[a_1 X_1 + a_2 X_2, b_1 X_1 + b_2 X_2] &= a_1 b_1 \text{Var}[X_1] + a_2 b_2 \text{Var}[X_2] \\ &\quad + (a_1 b_2 + a_2 b_1) \text{Cov}[X_1, X_2], \end{aligned} \quad (3.49)$$

which are useful results when studying portfolios of two risky assets. As an even more special case, for a random variable X and real numbers a and b , we have

$$\mathbb{E}[aX + b] = a \mathbb{E}[X] + b, \quad \text{Var}[aX + b] = a^2 \text{Var}[X], \quad \text{Std}[aX + b] = |a| \text{Std}[X], \quad (3.50)$$

where $|a|$ denotes the absolute value of a . The rules in (3.50) are useful for a portfolio of a risky asset and a riskfree asset. The next two examples provide some useful insights. The first example confirms that leverage increases the risk of an investment.

Example 3.5

Recall from Eq. (2.17) in Section 2.7 that the return on a leveraged investment in a stock is

$$r = r_{\text{stock}} + \frac{L_0}{E_0} (r_{\text{stock}} - r_{\text{loan}}) = \left(1 + \frac{L_0}{E_0}\right) r_{\text{stock}} - \frac{L_0}{E_0} r_{\text{loan}}, \quad (3.51)$$

where $E_0 > 0$ is the investor's own investment, $L_0 \geq 0$ is the borrowed amount, r_{loan} is the borrowing rate, and r_{stock} is the rate of return on the stock. Only the latter is a random

variable. Hence, the expected rate of return is

$$E[r] = E[r_{\text{stock}}] + \frac{L_0}{E_0} (E[r_{\text{stock}}] - r_{\text{loan}}) = \left(1 + \frac{L_0}{E_0}\right) E[r_{\text{stock}}] - \frac{L_0}{E_0} r_{\text{loan}}, \quad (3.52)$$

so by leveraging up you expect a larger rate of return provided the stock's expected return exceeds the borrowing rate. The leveraged rate of return has a variance and standard deviation of

$$\text{Var}[r] = \left(1 + \frac{L_0}{E_0}\right)^2 \text{Var}[r_{\text{stock}}], \quad \text{Std}[r] = \left(1 + \frac{L_0}{E_0}\right) \text{Std}[r_{\text{stock}}].$$

These formulas confirm that leverage increases the risk of the investment. In Exercise 3.8 you are asked to compare the Sharpe ratio of a leveraged position in a risky asset to the Sharpe ratio of the risky asset itself.

The second example illustrates why investors often use the Sharpe ratio to rank assets.

Example 3.6

Suppose you want to compare two assets. Asset 1 has both a higher expected return and a higher standard deviation than asset 2. Then it seems difficult to decide which asset is better to invest in. To facilitate a comparison, you can lever up asset 2 until you get the a standard deviation equal to that of asset 1. Formally, you construct a portfolio with a weight of $w = \text{Std}[r_1]/\text{Std}[r_2]$ in asset 2 and a weight of $1 - w$ in the riskfree asset. Given our assumption that $\text{Std}[r_1] > \text{Std}[r_2]$, we have $w > 1$ and thus $1 - w < 0$. The rate of return on the portfolio is

$$r_p = wr_2 + (1 - w)r_f.$$

Given the computational rules above, the expectation and the standard deviation of the portfolio return are

$$\begin{aligned} E[r_p] &= w E[r_2] + (1 - w)r_f = r_f + w (E[r_2] - r_f) = r_f + \frac{\text{Std}[r_1]}{\text{Std}[r_2]} (E[r_2] - r_f), \\ \text{Std}[r_p] &= w \text{Std}[r_2] = \text{Std}[r_1]. \end{aligned}$$

By construction, the portfolio has a standard deviation equal to that of asset 1, so now we can directly compare the expected returns of the portfolio and asset 1. We see that the portfolio has a larger expected return than asset 1 if

$$r_f + \frac{\text{Std}[r_1]}{\text{Std}[r_2]} (E[r_2] - r_f) > E[r_1] \quad \Leftrightarrow \quad \frac{E[r_2] - r_f}{\text{Std}[r_2]} > \frac{E[r_1] - r_f}{\text{Std}[r_1]},$$

i.e., if the Sharpe ratio is larger for asset 2 than asset 1. Therefore, the Sharpe ratio is often used to rank assets.

Note that correlations do not satisfy the same relations that covariances do. For example, while covariances satisfy

$$\text{Cov}[a_1 X_1 + a_2 X_2, Y] = a_1 \text{Cov}[X_1, Y] + a_2 \text{Cov}[X_2, Y], \quad (3.53)$$

the same relations is not true for correlations where we generally have

$$\text{Corr}[a_1 X_1 + a_2 X_2, Y] \neq a_1 \text{Corr}[X_1, Y] + a_2 \text{Corr}[X_2, Y] \quad (3.54)$$

with an equality only in very special cases.

Finally, consider the case where X_1, \dots, X_n are independent random variables that are identically distributed each with a mean of μ , a standard deviation of σ , a skewness of S , and a kurtosis of K . Then we have

$$\mathbb{E} \left[\sum_{i=1}^n X_i \right] = n\mu, \quad \text{Std} \left[\sum_{i=1}^n X_i \right] = \sqrt{n}\sigma \quad (3.55)$$

which follows from (3.43) and (3.44) letting all $a_i = 1$ and noting that all covariances are zero due to the assumed independence of X_1, \dots, X_n . For the skewness and kurtosis, it can be shown that

$$\text{Skew} \left[\sum_{i=1}^n X_i \right] = \frac{S}{\sqrt{n}}, \quad \text{Kurt} \left[\sum_{i=1}^n X_i \right] = \frac{K}{n}. \quad (3.56)$$

3.4.2 Distribution

If we define a random variable as a function of one random variable $Y = g(X)$ or several random variables $Y = g(X_1, \dots, X_n)$, the distribution of Y can be computed from the distribution of X or X_1, \dots, X_n . In general Y will follow a different type of distribution than X , but there are some important exceptions that are highly relevant for portfolio applications:

Theorem 3.3

Linear transformations preserve normality of probability distributions. More precisely, we have

- (a) If X is a normally distributed random variable, then any linear transformation $Y = aX + b$ where a and b are numbers (constants) is also normally distributed.
- (b) If X_1, \dots, X_n are jointly normally distributed random variables and a_1, \dots, a_n are numbers, then $Y = a_1 X_1 + \dots + a_n X_n$ is also normally distributed.

These results have the following implications for the lognormal distribution:

- (c) If X is a lognormally distributed random variable, then any power $Y = X^a$ where a is a number is also lognormally distributed.
- (d) If X_1, \dots, X_n are jointly lognormally distributed, then the product $Y = X_1 \times \dots \times X_n$ is also lognormally distributed.

These facts have implications for portfolio returns that are important enough to state in a theorem. Recall that the rate of return on a portfolio has the form $r_p = \sum_{i=1}^N \pi_i r_i$, where the portfolio weights π_i are known numbers and the asset returns r_i are random variables. The log-return on the portfolio is $r_p^{\log} = \ln(1 + r_p)$.

Theorem 3.4

This theorem considers buy-and-hold portfolios of N assets.

- (a) If the rates of return of the N assets are jointly normally distributed, then the rate of return on a buy-and-hold portfolio of the N assets is normally distributed.
- (b) If the log-returns of the N assets are jointly normally distributed, then the log-return on a buy-and-hold portfolio of the N assets is *not* normally distributed.

To understand part (b) of Theorem 3.4, consider a two-asset portfolio with gross return $R_p = wR_1 + (1 - w)R_2$. Then the log-return on the portfolio is $\ln R_p$, but since

$$\ln(wR_1 + (1 - w)R_2) \neq w\ln R_1 + (1 - w)\ln R_2,$$

the log-return on the portfolio is not a linear transformation of the log-returns on the two assets in the portfolio. If the log-returns of the assets are normally distributed, then the log-return of the portfolio is not normally distributed.

Over a very short period of time the gross returns both on the portfolio and the individual assets are close to 1 so the approximation $\ln R \approx R - 1$ is quite precise. Applying this approximation twice, we get

$$\begin{aligned}\ln R_p &= \ln(wR_1 + (1 - w)R_2) \approx wR_1 + (1 - w)R_2 - 1 \\ &= w(R_1 - 1) + (1 - w)(R_2 - 1) \approx w\ln R_1 + (1 - w)\ln R_2,\end{aligned}$$

so over very short horizons the log-return on the portfolio is close to the weighted average of the log-returns on the assets in the portfolio. If the assets' log-returns over a short horizon are normally distributed, then the log-return on the portfolio is close to being normal. The shorter the horizon, the better the approximation. We could rebalance the portfolio continuously (in reality, this means quite frequently) so that asset 1 will always have a weight of w . If the log-returns of both assets are normally distributed over any period, then the log-return of the continuously rebalanced, constant-weight portfolio is also normally distributed. This is discussed further in Section 8.1.1.

Theorem 3.3 has also implications for the compounding of returns which we explore in Section 3.6.

3.5 Tail risk measures

Many investors are particularly concerned about the left tail of return distributions as it represents the largest possible losses. Consequently, various popular risk measures focus on the left tail. Below we describe the two pre-dominant tail risk measures.

3.5.1 Value at risk

The most important of the tail risk measures is the value at risk, often abbreviated VaR. The value at risk is defined as the maximum loss on an asset or portfolio over a certain time period T with a certain probability p . The loss exceeds the value at risk only with a probability of p . The loss—and therefore also the value at risk—can be quantified either in monetary units (e.g., dollars) or as a percentage of the initial value of the asset or portfolio. The time period is usually one day, one week, two weeks, or a month, but the VaR can be computed for any time horizon. The probability is often chosen to be either 1% or 5%, but other values are also used. Sometimes $1 - p$ is said to be the confidence

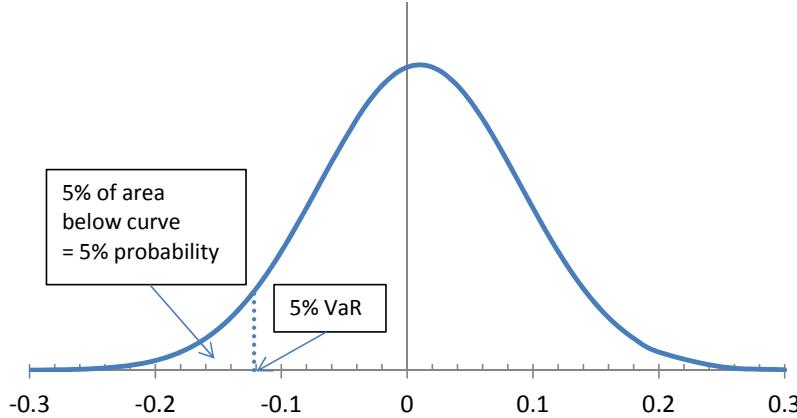


Figure 3.8: Value at risk.

The curve is the probability density function for a normal distribution with mean 0.01 and standard deviation 0.08. In this case the 5% value at risk is -0.1216 or -12.16% .

level because with that probability you lose less than the value at risk. For $p = 0.05$ we will use the term 5% value at risk, but others emphasize the confidence level and call it the 95% value at risk. Similarly for other values of p .

As an example, Figure 3.8 illustrates the 5% value at risk for a normally distributed rate of return, which is assumed to have a mean of 0.01 or 1% and a standard deviation of 0.08 or 8% as might be reasonable for a one-month horizon. Here the value at risk turns out to be -0.1216 ; we explain how to compute this below. Over the given period, there is only a 5% probability that the rate of return is more negative than -12.16% .

The value at risk plays a crucial role in the regulations of the risk-taking of financial institutions, which are required to maintain a capital buffer larger than, for example, the 5% value at risk over a one-month interval so that institution can cover the potential losses with 95% probability. Of course, this implies that there is a 5% risk that the loss exceeds the value at risk, which can be expected to happen in one out of 20 months.

Let V denote the value of the asset or portfolio. We can then define the value at risk in monetary units at time 0 for a time horizon of T as the smallest value of v for which

$$\text{Prob}(V_T - V_0 \leq v) = p.$$

If V_T follows a continuous distribution, as for example the normal distribution, there is exactly one value of v that satisfies this, so the value-at-risk is defined through the equation

$$\text{Prob}\left(V_T - V_0 \leq \text{VaR}^{\$}(p, T)\right) = p. \quad (3.57)$$

In other words, the value at risk $\text{VaR}^{\$}(p, T)$ is thus equivalent to the $100 \times p\%$ percentile in the distribution of the gain $V_T - V_0$. The corresponding value at risk in percentage terms is the smallest value of w for which

$$\text{Prob}\left(\frac{V_T - V_0}{V_0} \leq w\right) = p.$$

Again, with a continuous distribution, this implies that

$$\text{Prob} \left(\frac{V_T - V_0}{V_0} \leq \text{VaR}^{\%}(p, T) \right) = p, \quad (3.58)$$

so the percentage value-at-risk is equivalent to the $100 \times p\%$ percentile in the distribution of the rate of return $(V_T - V_0)/V_0$. The percentage value of risk equals the ratio of the monetary value at risk to the initial value,

$$\text{VaR}^{\%}(p, T) = \frac{\text{VaR}^{\$}(p, T)}{V_0}. \quad (3.59)$$

Often the probability level p and the time horizon T are clear from the context and are notationally omitted, and it is typically obvious whether a reported value at risk is in monetary units or percentage terms.

If the value V_T of the asset or portfolio is normally distributed with a known mean and known standard deviation, the value at risk is easy to compute. If $x \sim N(m, s^2)$, then $(x - m)/s \sim N(0, 1)$, cf. (3.25), so

$$\text{Prob}(x \leq v) = \text{Prob} \left(\frac{x - m}{s} \leq \frac{v - m}{s} \right) = N \left(\frac{v - m}{s} \right),$$

where N is the cumulative distribution function of the standard $N(0,1)$ normal distribution. Assume that $V_T \sim N(V_0 + \mu_V, \sigma_V^2)$, so that $V_T - V_0 \sim N(\mu_V, \sigma_V^2)$ and μ_V is the expected dollar gain. Then the value at risk satisfies

$$p = \text{Prob} \left(V_T - V_0 \leq \text{VaR}^{\$}(p, T) \right) = N \left(\frac{\text{VaR}^{\$}(p, T) - \mu_V}{\sigma_V} \right),$$

which implies that

$$\text{VaR}^{\$}(p, T) = \mu_V + \sigma_V \times N^{-1}(p), \quad (3.60)$$

where $N^{-1}(p)$ is the $100p\%$ percentile in the standard normal distribution, cf. Table 3.2. For example, the 5% value at risk is

$$\text{VaR}^{\$}(5\%, T) = \mu_V - 1.645 \sigma_V. \quad (3.61)$$

Note that Eq. (3.60) does not clearly show the role of the time horizon as T is not appearing on the right-hand side. Typically, the standard deviation increases with T and so does the expected gain provided that the expected return is positive.

If the time horizon is short, the expected gain μ_V in the value is small and often neglected, and then the value at risk is

$$\text{VaR}^{\$}(p, T) = \sigma_V \times N^{-1}(p). \quad (3.62)$$

Ignoring the expected change, the 5% value at risk therefore equals

$$\text{VaR}^{\$}(0.05, T) = -1.645 \sigma_V. \quad (3.63)$$

Similar expressions for the percentage value at risk exist if the rate of return on the asset or portfolio over the period is normally distributed. Assuming that the rate of

return $(V_T - V_0)/V_0 \sim N(\mu_r, \sigma_r^2)$, the percentage value at risk is

$$\text{VaR}^{\%}(p, T) = \mu_r + \sigma_r \times N^{-1}(p) \approx \sigma_r \times N^{-1}(p). \quad (3.64)$$

Example 3.7

Suppose you own a portfolio currently worth \$1,000,000. You believe that the portfolio's rate of return over the next month is normally distributed with a mean of 1% and a standard deviation of 8%. The 5% value at risk is then

$$\text{VaR}^{\%} = 0.01 - 0.08 \times 1.645 = -0.1216 = -12.16\%,$$

or -13.16% if the mean is neglected. The mean change in the portfolio value is 1% of \$1,000,000, that is \$10,000, and the standard deviation of the portfolio value is 8% or \$80,000. In monetary terms the value at risk is thus

$$\text{VaR}^{\$} = \$10,000 - \$80,000 \times 1.645 = -\$121,600,$$

which is exactly the percentage value at risk multiplied by the current portfolio value.

Note that with our above definition, the value at risk is negative, at least for the low p -values typically used. Often the negative sign is dropped and the value at risk is reported as a positive number in line with the notion of a maximum loss. In the preceding example the value at risk would then be quoted as \$121,600 or 12.16%.

For a normal distribution the value at risk is proportional to the standard deviation (if we ignore the mean) and therefore it does not contain any extra information. As an additional risk measure, the value at risk is therefore mostly relevant for distributions with a left tail that differs markedly from that of the normal distribution. Other things equal, the value at risk for the typically used probability levels is larger (to be precise: more negative) for distributions with a fatter left tail. Recall Figure 3.3 which shows a distribution with a positive kurtosis (that is, fat tails) and a distribution with a negative kurtosis (slim tails), together with the normal distribution. By construction, the three distributions have identical standard deviations. Panel (b) of this figure clearly shows that the value at risk is largest for the fat-tailed distribution and smallest for the slim-tailed distribution.

For most non-normal distributions a nice closed-form expression for the value at risk is not available, unfortunately. In cases where the probability density function is a well-known function, a large number of samples of the distribution can be simulated and then the relevant percentile can be estimated as the corresponding cut-off in the distribution of sampled values. For example, the 5% value at risk is then simply the simulated value that exceeds exactly 5% of all simulated values. This is the so-called *Monte Carlo simulation approach*. Practitioners often compute the value at risk by the so-called *historical simulation approach*: they find the appropriate cut-off in a distribution of realized values of, say, the return on the asset in a number of previous periods.

3.5.2 Expected shortfall

The 5% value at risk is silent about the magnitude of the losses in the 5% of all cases where the loss exceeds the value at risk. You can easily have two probability distributions

with the same 5% percentile, but where the distributions to the left of that percentile differ (see the example below). The magnitude of the extreme losses also matters. Therefore, as a supplement to the value at risk, the expected shortfall is often used. The expected shortfall is defined as the expected loss conditional on the loss exceeding the value at risk. Like the value at risk, we can express the expected shortfall either as a percent of the current value, denoted by $ES^{\%}$, or as an amount in the given currency, denoted by $ES^{\$}$. By definition, the expected shortfall is larger than the value at risk (in absolute value terms). Some use the terms *conditional value at risk* or *expected tail loss* instead of expected shortfall.

Example 3.8

Consider the probability distributions in Table 3.5. For distribution 1 (see second column) there is a 1% risk of a 30% loss, a 2% risk of both a 20% and a 10% loss, whereas with a probability of 95%, you lose less than 10% or even make a profit. The 5% value at risk is therefore -10%. For distribution 2 (see fourth column) there is a 2% risk of both a 50% loss and a 40% loss, a 1% risk of a 10% loss, whereas with 95% probability you lose less than 10% or you make a profit. Also for this distribution, the 5% value at risk is thus -10%.

While the two return distributions have identical 5% value at risk, the losses in these worst 5% of all cases are different. To calculate the expected shortfall, we need the conditional probabilities of the extreme losses. For distribution 1, the conditional probability of a return of -30% given that the return is lower than or equal to the value at risk (-10%) is $0.01/0.05 = 0.2$. For a return of -20% the conditional probability of a return of -20% is $0.02/0.05 = 0.4$. The same is true for a return of -10%. The expected shortfall for distribution 1 is thus

$$ES_1^{\%} = 0.2 \times (-30\%) + 0.4 \times (-20\%) + 0.4 \times (-10\%) = -18\%.$$

For distribution 2 similar considerations lead to an expected shortfall of

$$ES_2^{\%} = 0.4 \times (-50\%) + 0.4 \times (-40\%) + 0.2 \times (-10\%) = -38\%.$$

The two distributions have markedly different expected shortfalls.

If V still represents the value of the asset or portfolio, then the expected value shortfall at a probability level of p over a horizon of T can be formally defined as the conditional expectation

$$ES^{\$}(p, T) = E \left[V_T - V_0 \mid V_T - V_0 \leq \text{VaR}^{\$}(p, T) \right]. \quad (3.65)$$

If $V_T \sim N(V_0 + \mu_V, \sigma_V^2)$, the expected shortfall can be shown to equal

$$ES^{\$}(p, T) = \mu_V - \frac{\sigma_V}{p\sqrt{2\pi}} e^{-[N^{-1}(p)]^2/2}. \quad (3.66)$$

A similar expression holds for the expected shortfall in percent if the rate of return is normally distributed:

$$ES^{\%}(p, T) = \mu_r - \frac{\sigma_r}{p\sqrt{2\pi}} e^{-[N^{-1}(p)]^2/2}.$$

Return	Distribution 1		Distribution 2	
	Probability	Conditional	Probability	Conditional
-50%	0.00	0.00	0.02	0.40
-40%	0.00	0.00	0.02	0.40
-30%	0.01	0.20	0.00	0.00
-20%	0.02	0.40	0.00	0.00
-10%	0.02	0.40	0.01	0.20
Larger than -10%	0.95		0.95	
Var [%]	-10%		-10%	
ES [%]	-18%		-38%	

Table 3.5: Calculation of the value at risk and the expected shortfall.

The table lists the probability distributions used in Example 3.8.

For most non-normal distributions such a nice closed-form solution is unavailable so that the expected shortfall is computed as in the above example but based either on a set of simulated or historical values.

3.6 Return distributions and the investment horizon

Suppose we make some assumption about the distribution or some moments of monthly returns. What can we then say about the distribution and moments of returns over many months? What is the expected return, the standard deviation of returns, the tail risks, etc. for long-term investments? For such considerations, log-returns are more tractable since the log-return over T periods is just the sum of the log-returns in the different periods.

3.6.1 Distribution and moments of multi-period log-returns

The log-return over two periods is the sum of the log-returns over each of the periods,

$$r_{0,2}^{\log} = r_{0,1}^{\log} + r_{1,2}^{\log}.$$

This is just a simple special case of Eq. (2.27).

Suppose that the log-return over each period has the same mean m and the same variance s^2 , i.e., for every t we have

$$\mathbb{E}[r_{t,t+1}^{\log}] = m, \quad \text{Var}[r_{t,t+1}^{\log}] = s^2.$$

Then the expectation of the log-return over the two periods is obviously just twice the expected log-return over a single period since

$$\mathbb{E}[r_{0,2}^{\log}] = \mathbb{E}\left[r_{0,1}^{\log} + r_{1,2}^{\log}\right] = \mathbb{E}\left[r_{0,1}^{\log}\right] + \mathbb{E}\left[r_{1,2}^{\log}\right] = m + m = 2m.$$

The variance is a bit more involved:

$$\begin{aligned} \text{Var}\left[r_{0,2}^{\log}\right] &= \text{Var}\left[r_{0,1}^{\log} + r_{1,2}^{\log}\right] = \text{Var}\left[r_{0,1}^{\log}\right] + \text{Var}\left[r_{1,2}^{\log}\right] + 2\text{Cov}\left[r_{0,1}^{\log}, r_{1,2}^{\log}\right] \\ &= s^2 + s^2 + 2\rho s s = 2(1 + \rho)s^2, \end{aligned} \tag{3.67}$$

where ρ is the first-order *serial correlation* or *autocorrelation*

$$\rho = \text{Corr} [r_{0,1}^{\log}, r_{1,2}^{\log}],$$

and we have applied Eq. (3.41). A zero serial correlation means that the distribution of the return next period is unaffected by the realized return in this period. If the returns in different periods are independent of each other, the serial correlation is indeed zero so that $\text{Var}[r_{0,2}^{\log}] = 2s^2$.

In most of what follows, we will assume that returns in different periods are independent so that there is no autocorrelation in returns. Section 3.6.5 discusses autocorrelation and its implications for multi-period return distributions.

If we assume returns are independent over time and calculate the Sharpe ratio based on log-returns, the one-period Sharpe ratio is

$$\text{SR}_1^{\log} = \frac{\text{E}[r_{t,t+1}^{\log}] - r_f}{\text{Std}[r_{t,t+1}^{\log}]} = \frac{m - r_f}{s},$$

where r_f is the periodic log-riskfree rate, also known as the continuously compounded riskfree rate. Given the above findings, the two-period Sharpe ratio is then

$$\text{SR}_2^{\log} = \frac{\text{E}[r_{0,2}^{\log}] - 2r_f}{\text{Std}[r_{0,2}^{\log}]} = \frac{2m - 2r_f}{\sqrt{2}s} = \sqrt{2} \text{ SR}_1.$$

With the two returns being independent, one can also show that the skewness of the two-period log-return equals the skewness of the periodic log-return divided by $\sqrt{2}$, and the kurtosis of the two-period log-return equals the kurtosis of the periodic log-return divided by 2. We can generalize the above results to T periods as shown in the following theorem.

Theorem 3.5

Suppose that the periodic log-returns have mean m , variance s^2 , skewness S , and kurtosis K . Then the expected T -period log-return is

$$\text{E} [r_{0,T}^{\log}] = Tm. \quad (3.68)$$

If, furthermore, the periodic log-returns are independent, the variance, standard deviation, skewness, and kurtosis are given by

$$\text{Var} [r_{0,T}^{\log}] = Ts^2, \quad (3.69)$$

$$\text{Std} [r_{0,T}^{\log}] = \sqrt{T}s, \quad (3.70)$$

$$\text{Skew} [r_{0,T}^{\log}] = S/\sqrt{T}, \quad (3.71)$$

$$\text{Kurt} [r_{0,T}^{\log}] = K/T, \quad (3.72)$$

and the Sharpe ratio based on log-returns satisfies

$$\text{SR}_T^{\log} = \sqrt{T} \text{ SR}_1^{\log}. \quad (3.73)$$

Note that the expectation and variance increase proportionally with the horizon, whereas the standard deviation and the Sharpe ratio increase with the horizon at a square-root rate, still under the assumption of no serial correlation in returns. Here r_f is the one-period continuously compounded riskfree rate. When calculating the Sharpe ratios for different horizons in this way, we can see that if asset 1 has a larger one-period Sharpe ratio than asset 2, then the Sharpe ratio of asset 1 is larger than that of asset 2 for any horizon, at least if our assumption of independent or serially uncorrelated returns holds for both assets.

The expressions (3.68)–(3.70) show how expectations, variances, and standard deviations of log-returns can be annualized. For example, if m and s denote the expectation and standard deviation of the monthly log-return, then the annualized expectation, variance, and standard deviation are $12m$, $12s^2$ and $\sqrt{12}s \approx 3.5s$, respectively, again assuming that monthly log-returns are uncorrelated with each other.

3.6.2 Long-run returns with normally distributed short-run log-returns

The above considerations hold no matter what type of distribution the periodic log-returns follow. If we assume normally distributed log-returns, we can say a lot more about multi-period returns. Again, the multi-period log-return is the sum of the log-returns in the individual periods. If each of the periodic log-returns is normally distributed, then their sum is also normally distributed according to Theorem 3.3. Specifically, assume that the periodic log-returns are all normally distributed with mean m and variance s^2 and independent of each other. Then the log-return over T periods, $r_{0,T}^{\log}$, is normally distributed with mean Tm and variance Ts^2 . Based on this observation, we can apply Theorem 3.1 to characterize the distribution and moments of the T -period rate of return, $r_{0,T} = \exp\{r_{0,T}^{\log}\} - 1$. We summarize the results in the next theorem.

Theorem 3.6

Assume that the periodic log-returns $r_{0,1}^{\log}, \dots, r_{1,2}^{\log}, \dots, r_{T-1,T}^{\log}$ are all normally distributed with mean m and variance s^2 and independent of each other. Then the following holds:

- (a) The T -period log-return $r_{0,T}^{\log}$ is normally distributed with mean Tm and variance Ts^2 .
- (b) The cumulative distribution function for the T -period rate of return $r_{0,T}$ is

$$F_{r_{0,T}}(x) = N\left(\frac{\ln(1+x) - Tm}{s\sqrt{T}}\right), \quad (3.74)$$

and the percentiles are given by

$$p = \text{Prob}(r_{0,T} < k_p) \Leftrightarrow k_p = \exp\{\sqrt{T}sN^{-1}(p) + Tm\} - 1. \quad (3.75)$$

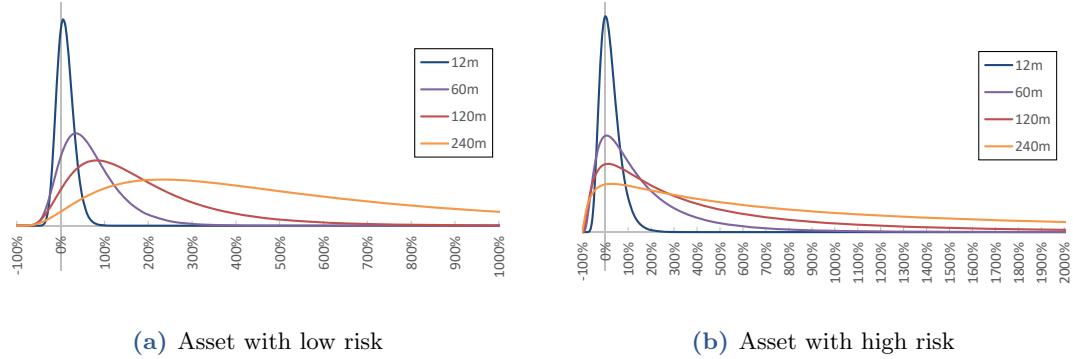


Figure 3.9: The rate of return distribution for different horizons.

Each panel shows the distribution of the rate of return over horizons of 12, 60, 120, and 240 months. All results assume that monthly log-returns are independent and normally distributed with mean m and s^2 . The left panel assumes $m = 0.005$ and $s^2 = 0.0025$ and is representative for a low-risk asset such as a stock market index. The right panel assumes $m = 0.001$ and $s^2 = 0.0105$ and is representative for a higher-risk asset such as an individual stock.

The key moments of $r_{0,T}$ are

$$\mathbb{E} [r_{0,T}] = e^{T(m + \frac{1}{2}s^2)} - 1, \quad (3.76)$$

$$\text{Med} [r_{0,T}] = e^{Tm} - 1, \quad (3.77)$$

$$\text{Var} [r_{0,T}] = e^{T(2m + \frac{1}{2}s^2)} (e^{Ts^2} - 1), \quad (3.78)$$

$$\text{Skew} [r_{0,T}] = (e^{Ts^2} + 2) \sqrt{e^{Ts^2} - 1}, \quad (3.79)$$

$$\text{Kurt} [r_{0,T}] = e^{4Ts^2} + 2e^{3Ts^2} + 3e^{2Ts^2} - 6. \quad (3.80)$$

Table 3.6 illustrates key moments for both log-returns and rates of return as well as selected percentiles for rates of return over various horizons under the assumption that the monthly log-returns are normally distributed with mean m and s^2 . The left part of the table assumes $m = 0.005$ and $s^2 = 0.0025$, whereas the right part assumes $m = 0.001$ and $s^2 = 0.0105$. In both cases, $m + \frac{1}{2}s^2 = 0.0625$ so the expected rate of return will be identical in the two case, no matter the investment horizon. The expected annual rate of return is $0.078 = 7.8\%$. In the left part, the standard deviation of the annual rate of return is $0.188 = 18.8\%$, which is in the range reasonable for a stock market index. In the right part, the standard deviation of the annual rate of return is $0.395 = 39.5\%$, a reasonable value for many individual stocks. Figure 3.9 depicts the probability distribution of the rate of return over horizons of 1, 5, 10, and 20 years.

The table shows that standard deviation of the rate of return grows a lot faster with the investment horizon for the high-risk asset than for the low-risk asset. While the standard deviation is about twice as big over a one-year horizon, it is almost four times as big over a 20-year horizon. Both the skewness and kurtosis of the rate of return distribution increase considerably with the horizon, especially for the high-risk asset.

The numbers in the row labeled ‘50 pct’ shows the median in the rate of return distri-

	Low risk (~ stock index)					High risk (~ individual stock)				
	1m	1y	5y	10y	20y	1m	1y	5y	10y	20y
Log-return										
Mean	0.005	0.060	0.300	0.600	1.200	0.001	0.012	0.060	0.120	0.240
Std	0.050	0.173	0.387	0.548	0.775	0.102	0.355	0.794	1.122	1.587
Riskfree	0.001	0.012	0.060	0.120	0.240	0.001	0.012	0.060	0.120	0.240
SR ^{log}	0.080	0.277	0.620	0.876	1.239	0.000	0.000	0.000	0.000	0.000
SR ^{log} / \sqrt{T}	0.080	0.080	0.080	0.080	0.080	0.000	0.000	0.000	0.000	0.000
Rate of return										
Mean	0.006	0.078	0.455	1.117	3.482	0.006	0.078	0.455	1.117	3.482
Std	0.050	0.188	0.585	1.252	4.064	0.103	0.395	1.363	3.364	15.15
Riskfree	0.001	0.012	0.062	0.127	0.271	0.001	0.012	0.062	0.127	0.271
SR	0.105	0.350	0.672	0.790	0.790	0.051	0.167	0.288	0.294	0.212
SR / \sqrt{T}	0.105	0.101	0.087	0.072	0.051	0.051	0.048	0.037	0.027	0.014
Skew	0.150	0.529	1.272	1.981	3.466	0.309	1.149	3.633	8.781	48.78
Kurt	0.040	0.501	3.008	7.706	27.08	0.171	2.434	30.24	273.4	28158
1 pct	-0.105	-0.290	-0.452	-0.490	-0.452	-0.211	-0.557	-0.832	-0.917	-0.968
5 pct	-0.074	-0.201	-0.286	-0.260	-0.071	-0.154	-0.436	-0.712	-0.822	-0.907
10 pct	-0.057	-0.150	-0.178	-0.097	0.230	-0.122	-0.358	-0.616	-0.732	-0.834
25 pct	-0.028	-0.055	0.040	0.259	0.969	-0.066	-0.203	-0.378	-0.471	-0.564
50 pct	0.005	0.062	0.350	0.822	2.320	0.001	0.012	0.062	0.127	0.271
75 pct	0.039	0.193	0.753	1.636	4.598	0.073	0.286	0.814	1.404	2.709
90 pct	0.072	0.326	1.217	2.676	7.959	0.141	0.595	1.936	3.752	8.722
95 pct	0.091	0.412	1.552	3.486	10.87	0.185	0.815	2.918	6.145	16.31
99 pct	0.129	0.589	2.323	5.516	19.13	0.270	1.311	5.729	14.35	50.06
$P(r < \text{mean})$	0.510	0.535	0.577	0.608	0.651	0.520	0.570	0.654	0.713	0.786
$P(r < \text{riskfree})$	0.468	0.391	0.268	0.190	0.108	0.500	0.500	0.500	0.500	0.500

Table 3.6: Return distributions for different horizons.

For various investment horizons, the upper panel lists key moments for log-returns and the lower panel both key moments and selected percentiles for rates of return. All results are based on the assumption that monthly log-returns are independent and normally distributed with mean m and s^2 . The left part of the table assumes $m = 0.005$ and $s^2 = 0.0025$ and is representative for an asset with a fairly low risk such as a stock market index. The right part assumes $m = 0.001$ and $s^2 = 0.0105$ and is representative for an asset with a fairly high risk such as an individual stock.

bution. The low and high percentiles for long horizons highlight the large skewness and kurtosis. For example, for the high-risk asset, there is a 10% probability that the rate of return is -83.4% or worse over a 20-year horizon. On the other hand, there is a 10% probability that the rate of return is 872.2% or higher. The second-to-last row shows the probability that the realized rate of return is lower than the expectation. Because of the skewed distribution, this differs from 50%. For long investment horizons, the probability that the return is lower than expected is substantially larger than 50%, especially for high-risk assets. In other words, it is very likely that the investor ends up disappointed. However, there is also a decent chance of getting a very high return.

Some of the numbers in the table involve a riskfree return. The monthly log-riskfree rate (also called the continuously compounded riskfree rate) is assumed to be 0.001. The log-riskfree rate over T months is then $T \times 0.001$. The riskfree rate of return over T periods is $\exp(T \times 0.001) - 1$. The last row of the table shows that with the assumed parameter values, there is a 50% chance that the high-risk asset is giving a higher rate of return than a riskfree investment over any horizon. In contrast, for an increasing horizon, it is getting less likely that the low-risk asset will underperform the riskfree asset. The Sharpe ratio based on the rate of return is first increasing with the horizon but at a slower rate than \sqrt{T} , and eventually the Sharpe ratio begins to decrease.

3.6.3 Moments of multi-period rates of return

Theorem 3.6 characterized the distribution and key moments of T -period rates of return when periodic log-returns are normally distributed. What can we say about the T -period rate of return under other distributional assumptions? Recall that the T -period rate of return is

$$r_{0,T} = (1 + r_{0,1})(1 + r_{1,2}) \dots (1 + r_{T-1,T}) - 1.$$

where $r_{t,t+1}$ is the periodic rate of return, i.e. in the period between t and $t + 1$. One idea is to assume that the periodic rates of return are jointly normally distributed. Then each of the terms $1 + r_{t,t+1}$ is also normally distributed. But the T -period rate of return $r_{0,T}$ is a product of such terms (and then the constant 1 is subtracted), and a product of normally distributed random variables is *not* normally distributed. However, no matter the distributional assumptions, as long as the periodic rates of return are independent of each other, we can still derive the expectation and the variance of $r_{0,T}$, as shown in the next theorem.

Theorem 3.7

- (a) If the periodic rates of return $r_{t,t+1}$ are normally distributed, then the T -period rate of return $r_{0,T}$ is not normally distributed.
- (b) If the periodic rates of return $r_{t,t+1}$ have mean μ and variance σ^2 and are independent of each other, then

$$\mathbb{E}[r_{0,T}] = (1 + \mu)^T - 1, \quad (3.81)$$

$$\text{Var}[r_{0,T}] = \left(\sigma^2 + (1 + \mu)^2\right)^T - (1 + \mu)^{2T}. \quad (3.82)$$

Proof

First, consider the expectation. Recall from Section 3.3 that when X_1 and X_2 are independent random variables, we have $E[X_1 X_2] = E[X_1] E[X_2]$. In our case, we get

$$E[r_{0,T}] = (1 + E[r_{0,1}]) (1 + E[r_{1,2}]) \dots (1 + E[r_{T-1,T}]) - 1,$$

and since $E[r_{0,1}] = E[r_{1,2}] = \dots = E[r_{T-1,T}] = \mu$, we get (3.81).

The calculation is trickier for the variance:

$$\begin{aligned} \text{Var}[r_{0,T}] &= \text{Var}[1 + r_{0,T}] = E[(1 + r_{0,T})^2] - (E[1 + r_{0,T}])^2 \\ &= E[(1 + r_{0,1})^2 \dots (1 + r_{T-1,T})^2] - ((1 + \mu)^T)^2 \\ &= E[(1 + r_{0,1})^2] \dots E[(1 + r_{T-1,T})^2] - (1 + \mu)^{2T} \\ &= (E[(1 + r_{t,t+1})^2])^T - (1 + \mu)^{2T} \\ &= (\text{Var}[1 + r_{t,t+1}] + (1 + E[r_{t,t+1}])^2)^T - (1 + \mu)^{2T} \\ &= (\sigma^2 + (1 + \mu)^2)^T - (1 + \mu)^{2T}. \end{aligned}$$

The expected T -period rate of return differs from $T\mu$ as we have to take the compounding of returns into account. Due to the compounding of returns, both the expectation and the variance of the rate of return grow more rapidly with the length of the investment period than the proportional growth seen for log-returns, cf. Eqs. (3.68) and (3.69). In particular, we can use Theorem 3.7 to annualize a monthly expected rate of return and variance as follows:

$$E[r_{\text{year}}] = (1 + E[r_{\text{month}}])^{12} - 1, \quad (3.83)$$

$$\text{Var}[r_{\text{year}}] = (\text{Var}[r_{\text{month}}] + (1 + E[r_{\text{month}}])^2)^{12} - (1 + E[r_{\text{month}}])^{24}, \quad (3.84)$$

where Eq. (3.83) is maybe what you have anticipated.

The left panel of Figure 3.10 shows how the annualized expected rate of return from Eq. (3.83) depends on the monthly expected rate of return. The dashed line is the approximation given by 12 times the monthly expected rate of return and thus ignoring the compounding of returns. The right panel of the figure does the same for the standard deviation. The red curves assume a 6% standard deviation of the monthly rate of return, whereas the green curves are based on 12%. The flat dashed lines represent the simple rule of multiplying by $\sqrt{12}$, which ignores compounding. The solid curves show the correctly annualized standard deviation obtained by taking square roots in (3.84). Note that the approximation can be quite far from the correct value unless the expected monthly rate of return is close to zero. For a well-diversified portfolio or a stock market index, the monthly rates of return could have a standard deviation of 0.06 and an expectation of 0.005. Then the annual rate of return has a standard deviation of 0.222 (compared to 0.208 when ignoring compounding) and an expectation of 0.0617 (compared to 0.06). An individual stock could have monthly rates of return with a standard deviation of 0.12 and an expectation of 0.008. Then the annual rate of return has a standard deviation of 0.472

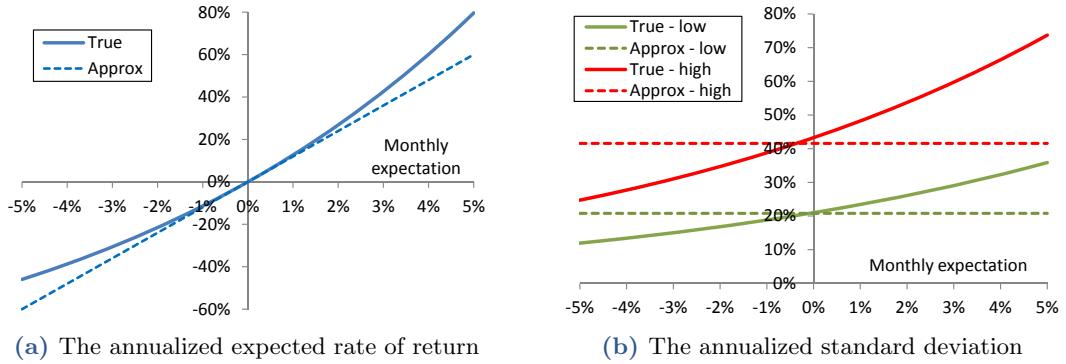


Figure 3.10: Annualization of expectation and standard deviation.

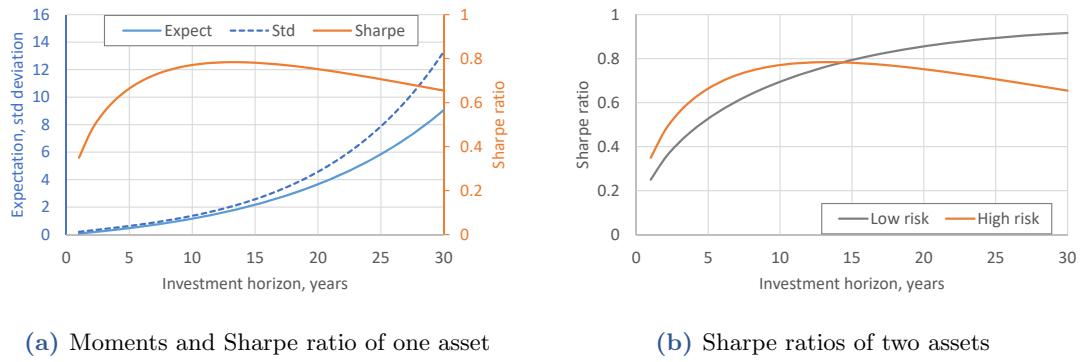
The left panel shows the annualized expected rate of return and the right panel the annualized standard deviation. Solid curves depict the values taking the compounding of returns into account, whereas the dashed curves are based on an approximation that ignores compounding. In the right panel the red curves are for the case of a monthly standard deviation of 6% and the green curves for 12%.

(compared to 0.416) and an expectation of 0.1003 (compared to 0.096).

With the above formulas for the expectation and variance of the multi-period rate of return, we can also consider how the Sharpe ratio

$$\text{SR}_{0,T} = \frac{\mathbb{E}[r_{0,T}] - [(1 + r_f)^T - 1]}{\text{Std}[r_{0,T}]} \quad (3.85)$$

depends on the length of the investment horizon, T . Here, r_f denotes the periodic riskfree rate of return. Suppose, for example, that annual returns have expectation 0.08 and standard deviation 0.20, roughly corresponding to a broad stock market index. Panel (a) of Figure 3.11 shows how the expectation and standard deviation of the rate of return increase with the length of the investment horizon. Assume a constant riskfree rate of return of $r_f = 0.01$ or 1% per year. Then we see that the Sharpe ratio is first increasing with the length of the investment horizon, but eventually the standard deviation grows so much faster than the expectation that the Sharpe ratio starts to decrease. This is in line with our observations from Table 3.6. Panel (b) of Figure 3.11 compares with another asset having annual returns with a lower standard deviation, 0.10, and a lower expectation, 0.035. For short horizons, the high-risk asset has a larger Sharpe ratio than the low-risk asset, but for longer horizons the opposite is true. The main driver of this result is that, due to the compounding of returns, the long-term return on an asset with a large short-term standard deviation becomes extremely risky, whereas the risk does not grow with the horizon as rapidly when the short-term standard deviation is low. This graph also shows that Sharpe ratios can be deceptive. One asset may seem more attractive than another asset if you compare their Sharpe ratios for a one-year horizon, but the other asset may appear best when comparing Sharpe ratios for a long investment horizon. Recall from Eq. (3.73) and the subsequent discussion that this mismatch is avoided when Sharpe ratios are calculated from log-returns.



(a) Moments and Sharpe ratio of one asset

(b) Sharpe ratios of two assets

Figure 3.11: Return moments and the investment horizon.

Annual returns are assumed independent. The left panel considers an asset, asset 1, with annual returns having an expectation of 0.08 or 8% and a standard deviation of 0.20 or 20%. The riskfree rate is 0.01 or 1% per year. The right panel compares the Sharpe ratios for different horizons of asset 1 (“high risk”) and asset 2 (“low risk”), where the annual returns on asset 2 have an expectation of 0.035 or 3.5% and a standard deviation of 0.10 or 10%.

3.6.4 Long-run returns with normally distributed short-run rates of return

Table 3.6 and Figure 3.9 showed higher-order moments and the probability distribution of multi-period rates of return under the assumption of normally distributed monthly log-returns. The skewness and kurtosis increase rapidly with the length of the investment horizon, especially for relatively risky assets. You might wonder whether this conclusion is special to the case of normally distributed log-returns. An obvious alternative assumption is to assume that the monthly rates of return are independent and normally distributed. Then the multi-period rates of return are not following a normal distribution, however, nor any other well-known probability distribution. Therefore, to study the properties of $r_{0,T}$ in this case, we employ *simulations* in which we draw many possible sequences or paths of periodic returns from the normal distribution and for each path we calculate $r_{0,T}$. By looking at the calculated values of $r_{0,T}$ for a large number of different paths, we obtain an approximation of the probability distribution.

Table 3.7 compares the mean, the standard deviation, and selected percentiles across the two assumptions, i.e. normally distributed monthly log-returns or normally distributed monthly rates of return. The results in the latter case are based on 500,000 simulated sequences of 240 monthly rates of return.⁴ The comparison is made both for an asset with relatively low risk and an asset with relatively high risk, as in Table 3.6. In each case the parameters μ and σ are derived from m and s such that the expectation and the variance of the monthly rate of return are the same whether the monthly log-returns are assumed $N(m, s^2)$ distributed or the monthly rates of return are assumed $N(\mu, \sigma^2)$ distributed. The overall conclusion from the table is that the long-run distribution of the rate of return—and in particular the mean and the standard deviation—are highly similar across the two cases. Also with normally distributed monthly rates of return, the long-run rates of return follow a highly skewed distribution with a long right tail. Due to the compounding of returns, the multi-period rate of return distribution is positively skewed

⁴Given the normality assumption, there is a tiny positive probability of a monthly rate of return being less than -1 , i.e. less than -100% , but none of the 120 million simulated monthly rates of return were that low.

	Low risk asset				High risk asset			
	12 months		240 months		12 months		240 months	
	$r_{\text{mon}}^{\log} \sim N$	$r_{\text{mon}} \sim N$						
Mean	0.078	0.077	3.482	3.476	0.078	0.077	3.482	3.469
Std dev	0.188	0.188	4.064	4.056	0.395	0.395	15.151	14.531
1%	-0.290	-0.295	-0.452	-0.461	-0.557	-0.574	-0.968	-0.972
5%	-0.201	-0.204	-0.071	-0.081	-0.436	-0.448	-0.907	-0.914
10%	-0.150	-0.151	0.230	0.219	-0.358	-0.366	-0.834	-0.845
25%	-0.055	-0.056	0.969	0.963	-0.203	-0.206	-0.564	-0.583
50%	0.062	0.063	2.320	2.317	0.012	0.015	0.271	0.242
75%	0.193	0.194	4.598	4.600	0.286	0.292	2.709	2.675
90%	0.326	0.325	7.959	7.963	0.595	0.597	8.722	8.731
95%	0.412	0.410	10.871	10.859	0.815	0.811	16.307	16.348
99%	0.589	0.582	19.125	19.134	1.311	1.281	50.058	50.513

Table 3.7: Moments and percentiles of multi-period rates of return.

For horizons of 12 and 240 months, the table lists the mean, the standard deviation, and selected percentiles for the rate of return both for the case where monthly log-returns are normally distributed, $r_{\text{mon}}^{\log} \sim N(m, s^2)$, and the case where monthly rates of return are normally distributed, $r_{\text{mon}} \sim N(\mu, \sigma^2)$. Here μ and σ^2 are chosen such that the expectation and the variance of the monthly rate of return are the same in the two cases. The left part of the table is representative for an asset with relatively low risk, where $m = 0.005, s^2 = 0.0025$ and $\mu \approx 0.00627, \sigma \approx 0.05034$. The right part is representative for an asset with relatively high risk, where $m = 0.010, s^2 = 0.0105$ and $\mu \approx 0.00627, \sigma \approx 0.10338$. The results based on normally distributed log-returns are determined analytically as explained in Section 3.6.2. The results based on normally distributed monthly rates of return are determined from 500,000 simulated sequences of 240 monthly returns.

and leptokurtic even though the single-period rate of return has zero skew and kurtosis.⁵

3.6.5 Autocorrelation in returns

The concept of autocorrelation in returns was briefly introduced in Section 3.6.1. Intuitively, a positive autocorrelation means that a positive return tends to be followed by another positive return, whereas a negative return tends to be followed by another negative return. This is often referred to as a *momentum* in return. Conversely, a negative autocorrelation represents a *mean reversion in prices* or *reversal in returns*: positive returns tend to be followed by negative returns and vice versa.

First, consider an autocorrelation in log-returns. As shown in the beginning of Section 3.6.1, the expected two-period log-return is twice the expected log-return per period, whether the log-returns are autocorrelated or not. But the variance is affected by the autocorrelation. If the standard deviation of the periodic log-return is s and the autocorrelation is ρ , then the variance of the two-period log-return is

$$\text{Var}[r_{0,2}^{\log}] = 2(1 + \rho)s^2,$$

see Eq. (3.67). With autocorrelation, the variance is thus $1 + \rho$ times what the variance would be without autocorrelation (which is $2s^2$). Hence, an autocorrelation implies that

⁵See Arditti and Levy (1975), Bessembinder (2018), and Farago and Hjalmarsson (2023) for a more detailed discussion.

the standard deviation of the two-period log-return is multiplied by $\sqrt{1+\rho}$. Note that $2(1+\rho)s^2 < s^2$ when $\rho < -1/2$. Hence, with a strong negative autocorrelation, the variance of the two-period log-return can be lower than the variance of the one-period log-return, but such strong autocorrelations are rarely seen in financial data.

The effect of autocorrelation on the risk is straightforward. A positive autocorrelation implies a larger chance of getting a very high two-period return or a very low two-period return, while the chance of a non-extreme two-period return is reduced. The larger probability of extreme outcomes leads to a larger variance. Conversely, a negative autocorrelation reduces the chance of extreme two-period returns and increases the chance of non-extreme two-period returns, which gives a lower variance. Let us consider a simple example.

Example 3.9

Suppose the log-return on the stock market in any given year can be either 20% or -10% . Since log-returns are additive over time, the log-return over a two-year period is thus either 40% (both years good), 10% (one good and one bad year), or -20% (two bad years). In year 1, the two outcomes are equally likely, and the log-return in year 1 has expectation 5% and standard deviation $0.15 = 15\%$.

With a *zero autocorrelation*, the two outcomes are also equally likely in year 2. The chance of two good years in a row is then $0.5 \times 0.5 = 0.25$ or 25%. Similarly, the chance of two bad years in a row is also 0.25, whereas the probability of one good and one bad year in any order is 0.5. See Panel (a) of Figure 3.12 for an illustration. The expectation and variance of the two-year log-return are then

$$\begin{aligned} E[r_{0,2}^{\log}] &= 0.25 \times 40\% + 0.5 \times 10\% + 0.25 \times (-20\%) = 10\%, \\ \text{Var}[r_{0,2}^{\log}] &= 0.25 \times (40\% - 10\%)^2 + 0.5 \times (10\% - 10\%)^2 + 0.25 \times (-20\% - 10\%)^2 = 450(\%)^2, \end{aligned}$$

so that the standard deviation is $\sqrt{450(\%)^2} \approx 21.21\%$.

As an example of a *positive autocorrelation*, assume that if year 1 is good, then the probability of year 2 also being good is 0.8 (instead of 0.5) and, thus, the probability of year 2 being bad is only 0.2. On the other hand, if year 1 is bad, then the probability of year 2 also being bad is 0.8 (instead of 0.5), while the probability of year 2 being good is 0.2. See Panel (b) of Figure 3.12. Now the expected log-return in year 2 increases with the realized log-return in year 1:

$$\begin{aligned} E[r_{1,2}^{\log} | r_{0,1}^{\log} = 20\%] &= 0.8 \times 20\% + 0.2 \times (-10\%) = 14\%, \\ E[r_{1,2}^{\log} | r_{0,1}^{\log} = -10\%] &= 0.2 \times 20\% + 0.8 \times (-10\%) = -4\%. \end{aligned}$$

Here, for example, $E[r_{1,2}^{\log} | r_{0,1}^{\log} = 20\%]$ is the expected log-return in year 2 *conditional* on the log-return being 20% in year 1. The probability of two good years in a row and thus a two-year log-return of 40% is now $0.5 \times 0.8 = 0.4$, which is identical to the probability of two bad years in a row with a two-year log-return of -20% . Finally, the probability of one year being good and the other bad (in any order) is 0.2. It can be shown that the autocorrelation in this case is 0.6. Changing the probabilities in the formulas above, we still get an expected two-year log-return of 10%, but the standard deviation is now larger, 26.83%, given the larger probabilities of extreme outcomes. Note that the 26.83% equals

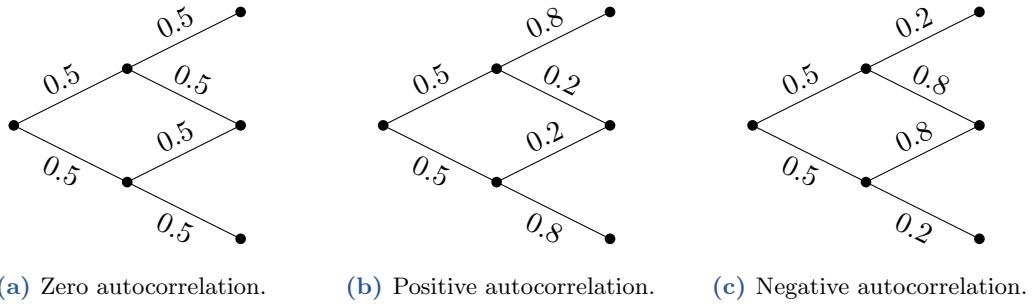


Figure 3.12: An example of autocorrelation in returns.

The figure refers to Example 3.9. An upward-sloping line between two nodes indicates a high return and a downward-sloping line a low return in a given year. The numbers along the lines represent the probability of the outcome.

the 21.21% from the uncorrelated case multiplied by the factor $\sqrt{1 + \rho} = \sqrt{1.6} \approx 1.265$.

To illustrate a *negative autocorrelation*, assume that if year 1 is good then the probability of year 2 being good is only 0.2, and similarly the probability is 0.2 that a bad year 1 is followed by a bad year 2, see Panel (c) of Figure 3.12. In this case, the expected log-return in year 2 is -4% following a large year 1 return and 14% following a low year 1 return. The probability of two good years in a row is only $0.5 \times 0.2 = 0.1$, which is identical to the probability of two bad years in a row, whereas the probability of one year being good and the other bad (in any order) is 0.8. In this case, the autocorrelation is -0.6 . The expected two-year log-return is still 10%, but the standard deviation is now only $13.42\% = \sqrt{1 - 0.6} \times 21.21\%$, reflecting the lower probabilities of extreme outcomes.

More generally, if periodic log-returns are positively autocorrelated, the variance of multi-period log-returns grows at a rate faster than the number of periods, and the Sharpe ratio based on log-returns increases by a rate slower than \sqrt{T} . If periodic log-returns are mildly negatively autocorrelated, the variance of multi-period log-returns grows at a rate slower than \sqrt{T} and the log-version of the Sharpe ratio grows at a rate faster than \sqrt{T} .

Next, let us turn to autocorrelations in rates of return. In this case, not only the variance, but also the expectation of the multi-period rate of return is affected by autocorrelation. With a positive autocorrelation, a high return this period tends to be followed by a high return next period, which also generates large returns on returns and thus a higher expected two-period rate of return. More formally, suppose the periodic rate of return has expectation μ and standard deviation σ , and let $\rho = \text{Corr}[r_{0,1}, r_{1,2}]$ be the autocorrelation which means that the autocovariance is

$$\text{Cov}[r_{0,1}, r_{1,2}] = \text{Corr}[r_{0,1}, r_{1,2}] \text{Std}[r_{0,1}] \text{Std}[r_{1,2}] = \rho\sigma^2.$$

Since the two-period rate of return is

$$r_{0,2} = (1 + r_{0,1})(1 + r_{1,2}) - 1 = r_{0,1} + r_{1,2} + r_{0,1}r_{1,2},$$

we get $E[r_{0,2}] = 2\mu + E[r_{0,1}r_{1,2}]$. From (3.38), we have $E[r_{0,1}r_{1,2}] = \text{Cov}[r_{0,1}, r_{1,2}] +$

$E[r_{0,1}] E[r_{1,2}] = \rho\sigma^2 + \mu^2$. Consequently,

$$E[r_{0,2}] = 2\mu + \rho\sigma^2 + \mu^2 = (1 + \mu)^2 - 1 + \rho\sigma^2. \quad (3.86)$$

If we compare this with (3.81) for $T = 2$, we see that the autocorrelation generates the term $\rho\sigma^2$. A positive autocorrelation in periodic rates of return lead to a higher expected multi-period rate of return. Conversely, a negative autocorrelation in periodic rates of return lead to a lower expected multi-period rate of return.

The variance is also affected by autocorrelation, as we saw already for log-returns, but with autocorrelation in rates of return, the variance of a multi-period rate of return becomes quite complicated, so we skip the exact formula. The intuition is similar as for log-returns: A positive autocorrelation in rates of return makes more extreme multi-period rates of return more likely and, thus, leads to an increase in variance. A negative autocorrelation leads to a reduction in variance.

Example 3.10

Suppose the assumptions in Example 3.9 hold for rates of return instead of log-returns. Given the compounding of returns, the two-year rate of return is then either $1.2 \times 1.2 - 1 = 0.44 = 44\%$, $1.2 \times 0.9 - 1 = 0.08 = 8\%$, or $0.9 \times 0.9 - 1 = -0.19 = -19\%$. With a zero autocorrelation, the probabilities of the three outcomes are 0.25, 0.5, and 0.25, respectively, so that the expectation and variance of the two-year rate of return are then

$$\begin{aligned} E[r_{0,2}] &= 0.25 \times 44\% + 0.5 \times 8\% + 0.25 \times (-19\%) = 10.25\%, \\ \text{Var}[r_{0,2}] &= 0.25 \times (44\% - 10.25\%)^2 + 0.5 \times (8\% - 10.25\%)^2 + 0.25 \times (-19\% - 10.25\%)^2 \\ &= 501.1875(\%)^2, \end{aligned}$$

so that the standard deviation is 22.39%. You can verify that these results are consistent with Eqs. (3.81) and (3.82) with $\mu = 0.1$ and $\sigma = 0.15$.

If the probabilities of the rate of return in year 2 depend on the realized rate of return in year 1 as shown in Panel (b) of Figure 3.12, we have a 0.6 autocorrelation in the rate of return. In this case, the two-year rate of return has expectation 11.60% and standard deviation 28.23%. On the other hand, with probabilities as shown in Panel (c), the autocorrelation is -0.6, and the two-year rate of return has expectation 8.90% and standard deviation 14.20%. Hence, with rates of return, both the expectation and the risk increase with the degree of autocorrelation.

In this section, we have only considered the first-order autocorrelation such as the correlation between returns in two subsequent months. Autocorrelations for other lags are defined similarly.

3.6.6 Covariances and correlations over different horizons

Next, we consider the covariance and the correlation between the return on two assets, say asset A and asset B. How does the covariance and correlation change with the length of the period over which we measure the returns? Again we distinguish between log-returns and rates of return.

First, let us focus on log-returns. Just as for the expectation and variance in Theorem 3.5, we get simple relations for how the covariance and correlation depend on the

investment horizon.

Theorem 3.8

Suppose that the periodic log-returns $r_{t,t+1}^{A,\log}$ and $r_{t,t+1}^{B,\log}$ on assets A and B are independent of each other and independent over time. Let $\sigma_{AB}^{\log} = \text{Cov}[r_{t,t+1}^{A,\log}, r_{t,t+1}^{B,\log}]$ and $\rho_{AB}^{\log} = \text{Corr}[r_{t,t+1}^{A,\log}, r_{t,t+1}^{B,\log}]$ denote the periodic covariance and correlation between the log-returns on the two assets. Then the covariance and correlation between the T -period log-returns on the two assets are

$$\text{Cov} [r_{0,T}^{A,\log}, r_{0,T}^{B,\log}] = T\sigma_{AB}^{\log}, \quad (3.87)$$

$$\text{Corr} [r_{0,T}^{A,\log}, r_{0,T}^{B,\log}] = \rho_{AB}^{\log}. \quad (3.88)$$

Proof

The correlation formula follows from the covariance formula together with the formula (3.70) for the standard deviation:

$$\begin{aligned} \text{Corr} [r_{0,T}^{A,\log}, r_{0,T}^{B,\log}] &= \frac{\text{Cov} [r_{0,T}^{A,\log}, r_{0,T}^{B,\log}]}{\text{Std} [r_{0,T}^{A,\log}] \times \text{Std} [r_{0,T}^{B,\log}]} \\ &= \frac{T\sigma_{AB}^{\log}}{\sqrt{T}\sigma_A^{\log} \times \sqrt{T}\sigma_B^{\log}} = \frac{\sigma_{AB}^{\log}}{\sigma_A^{\log} \times \sigma_B^{\log}} = \rho_{AB}^{\log}. \end{aligned}$$

To show the covariance formula, let us focus on the two-period case, where

$$r_{0,2}^{A,\log} = r_{0,1}^{A,\log} + r_{1,2}^{A,\log}, \quad r_{0,2}^{B,\log} = r_{0,1}^{B,\log} + r_{1,2}^{B,\log}.$$

Using the general computational rules for covariances, we get

$$\begin{aligned} \text{Cov} [r_{0,2}^{A,\log}, r_{0,2}^{B,\log}] &= \text{Cov} [r_{0,1}^{A,\log} + r_{1,2}^{A,\log}, r_{0,1}^{B,\log} + r_{1,2}^{B,\log}] \\ &= \text{Cov} [r_{0,1}^{A,\log}, r_{0,1}^{B,\log}] + \text{Cov} [r_{1,2}^{A,\log}, r_{1,2}^{B,\log}] \\ &\quad + \text{Cov} [r_{0,1}^{A,\log}, r_{1,2}^{B,\log}] + \text{Cov} [r_{1,2}^{A,\log}, r_{0,1}^{B,\log}]. \end{aligned}$$

If we assume that the log-return on one asset in one period is uncorrelated with the log-return on the other asset in other periods, the last two terms are zero. By assumption, each of the first two terms equals σ_{AB}^{\log} , so we get

$$\text{Cov} [r_{0,2}^{A,\log}, r_{0,2}^{B,\log}] = 2\sigma_{AB}^{\log}.$$

This procedure generalizes to the T -period case.

An implication of the theorem is that the covariance between the annual log-returns equals 12 times the covariance between the monthly log-returns, whereas the correlation

between the annual log-returns is equal to the correlation between the monthly log-returns. These results assume independence of returns both across periods and across assets.

Just as for variances, the relation between multi-period and single-period covariances is more complicated for rates of return than for log-returns, even when we assume that returns in different periods are uncorrelated with each other. We have the following result:

Theorem 3.9

Let $\sigma_{AB} = \text{Cov}[r_{t,t+1}^A, r_{t,t+1}^B]$ denote the covariance between the rates of return of assets A and B each period. Let $\mu_A = E[r_{t,t+1}^A]$ and $\mu_B = E[r_{t,t+1}^B]$ denote the expected periodic rate of return of the two assets. Assume returns are independent both across periods and across assets. Then the covariance between T -period rates of return is

$$\text{Cov}[r_{0,T}^A, r_{0,T}^B] = (\sigma_{AB} + (1 + \mu_A)(1 + \mu_B))^T - (1 + \mu_A)^T (1 + \mu_B)^T. \quad (3.89)$$

Proof

First note that

$$\begin{aligned} \text{Cov}[r_{0,T}^A, r_{0,T}^B] &= \text{Cov}[1 + r_{0,T}^A, 1 + r_{0,T}^B] \\ &= E[(1 + r_{0,T}^A)(1 + r_{0,T}^B)] - E[1 + r_{0,T}^A]E[1 + r_{0,T}^B]. \end{aligned}$$

Here, in the last term, we can apply Eq. (3.81) to get

$$E[1 + r_{0,T}^A] = 1 + E[r_{0,T}^A] = (1 + \mu_A)^T$$

and similarly for asset B. Furthermore, assuming that returns in different periods are independent of each other, we find that

$$\begin{aligned} E[(1 + r_{0,T}^A)(1 + r_{0,T}^B)] &= E[(1 + r_{0,1}^A) \dots (1 + r_{T-1,T}^A) \times (1 + r_{0,1}^B) \dots (1 + r_{T-1,T}^B)] \\ &= E[(1 + r_{0,1}^A)(1 + r_{0,1}^B)] \times \dots \times E[(1 + r_{T-1,T}^A)(1 + r_{T-1,T}^B)] \\ &= (E[(1 + r_{t,t+1}^A)(1 + r_{t,t+1}^B)])^T \\ &= (E[1 + r_{t,t+1}^A + r_{t,t+1}^B + r_{t,t+1}^A r_{t,t+1}^B])^T \\ &= (1 + E[r_{t,t+1}^A] + E[r_{t,t+1}^B] + E[r_{t,t+1}^A r_{t,t+1}^B])^T \\ &= (1 + \mu_A + \mu_B + \sigma_{AB} + \mu_A \mu_B)^T \\ &= (\sigma_{AB} + (1 + \mu_A)(1 + \mu_B))^T. \end{aligned}$$

Combining these findings, we get (3.89).

Note that the covariance expression in Eq. (3.89) is considerably more complicated than $T\sigma_{AB}$. Given Eq. (3.89) for the covariance and Eq. (3.82) for the variance of annual rates

of return, it is clear that the correlation between T -period rates of return generally differs from the correlation between one-period rates of return.

As an example, suppose that the monthly rates of return on both assets A and B have an expectation of 0.008 and a standard deviation of 0.12, and that the covariance between the monthly rates of return is 0.00864, corresponding to a correlation of 0.6. Then the annual rates of return have a covariance of 0.12949, larger than $12 \times 0.00864 \approx 0.10368$, and a correlation of 0.581 and thus different—but close to—the monthly correlation.

3.7 Using historical returns

When we invest money in financial markets, we care about future returns. However, to the extent that the future resembles the past, we might learn a lot about the possible future returns from looking at past returns. Analysts often use time series of past returns to estimate the type or shape of the distribution as well as the key moments of future returns.

Some return time series are freely available and easily accessible online. One provider is [Yahoo Finance](http://finance.yahoo.com) that offers lots of information on their homepage <http://finance.yahoo.com>. The homepages of finance professors Aswath Damodaran at Stern School of Business, New York University⁶ and Kenneth R. French at Tuck School of Business, Dartmouth College⁷ provide access to an abundance of financial data. Many central banks publish free data on exchange rates and government bonds on their homepages, see for example the homepage <http://www.federalreserve.gov/> of the Federal Reserve, the central bank of the United States. Furthermore, many business schools and universities subscribe to a number of financial databases and often grant students access. An example is the CRSP database from the Center for Research in Security Prices at the University of Chicago's Booth School of Business. CRSP contains data mainly on U.S. stocks, mutual funds, and Treasury bonds.

3.7.1 Constructing empirical distribution of returns

The empirical distribution of past returns calculated for some period length—typically a year, a quarter, a month, a week, or a day—might represent a good guess for the probability distribution of the future return over a period of the same length. Given a time series of returns, the empirical distribution in the form of a histogram is easy to construct in Excel once the Data Analysis toolpak has been installed. Choose ‘Data Analysis’, then ‘Histogram’, and click ‘OK’. Enter the ‘Input Range’ of your time series which should be written in a column range of cells in your worksheet. A histogram counts and displays the number of observations in various non-overlapping intervals or bins. In your worksheet you need to write the division points of these bins in a range of cells. Enter this as the ‘Bin Range’. Check the box ‘Chart Output’ and choose where you want the output placed. Then click ‘OK’ and Excel produces the histogram.

As an example, we consider the annual rates of return on the S&P 500 index from 1946 to 2022. The 77 return observations range from -36.46% in 2008 to 52.65% in 1954 with an arithmetic average (see below for the definition) of 12.33% . Figure 3.13 displays two histograms generated by Excel based on this time series. The histogram to the left is produced using equal-sized bins defined by $-15\%, -10\%, -5\%, \dots, 40\%$, whereas the histogram to the right uses equal-sized bins defined by $-16\%, -8\%, -0\%, \dots, 40\%$. For example, the left-most bar in the left histogram shows that in 4 of the 77 years, the rate

⁶<http://pages.stern.nyu.edu/~adamodar/>

⁷<http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>

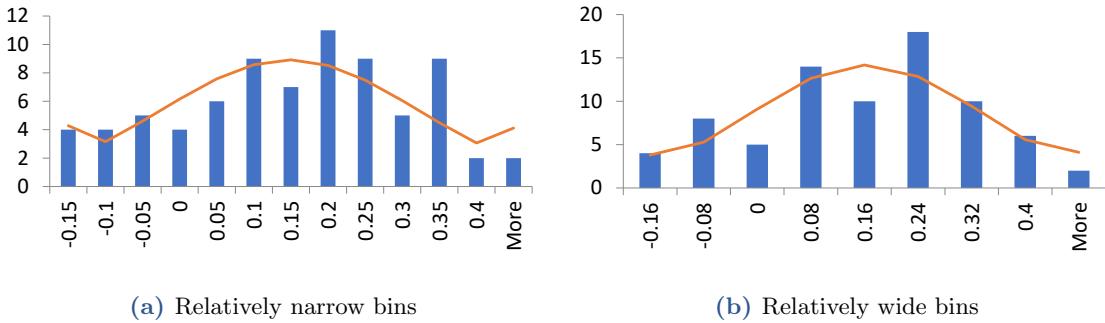


Figure 3.13: Historical stock returns.
The graphs present histograms of the annual rates of return on the S&P 500 stock index from 1946 to 2022. The orange curve depicts the expected number of observations in each bin if the observations were drawn from a normal distribution with the same mean and variance as the data. Data source: CRSP, downloaded via WRDS on September 19, 2023.

of return was lower than -15% . The next bar in the left histogram shows that in 4 of the 77 years, the rate of return was between -15% and -10% , etc. The two histograms give somewhat different impressions of the distribution. The orange curve shows the expected number of observations in each bin if the observations were drawn from a normal distribution with a mean and a variance equal to the sample mean and sample variance. A visual comparison of the bars and the curves suggests that the empirical distribution resembles a normal distribution somewhat but also that the match is not perfect.

When doing such a visual comparison with the normal distribution, note that if you draw a modest quantity of random numbers from a normal distribution, a histogram generated from these numbers may look very different from a normal distribution. As an illustration, the blue columns in Figure 3.14 show the distribution of the 120 monthly returns on Walmart stocks from October 2006 to September 2016. For example, two months showed a return less than -10% , one month between -10% and -8% , etc. The sample mean and standard deviation (see below) were 0.62% and 4.62% , respectively. The gray, red, and green columns show three histograms each based on 120 draws from the normal distribution with this mean and standard deviation. These histograms do not closely resemble the bell-shaped curve of the underlying normal distribution, and the empirical Walmart return distribution does not stand out from the three simulated histograms. It is just difficult to tell from 120 observed returns whether or not the returns are coming from a normal distribution.

3.7.2 Computing summary statistics

Given a time series of rates of return, real or nominal, r_1, r_2, \dots, r_T over T time periods of equal length, we are often interested in the average return as this might be a good estimate of the return that we can expect from the asset in the future. You can compute the *arithmetic average* rate of return as

$$\bar{r}_{\text{arith}} = \frac{1}{T} \sum_{t=1}^T r_t, \quad (3.90)$$

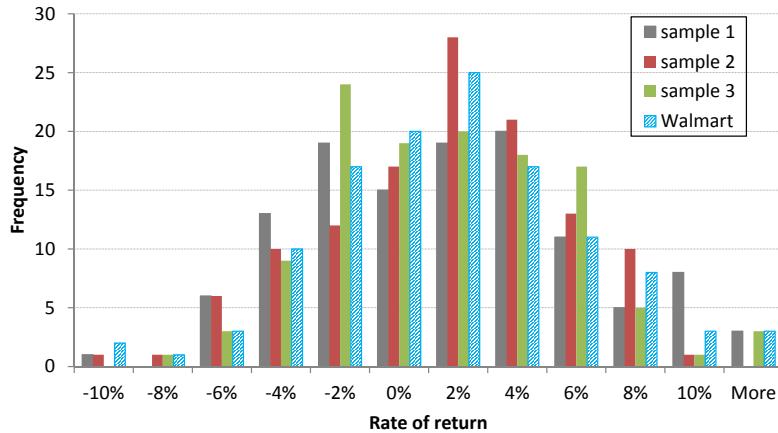


Figure 3.14: Is Walmart normal?

The blue columns constitute a histogram of the 120 observed monthly returns on Walmart stocks from October 2006 to September 2016. The returns were calculated from monthly adjusted closing prices downloaded from Yahoo finance. The gray, the red, and the green columns form three histograms each generated by drawing 120 samples from a normal distribution with the same mean and standard deviation as the observed Walmart returns.

which is the sample mean of the time series. This may be a good measure of the return you can expect over any period of the same length as covered by each of the returns in the sample. On the other hand, it is not necessarily a good measure for the return you can expect to get over longer periods. This is due to the fact that returns are compounded as explained in Section 2.2. If you invest 1 unit of account (1 dollar if dividends and prices are measured in dollars) at the starting date of the time series and keep reinvesting dividends by purchasing additional units of the asset, then you will end up with

$$(1 + r_1)(1 + r_2) \dots (1 + r_T) \quad (3.91)$$

after the last period. The *geometric average* rate of return is computed as

$$\bar{r}_{\text{geo}} = [(1 + r_1)(1 + r_2) \dots (1 + r_T)]^{1/T} - 1, \quad (3.92)$$

implying that

$$(1 + \bar{r}_{\text{geo}})^T = (1 + r_1)(1 + r_2) \dots (1 + r_T) \quad (3.93)$$

so in that sense you can think of \bar{r}_{geo} as the average periodic return over longer periods. Note that the right-hand side of (3.93) is simply the gross return over all T periods. If dividends are reinvested along the way, this is simply the final value divided by the initial value of the asset. For the geometric average return, it does not matter which sequence of periodic returns led to this gross return over the entire sample period. The geometric average return is closely related to the arithmetic average of the log-returns $r_t^{\log} = \ln(1 + r_t)$:

$$\frac{1}{T} \sum_{t=1}^T r_t^{\log} = \ln(1 + \bar{r}_{\text{geo}}), \quad (3.94)$$

which follows from the calculation

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T r_t^{\log} &= \frac{1}{T} \sum_{t=1}^T \ln(1 + r_t) = \frac{1}{T} (\ln(1 + r_1) + \ln(1 + r_2) + \dots + \ln(1 + r_T)) \\ &= \frac{1}{T} \ln \{(1 + r_1)(1 + r_2) \dots (1 + r_T)\} \\ &= \ln \left\{ ((1 + r_1)(1 + r_2) \dots (1 + r_T))^{1/T} \right\} = \ln (1 + \bar{r}_{\text{geo}}),\end{aligned}$$

where we have used the rules $\ln(xy) = \ln(x) + \ln(y)$ and $a \ln(x) = \ln(x^a)$.

The geometric average and the arithmetic average can be very different for a given sample. As a simple example, assume that the percentage return of an asset is 100% in year 1 and -50% in year 2. Then the arithmetic average return is $(100\% + (-50\%))/2 = 25\%$ and the geometric average return is $[(1+1)(1-0.5)]^{1/2} - 1 = 1 - 1 = 0$, that is 0%. An investment of 1 in the beginning of the first year has grown to 2 at the end of year 1 and then dropped to $2 \times 0.5 = 1$ at the end of year 2. The zero return over the two periods is better reflected by the geometric average than the arithmetic average.

The geometric average is always lower than the arithmetic average, and the difference between the two is larger for a very variable series of returns. The difference is sometimes measured by the approximation

$$\bar{r}_{\text{arith}} \approx \bar{r}_{\text{geo}} + \frac{1}{2} \tilde{\sigma}^2, \quad (3.95)$$

where $\tilde{\sigma}^2$ is the sample variance of the returns r_1, r_2, \dots, r_T (defined below). For some samples the approximation is quite imprecise. We can make some sense of the relation if we assume that the log-returns r_t^{\log} are sampled from a normal distribution. In that case we have from (3.31) that

$$E[1 + r] = \exp \left\{ E[r^{\log}] + \frac{1}{2} \text{Var}[r^{\log}] \right\} = \exp \left\{ E[r^{\log}] \right\} \times \exp \left\{ \frac{1}{2} \text{Var}[r^{\log}] \right\}. \quad (3.96)$$

For a given sample we can estimate the left-hand side by

$$E[1 + r] = 1 + E[r] \approx 1 + \frac{1}{T} \sum_{t=1}^T r_t = 1 + \bar{r}_{\text{arith}}.$$

On the right-hand side we can similarly estimate the expectation by

$$E[r^{\log}] \approx \frac{1}{T} \sum_{t=1}^T r_t^{\log} = \ln (1 + \bar{r}_{\text{geo}}),$$

where the equality follows from (3.94). Hence, Eq. (3.96) leads to

$$1 + \bar{r}_{\text{arith}} \approx (1 + \bar{r}_{\text{geo}}) \times \exp \left\{ \frac{1}{2} \text{Var}[r^{\log}] \right\}.$$

If we use the approximations $e^x \approx 1 + x$ (good for x near zero) and $\text{Var}[r^{\log}] \approx \text{Var}[r]$ and

replace the latter with the sample variance $\tilde{\sigma}^2$, then we get

$$1 + \bar{r}_{\text{arith}} = (1 + \bar{r}_{\text{geo}}) \times \left(1 + \frac{1}{2}\tilde{\sigma}^2\right) \approx 1 + \bar{r}_{\text{geo}} + \frac{1}{2}\tilde{\sigma}^2,$$

where the last approximation simply ignores the product $\frac{1}{2}\bar{r}_{\text{geo}} \times \tilde{\sigma}^2$ which tends to be much smaller than the other terms. Finally, we can subtract 1 from both sides of the equation to get (3.95). Note that this derivation and thus the conclusion are based on various approximations and the assumption of normally distributed log-returns.

In Excel, the arithmetic average is computed with the function **AVERAGE** applied to the range of cells containing all the rates of return. To compute the geometric average with Excel, you need to have the gross returns $1 + r_1, 1 + r_2, \dots$ in a range of cells, apply the function **GEOMEAN** to that and subtract 1. Alternatively, calculate the geometric average return directly by raising the ratio of the final value over the initial value to the power of $1/T$ and subtract one. For the 77 annual S&P 500 returns considered above, the arithmetic average rate of return is 12.32%, whereas the geometric average is 10.94%.

If we let $\tilde{\mu}$ denote the sample mean (the arithmetic average) of the time series,

$$\tilde{\mu} = \frac{1}{T} \sum_{t=1}^T r_t, \quad (3.97)$$

the sample variance of the time series $\tilde{\sigma}^2$ is given by

$$\tilde{\sigma}^2 = \frac{1}{T-1} \sum_{t=1}^T (r_t - \tilde{\mu})^2. \quad (3.98)$$

The division by $T-1$ instead of T ensures that the sample variance is an unbiased estimate of the unknown population variance. The sample standard deviation is just the square root of the sample variance. In Excel the sample variance is computed with the function **VAR.S** and the sample standard deviation with **STDEV.S**. Alternatively, if you have the Data Analysis toolpak installed, just choose ‘Data Analysis’ under the ‘Data’ tab and then pick ‘Descriptive Statistics’ which gives you a bunch of summary statistics. For our 1946-2022 time series of S&P 500 returns, the sample variance is 0.02945 or 294.5(%)², and the sample standard deviation is 0.1716 or 17.16%.

Typically, the sample mean is seen as an estimate of the unknown mean of a distribution. To be more specific, suppose you think that the rate of return r on a given asset over some future period is normally distributed with a mean of μ and a variance of σ^2 , but you do not know the values of μ and σ^2 . However, you have a number of observed rates of return r_1, r_2, \dots, r_T on the same asset over past periods of the same length. If you believe that all those returns were independent draws from the same normal distribution, you can use the sample mean $\tilde{\mu}$ in (3.97) as an estimate of the true mean μ . But it is just an estimate. The sample variance is informative about the precision of the estimated mean. It can be shown that a 95% confidence interval of the true mean μ is the interval

$$95\% \text{ confidence interval for } \mu: \left[\tilde{\mu} - t_{(T-1, 0.975)} \times \frac{\tilde{\sigma}}{\sqrt{T}}, \tilde{\mu} + t_{(T-1, 0.975)} \times \frac{\tilde{\sigma}}{\sqrt{T}} \right],$$

where T is the number of observations and $\tilde{\sigma}$ is the sample standard deviation defined above. The symbol $t_{(T-1, 0.975)}$ represents the 97.5% percentile of a t -distribution with $T-1$ degrees of freedom. When the degrees of freedom increases towards infinity, the

t -distribution converges to the normal distribution which has a 97.5% percentile equal to 1.960, cf. Table 3.2. With a finite sample and thus finite degrees of freedom, the t -distribution percentile exceeds the corresponding percentile of the normal distribution somewhat. For example, with 25, 50, 75, 100, and 200 degrees of freedom, respectively, the 97.5% t -percentile is 2.060, 2.009, 1.992, 1.984, and 1.972. The t -distribution percentile can be computed in Excel as `T.INV(0.975;df)`, where ‘ df ’ is the degrees of freedom.

Let us take the S&P 500 returns as an example. As explained above, the sample mean is 12.32% and the sample standard deviation is 17.16%. With $T = 77$ observations, there are $T - 1 = 76$ degrees of freedom. Since $t_{(76,0.975)} = 1.9917$, the 95% confidence interval for the mean return of the S&P 500 index is

$$\left[12.32\% - 1.9917 \times \frac{17.16\%}{\sqrt{77}}, 12.32\% + 1.9917 \times \frac{17.16\%}{\sqrt{77}} \right] = [8.43\%, 16.22\%].$$

In words, we are 95% certain that the true mean return is between 8.43% and 16.22%. Even with 77 years of data, our estimate of the mean is therefore very imprecise. And this is even under the courageous assumption that all 77 annual returns are drawn from the same distribution, meaning that the mean and variance of the annual return have not changed during this long period.

Note that the width of the confidence interval increases with the sample standard deviation. If the standard deviation had been 40% instead of 17.16%, the 95% confidence interval of the mean would be [3.25%, 21.41%], which is very wide. A standard deviation of 40% or more is not unusual for individual stocks, so it is extremely difficult to estimate their expected return from past returns.

Another implication is that it is difficult to distinguish skill from luck in investment decisions. To evaluate a portfolio manager we typically only have a short sample of returns with a relatively high standard deviation, which leads to a very wide confidence interval for the mean return of the manager. It is hard to tell whether a large average return over a period of few years is due to the manager being lucky or skillful in making the investment decisions. More on the evaluation of investment performance in Chapter 13.

The problem of the imprecise mean estimate cannot be solved by sampling returns more frequently, e.g., by looking at monthly returns instead of annual returns. If we disregard compounding (or work with log-returns), the estimate of the mean monthly return would be $\tilde{\mu}/12$ with a standard deviation of $\tilde{\sigma}/\sqrt{12}$, cf. (3.68) and (3.70). The number of observations would then be $12 \times T$, where T is the number of years in the sample. Hence, the lower bound of the 95% confidence interval for the mean monthly return is

$$\frac{\tilde{\mu}}{12} - t \times \frac{\tilde{\sigma}/\sqrt{12}}{\sqrt{12T}} = \frac{\tilde{\mu}}{12} - t \times \frac{\tilde{\sigma}}{12\sqrt{T}} = \frac{1}{12} \left(\tilde{\mu} - t \times \frac{\tilde{\sigma}}{\sqrt{T}} \right),$$

which is simply the lower bound on the annual mean converted into a monthly frequency. Similarly for the upper bound. Hence, we arrive at the same confidence interval if we ignore the very small change in the relevant t -percentile.

The sample variance $\tilde{\sigma}^2$ is an estimate of the true return variance σ^2 . Under the assumption that all the observed returns are independent draws from the same normal distribution, the 95% confidence interval of the variance is

$$95\% \text{ confidence interval for } \sigma^2: \left[\frac{(T-1)\tilde{\sigma}^2}{\chi^2_{(T-1,0.975)}}, \frac{(T-1)\tilde{\sigma}^2}{\chi^2_{(T-1,0.025)}} \right],$$

where $\chi^2_{(T-1,p)}$ is the p -percentile of the so-called chi-square distribution with $T-1$ degrees

of freedom. The percentiles can be computed in Excel as `CHISQ.INV(p;df)`, where ‘df’ is again the degrees of freedom. Based on the 77 annual S&P 500 returns, a 95% confidence interval of the return variance is

$$\left[\frac{76 \times 294.5(\%)^2}{102.00}, \frac{76 \times 294.5(\%)^2}{53.78} \right] = [219.4(\%)^2, 416.1(\%)^2],$$

since the 97.5% and 2.5% percentiles of the chi-square distribution with 76 degrees of freedom are 102.00 and 53.78, respectively. Note that this interval is not symmetric around the sample variance of $294.5(\%)^2$ since the chi-square distribution is asymmetric. Taking square roots of the endpoints of the above interval, we obtain a 95% confidence interval for the standard deviation of $[14.81\%, 20.40\%]$, which again is not symmetric around the sample standard deviation of 17.16%. Note that the confidence interval for the standard deviation is narrower than that for the mean, which is often the case when working with returns. It can also be shown that the confidence interval for the standard deviation does get narrower if we increase the sampling frequency, for example by working with monthly returns instead of annual returns. With more frequent observations we can better estimate the variance and standard deviation, but not the mean.

Turning to higher moments, we can calculate the sample skew and kurtosis as

$$\widetilde{\text{Skew}} = \frac{\frac{1}{T} \sum_{t=1}^T (r_t - \tilde{\mu})^3}{\left(\frac{1}{T} \sum_{t=1}^T (r_t - \tilde{\mu})^2 \right)^{3/2}}, \quad \widetilde{\text{Kurt}} = \frac{\frac{1}{T} \sum_{t=1}^T (r_t - \tilde{\mu})^4}{\left(\frac{1}{T} \sum_{t=1}^T (r_t - \tilde{\mu})^2 \right)^2} - 3.$$

As for the variance, the sample skew and sample kurtosis are sometimes scaled by specific multipliers slightly different from 1 to get closer to unbiased estimates of the unknown population skew and population kurtosis. In Excel the sample skew and kurtosis are computed using the functions `SKEW` and `KURT`. For the time series of annual S&P 500 returns discussed above, the sample skew of -0.377 indicates that the empirical distribution leans somewhat to the right, which also can be detected from the histograms in Figure 3.13. The sample kurtosis of 0.009 is virtually zero and no fat tails are visible.

If you are interested in the covariance between the future returns on two stocks, you might want to know how the two stocks have covaried in the past. Suppose that for any period $t = 1, 2, \dots, T$ you have rates of return r_{1t} and r_{2t} on stock 1 and stock 2 respectively. If $\tilde{\mu}_1$ and $\tilde{\mu}_2$ denote the sample means (arithmetic average returns) of the two stocks, a sample covariance is typically computed by

$$\widetilde{\text{Cov}} = \frac{1}{T-1} \sum_{t=1}^T (r_{1t} - \tilde{\mu}_1)(r_{2t} - \tilde{\mu}_2). \quad (3.99)$$

In Excel the sample covariance can be computed using the function `COVARIANCE.S` that takes the two time series as inputs. The sample correlation between the two returns is

$$\widetilde{\text{Corr}} = \frac{\sum_{t=1}^T (r_{1t} - \tilde{\mu}_1)(r_{2t} - \tilde{\mu}_2)}{\sqrt{\sum_{t=1}^T (r_{1t} - \tilde{\mu}_1)^2} \sqrt{\sum_{t=1}^T (r_{2t} - \tilde{\mu}_2)^2}} = \frac{\widetilde{\text{Cov}}}{\tilde{\sigma}_1 \tilde{\sigma}_2}, \quad (3.100)$$

where the last expression is the sample covariance divided by the product of the sample standard deviations. In Excel the sample correlation can be computed directly using the function `CORREL` with the two time series as inputs.

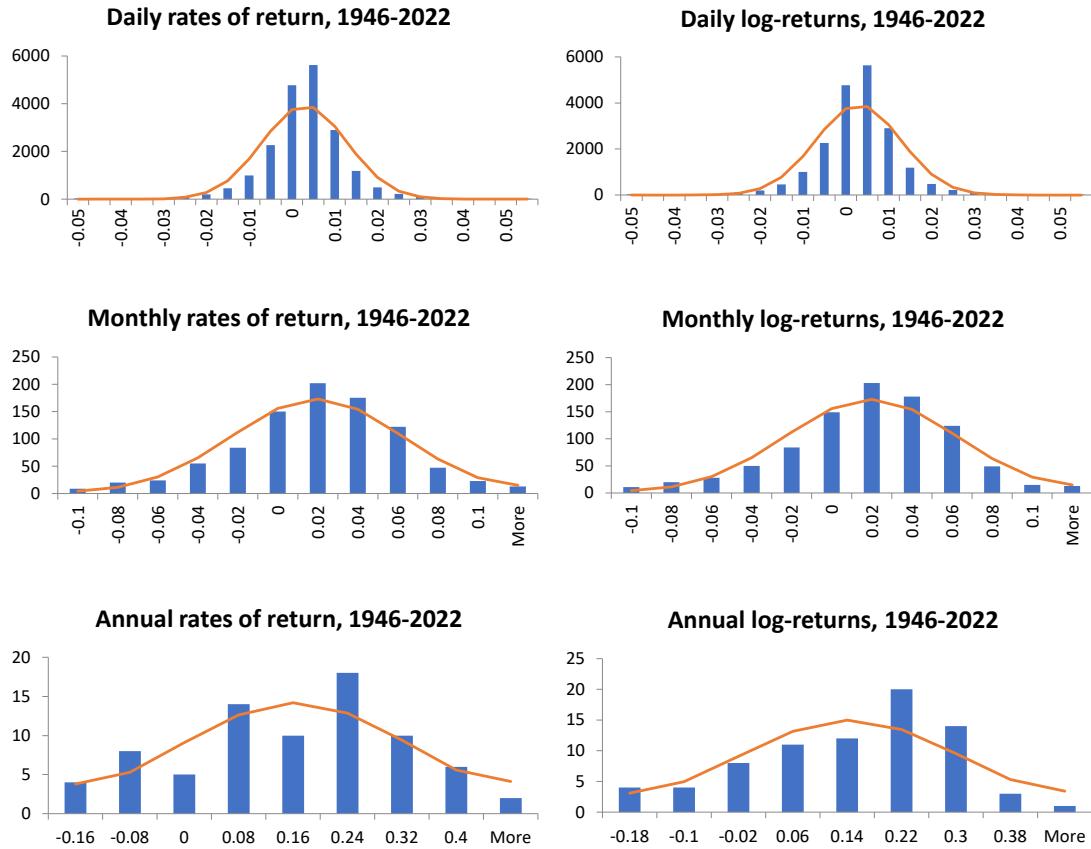


Figure 3.15: Historical returns over different investment horizons.

The graphs present histograms of observed returns on the S&P 500 stock index from 1946 to 2022 on either a daily, monthly, or annual horizon. The graphs in the left column are for rates of return and the graphs in the right column for log-returns. The orange curve depicts the expected number of observations in each bin if the observations were drawn from a normal distribution with the same mean and variance as the data. Data source: CRSP, downloaded via WRDS on September 19, 2023.

3.7.3 Past returns over different investment horizons

In the preceding subsections we have considered past annual returns on the S&P 500 index from 1946 to 2022. What about returns over different investment horizons? Figure 3.15 shows histograms of daily, monthly, and annual returns on the S&P 500 over the same sample period. The graphs to the left depict rates of return, the graphs to the right show log-returns. The 924 monthly rates of return and log-returns seem to fit the normal distribution well. The 77 annual rates of return and log-returns seem to give a poorer fit, but they are also based on relatively few observations. The 19585 daily rates of return and log-returns, on the other hand, appear to fit the normal distribution relatively poorly with too many observations very close to the mean and too few observations in a modest distance from the mean.

Table 3.8 reports selected summary statistics on the observed S&P 500 returns. All six data series have a slightly negative skewness indicating a peak to the right of the mean and a longer left tail than right tail, see Figure 3.2. The kurtosis is positive for all six series: small for the annual series, somewhat larger for the monthly series, and

high for the daily series. The positive kurtosis is often associated with fat tails, but the lower part of Table 3.8 reveals that the tails are not excessively fat. For a normal distribution, 2.28% of the observations should be two standard deviations below the mean and 2.28% two standard deviations above the mean. The return series have a slightly larger fraction of observations more than two standard deviations below the mean and a slightly smaller fraction more than two standard deviations above the mean. When looking more than three standard deviations away from the mean, a normal distribution has 0.13% of observations in each tail, but most of the six series have more observations in the left tail and the daily series also have more observations in the right tail. All in all, the return distributions have a somewhat larger fraction of extreme observations, in particular extreme negative observations, than the normal distribution.

The first-order autocorrelations of all six return series are close to zero. For annual returns, there is a slight indication of a negative autocorrelation, i.e., a weak tendency that a good year is followed by a bad year and a bad year is followed by a good year.

To sum up, our empirical analysis of S&P 500 returns indicate that assuming that either rates of return or log-returns over a monthly or annual horizon are normally distributed is a reasonably good approximation of reality. We can accept some deviations from reality because the normal distribution is much more tractable than any alternative distributions that may fit the return data even better. For returns over multiple years, we have too few (non-overlapping) observations to say something meaningful about the distribution type. Here, we have to rely on the implications that normality of short-term rates of return of log-returns have for longer-term returns as studied in Section 3.6.

Only few recent research papers have considered return distributions of individual stocks. [Jondeau, Zhang, and Zhu \(2019\)](#) show, among other things, that individual stocks tend to have a positive skewness, whereas the index (as shown above) has a negative skewness. Small stocks in terms of market capitalization tend to have higher skewness and higher kurtosis than large stocks, as we shall see some support of in Chapter 6. An important difference between individual stocks and a stock index is that individual stocks can be removed from the stock market if the company issuing the stock is acquired by a different company or if goes bankrupt. In the latter case, the stock price typically ends up at zero so that a rate of return of -100% is realized. As shown by [Bessembinder \(2018\)](#) and discussed further in Chapter 6, this occurs more often than you would probably think.

3.7.4 Predictions

It would be great if we could predict the return on an asset over a future period from the information available today. Investors, analysts, and academics have tried for decades to find ways to predict stock returns. Often the endeavour consists of running a linear regression of returns on variables you think might predict the returns.

Suppose r_{t+1} denotes the rate of return (or excess return) on some asset or portfolio over period $t + 1$, which is the time interval from time t to time $t + 1$ measured in some time unit. Let x_t denote the value at time t of some candidate predictor. Given a time series of observations of both the return and the candidate predictor, you can run a linear regression of the form

$$r_{t+1} = a + bx_t + \varepsilon_{t+1}, \quad t = 0, 1, \dots, T - 1, \tag{3.101}$$

where a and b are constants to be estimated and ε_{t+1} is the mean-zero residual. Note the time subscripts in the equation. The known value of the predictor at time t is related to the return over the following period. If the regression estimate of the coefficient b is

	Rates of return			Log-returns		
	Annual	Monthly	Daily	Annual	Monthly	Daily
Mean	0.1233	0.0096	0.00046	0.1038	0.0087	0.00041
Std dev	0.1716	0.0422	0.0098	0.1622	0.0423	0.0099
Skew	-0.3773	-0.4094	-0.6038	-0.8588	-0.6312	-0.9262
Kurt	0.0089	1.4818	18.1720	0.9898	2.0764	22.2940
Autocorr, lag 1	-0.0662	0.0149	0.0095	-0.0539	0.0244	0.0104
Min	-0.3646	-0.2158	-0.1946	-0.4535	-0.2431	-0.2164
Max	0.5265	0.1681	0.1151	0.4230	0.1554	0.1090
Fraction of observations 2 std dev away from mean (normal: 2.28%)						
Below	2.60%	3.57%	2.53%	3.90%	3.79%	2.54%
Above	1.30%	1.84%	2.20%	0.00%	1.52%	2.11%
Fraction of observations 3 std dev away from mean (normal: 0.13%)						
Below	0.00%	0.54%	0.80%	1.30%	0.54%	0.83%
Above	0.00%	0.11%	0.67%	0.00%	0.11%	0.63%

Table 3.8: Historical returns over different investment horizons.

The table shows summary statistics on observed returns on the S&P 500 stock index from 1946 to 2022 on either a daily, monthly, or annual horizon. The left part of the table considers rates of return and the right part log-returns. The lower part of the table shows the fraction of observations that are more than two or three standard deviations away from the mean, which for a normal distribution is 2.28% and 0.13%, respectively. Data source: CRSP, downloaded via WRDS on September 19, 2023.

statistically significant, the conclusion is that x is a useful return predictor. If the linear relation also holds in the future with the same value of b , then we can use the current value of the predictor to assess the future return on the asset. More precisely, the expected return in the next period is $a + bx_t$, but of course the residual ε_{t+1} might turn out positive or negative so that the realized return differs from the expected.

In practice, the residual risk in such regressions is substantial so that it is difficult to predict the return with great precision. This is not really surprising. As discussed above, it is very difficult to estimate a constant expected return with high precision. Intuitively, it is even more difficult to estimate an expected return that varies over time. Moreover, it often turns out that the relation between returns and potential predictors is quite unstable over time, meaning that we should not put too much trust into the estimate of b . We return to the discussions of return predictability in subsequent chapters.

3.8 Exercises

Exercise 3.1. Consider a financial market with two stocks, A and B, and three possible states (outcomes) for the economy at year-end. The following table provides the state-dependent rates of return.

State of economy	Probability	Return stock A	Return stock B
Good	0.25	20%	18%
Average	0.50	10%	0%
Bad	0.25	-5%	10%

- (a) Determine the expected return $E[r]$ for each stock.

- (b) Determine the return variance $\text{Var}[r]$ for each stock.
- (c) Determine the expected squared return $E[r^2]$ for each stock and verify that $\text{Var}[r] = E[r^2] - (E[r])^2$.
- (d) Determine the return standard deviation for each stock.
- (e) Determine the covariance and the correlation between the returns of the two stocks.

Exercise 3.2. Let r be the rate of return on a given financial asset. Assume that r is normally distributed with mean μ and variance σ^2 .

- (a) How can you use the cumulative distribution function $N(x)$ for the standard $N(0,1)$ normal distribution to compute the following probabilities?
 - (i) the probability that the rate of return is below some fixed level \underline{R} , that is $\text{Prob}(r \leq \underline{R})$
 - (ii) the probability that the rate of return is above some fixed level \bar{R} , that is $\text{Prob}(r \geq \bar{R})$
 - (iii) the probability that the rate of return falls between two fixed levels \underline{R} and \bar{R} (assuming $\underline{R} < \bar{R}$), that is $\text{Prob}(\underline{R} < r < \bar{R})$.
- (b) Suppose now that $\mu = 0.10 = 10\%$ and $\sigma = 0.40 = 40\%$. To answer the following questions you might want to use the Excel functions **NORM.DIST** or **NORM.S.DIST**.
 - (i) What is the probability that the rate of return on the asset is below -10% ?
 - (ii) What is the probability that the rate of return on the asset is above 30% ?
 - (iii) What is the probability that the rate of return on the asset is between 0% and 20% ?

Exercise 3.3. Let r_1 and r_2 denote the rates of return on two investments over a one-year period. Suppose that r_1 and r_2 are both normally distributed with means $\mu = 0.1$ and standard deviation $\sigma = 0.4$ and that the correlation between the two returns is $\rho = 0.2$.

- (a) What is the probability distribution of the return difference $r_1 - r_2$?
- (b) What is the probability that investment 1 outperforms investment 2 over the next year, i.e., that the realized return on investment 1 is larger than the realized return on investment 2? How does that probability depend on μ , σ , and ρ ? Explain!
- (c) What is the probability that investment 1 over the next year outperforms investment 2 by 5 percentage points? By 10 percentage points?
- (d) Answer the preceding question for all 25 combinations of standard deviations σ in the set $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ and correlations ρ in $\{-0.8, -0.4, 0, 0.4, 0.8\}$. Keep $\mu = 0.1$ in all cases. Explain how the values of σ and ρ affect the answer.
- (e) Discuss what the above analysis implies for the performance evaluation and comparison of portfolio managers.

Exercise 3.4. SPDR S&P 500 is the name of an exchange-traded fund designed to track the S&P 500 index of U.S. stocks. The ‘DR’ in the name stands for ‘Depositary Receipts’. The fund is listed on the New York Stock Exchange (NYSE) under the ticker symbol ‘SPY’. The supplementary material for these lecture notes includes an Excel file **Exercise3_4_data.xlsx** which contains monthly data on SPY from December 2010 to December 2020, downloaded in August 2021 from Yahoo Finance. Recall that the entry dated 01-12-2010 really covers the entire month of December 2010. In particular, the adjusted closing price of 125.75 shown for that date is in fact the adjusted closing price on the last trading date in December 2010. Note the very high average daily trading volume.

- (a) Download the same data from Yahoo Finance and import them into an Excel file. Is there any difference between the adjusting closing prices in the file **Exercise3_4_data.xlsx** and in your file? Why can this happen? Calculate monthly rates of return in both files. Are they identical? (In case that they are not, please use the monthly rates of return based on the data in the file **Exercise3_4_data.xlsx** to answer the questions below.)
- (b) What is the minimum and the maximum rates of return and when did they occur?
- (c) Construct a histogram of the monthly rates of return.
- (d) Use the Excel functions to compute both the arithmetic and the geometric average monthly return as well as the variance and the standard deviation of the monthly returns. How would you annualize the average and the standard deviation of the monthly returns?

We will analyze how well the normal distribution assumption holds for the returns on SPY. To answer some of the following questions you might want to use the Excel functions `NORM.DIST`, `NORM.S.DIST`, `NORM.INV`, or `NORM.S.INV`.

- (e) Under the assumption that returns are normally distributed, find the 10% value at risk (i.e., the maximum loss with 90% probability) in a given month. In how many months are the observed returns in fact below this value? Is this in accordance with the assumption of normality?
- (f) Under the assumption that returns are normally distributed, find the 5% value at risk in a given month. In how many months are the observed returns in fact below this value? Is this in accordance with the assumption of normality?
- (g) Under the assumption of normality, how often would it occur on average that the monthly return is lower than the minimum monthly return observed in the data set?
- (h) Under the assumption of normality, how often would it occur on average that the monthly return is higher than the maximum monthly return observed in the data set?

Exercise 3.5. In this problem you have to look at the stocks of three large U.S. companies: The Coca-Cola Company (ticker symbol KO), Pepsico, Inc (PEP), and The Home Depot, Inc (HD).

- (a) Download from Yahoo Finance the adjusted closing prices for each of the three stocks in every month from December 2013 to December 2023. Use the adjusted closing prices to calculate the rates of return for each of the three stocks in every month from January 2014 to December 2023.
- (b) What are the minimum and maximum monthly returns for each stock?
- (c) Construct for each stock two histograms of the monthly rates of return: one with bins given by $-8\%, -7\%, -6\%, \dots, 8\%$ and the other with bins given by $-10\%, -8\%, -6\%, \dots, 12\%$. Do the histograms resemble the probability density function of a normal distribution?
- (d) Compute for each stock the arithmetic and geometric average monthly rate of return as well as the sample variance and standard deviation. Annualize these statistics appropriately.
- (e) Compute for every pair of stocks the sample covariance and correlation between the monthly rates of return.

Exercise 3.6. The expectation and standard deviation of the future returns on a stock are sometimes estimated from a time series of past returns on the stock. This exercise investigates the role of the length of the time series (here: 5 years vs. 10 years) and the frequency of the observations in the time series (here: weekly vs. monthly).

The analysis is to be made on the stocks of the Danish pharmaceutical company Novo Nordisk that are listed on Nasdaq in New York under the ticker symbol NVO (the company's stocks are also listed on the Nasdaq Copenhagen stock exchange). The relevant stock prices for answering the following questions can be downloaded from Yahoo Finance; use the adjusted closing prices.

- (a) Download monthly stock prices for Novo Nordisk for the period from end of December 2013 to end of December 2023 and import them into an Excel file. Construct a graph showing how the stock price has changed over the period. Compute the rate of return for each month from January 2014 to December 2023.
- (b) Based on the 10 years of monthly returns, compute the sample average, standard deviation, kurtosis, and skewness (or simply get all descriptive statistics at once using 'Descriptive Statistics' from the 'Data Analysis' command under the 'Data' tab in Excel). Annualize your estimates of the expected return and standard deviation. Construct a histogram of the monthly returns.
- (c) Answer question (b) again, but only using the returns in the most recent 5 years, i.e., the 60 months from January 2019 to December 2023. Compare with your results in (b).
- (d) Now download weekly stock prices for the period from end of December 2013 to end of December 2023. Answer question (b) again using these weekly returns. Compare with the results based on the monthly returns over the same period.

Exercise 3.7. You manage a portfolio valued at \$100 mill. The standard deviation of the portfolio return is 8% per year. Assume that returns are normally distributed. You strive to deliver a positive return with at least 80% probability. What must be the expected return of the portfolio?

Exercise 3.8. The return on a levered position in a stock was derived in Section 2.7. Suppose that you can borrow at the riskfree rate, i.e., $r_{\text{loan}} = r_f$. Show that the Sharpe ratio of a levered position is then independent of the leverage ratio and therefore equal to the Sharpe ratio of the stock. How is the Sharpe ratio of the levered position different from that of the stock if the loan rate exceeds the riskfree rate?

Exercise 3.9. Suppose that X and Y are random variables related through the equation $X = a + bY$, where a and b are constants. By using (3.41), show that $\text{Corr}[X, Y] = 1$ if $b > 0$ and $\text{Corr}[X, Y] = -1$ if $b < 0$.

Exercise 3.10. If an asset over a period of some length T has a rate of return r , which is normally distributed with mean μ_r and standard deviation σ_r , then the value at risk is defined by $\text{VaR} = \mu_r + \sigma_r \times N^{-1}(p)$. Here p is the probability that the return is lower (more negative) than VaR , and $N^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution.

- (a) Over the next month, the rate of return on stocks in the company InYaDreams is normally distributed with an expectation of 2% and a standard deviation of 6%. What is the probability that the return is negative? What is the value at risk on this stock over the next month if $p = 5\%$? What if $p = 1\%$?

Let r_s denote the rate of return on a given stock over a given period, and assume that $r_s \sim N(\mu_s, \sigma_s^2)$. Let VaR_s denote the value at risk for this stock. In the following consider a levered investment in the stock over the same period. You invest E of your own money and borrow L at an interest rate of r_l over the period.

- (b) Explain why the value at risk of the levered investment is given by

$$\text{VaR} = \left(1 + \frac{L}{E}\right) \text{VaR}_s - \frac{L}{E} r_l.$$

- (c) Suppose you take a levered position in InYaDreams stocks with the characteristics stated in (a). The borrowing rate is $r_l = 0.2\%$ per month. With $p = 5\%$, what is the value at risk of your levered position if $L/E = 1$? What if $L/E = 5$? Discuss your results.

Exercise 3.11. Let r denote the rate of return on an asset over a month. Recall that the 5% one-month Value-at-Risk is defined as the value VaR for which $\text{Prob}(r \leq \text{VaR}) = 0.05$ and that $N^{-1}(0.05) = -1.645$. Assume that the monthly log-return is normally distributed with expectation m and variance s^2 , i.e. $\ln(1+r) \sim N(m, s^2)$.

- (a) Show that the 5% one-month Value-at-Risk is given by

$$\text{VaR} = \exp\{m - 1.645 s\} - 1.$$

- (b) Assume that returns in different months are independent of each other. Explain why the 5% Value-at-Risk over a period of T months is given by

$$\text{VaR}_T = \exp\{m T - 1.645 s \sqrt{T}\} - 1.$$

Exercise 3.12. Suppose the rate of return on Disney stocks in any given year is normally distributed with mean 0.10 (or 10%) and standard deviation 0.30 (or 30%). The returns on Disney stocks in different years are independent of each other.

- (a) What is the probability that the rate of return on Disney stocks over the next year is between 0 and 0.2?
- (b) What is the probability that the log-return on Disney stocks over the next year is positive?
- (c) What is the probability that the square of the rate of return on Disney stocks over the next year is larger than 0.25?
- (d) What is the expected rate of return on Disney stocks over a 5-year period?

Suppose the rates of return on Disney and Netflix stocks in any given year are jointly normally distributed. For Disney, the expectation and standard deviation are as stated above. For Netflix, the rate of return has an expectation of 0.06 (or 6%) and a standard deviation is 0.40 (or 40%). The correlation between the rates of return on the two stocks is 0.5.

- (e) What is the probability that, over the next year, the rate of return on Disney is more than two times the rate of return on Netflix?

CHAPTER 4

Portfolios

Most investors hold a portfolio of several assets. The main motive for doing so is diversification: the return on a portfolio generally has a lower risk than the typical asset in the portfolio. The intuition is simple. Should one of the assets turn out to deliver a highly negative return, there is a good chance that some of the other assets deliver better returns so that the overall portfolio return is not so bad. This reduces the risk of ending up with a highly negative return. Of course, the mechanism works the other way as well so by owning a portfolio of different assets, the chance of getting a very large return is generally smaller than for a single asset. Many investors are willing to give up some chance of very large returns to reduce the risk of highly negative returns. Investors not only care about risk but also about expected returns. Ultimately they want to know which portfolio offers the most attractive risk-return tradeoff.

Before we can start searching for the optimal portfolio of a given investor, we need to know more about how a given portfolio's expected return, return variance, and other risk measures are determined from the individual assets' expected returns, variances, etc. For simplicity, we first look at portfolios of two assets in Section 4.1. The step to a general number of assets is taken in Section 4.2. High-dimensional portfolio mathematics is simplified by the use of vectors and matrices and these concepts are also very useful for handling large portfolios in Excel and similar computational software. Therefore, Section 4.2 also introduces vectors and matrices as well as the associated computational rules needed in portfolio theory. Section 4.3 briefly considers higher-order moments of portfolio returns. Section 4.4 focuses on the idea of diversification and provides a preliminary discussion of how much the risk can be reduced by forming portfolios. Finally, Section 4.5 presents the concepts of replicating portfolios and tracking portfolios, which are portfolios constructed to match the value or return on some benchmark as closely as possible. The important concept of an arbitrage is also defined in this section.

4.1 Two-asset portfolio mathematics

This section studies the properties of buy-and-hold portfolios of two assets. First, we consider the case where both assets are risky. Subsequently, we study the special case where one of the assets is riskfree.

4.1.1 Portfolios of two risky assets

Recall from Eq. (2.20) that the rate of return on a portfolio of two assets is

$$r(w) = wr_1 + (1 - w)r_2, \quad (4.1)$$

where r_i is the rate of return on asset i and w is the fraction of the total value of the portfolio which is invested in asset 1, implying that a fraction $1 - w$ is invested in asset 2. In words, the rate of return on a portfolio is a value-weighted average of the rates of return on the assets in the portfolio. This is true both if you use known, historical returns and if you use possible future returns. In the latter case, this implies that the relation (4.1) holds no matter what the returns of the two assets turn out to be.

We focus for now on the expectation (or mean) and the variance of the portfolio return. Let $\mu_1 = E[r_1]$ and $\sigma_1^2 = \text{Var}[r_1]$ denote the expectation and variance, respectively, of the rate of return of asset 1. Similarly, μ_2 and σ_2^2 denote the expectation and the variance of the return of asset 2. We assume that both σ_1 and σ_2 are strictly positive so that the assets indeed are risky. We let ρ be the correlation between the returns of the two assets so that the covariance is $\text{Cov}[r_1, r_2] = \rho\sigma_1\sigma_2$. The next theorem provides formulas for the expectation, variance, and standard deviation of a two-asset portfolio.

Theorem 4.1

Consider a buy-and-hold portfolio of two risky assets. Let w denote the portfolio weight of asset 1 so that $1 - w$ is the portfolio weight of asset 2. Then the expectation $\mu(w)$, the variance $\sigma^2(w)$, and the standard deviation $\sigma(w)$ of the portfolio's rate of return are

$$\mu(w) = w\mu_1 + (1 - w)\mu_2, \quad (4.2)$$

$$\sigma^2(w) = w^2\sigma_1^2 + (1 - w)^2\sigma_2^2 + 2w(1 - w)\rho\sigma_1\sigma_2, \quad (4.3)$$

$$\sigma(w) = \sqrt{w^2\sigma_1^2 + (1 - w)^2\sigma_2^2 + 2w(1 - w)\rho\sigma_1\sigma_2}. \quad (4.4)$$

In particular if $0 \leq w \leq 1$, the standard deviation satisfies

$$\sigma(w) \leq w\sigma_1 + (1 - w)\sigma_2. \quad (4.5)$$

Proof

By using the rule (3.47) for expectations of a sum, we find that the expected return on the portfolio is

$$\mu(w) = E[r(w)] = E[wr_1 + (1 - w)r_2] = wE[r_1] + (1 - w)E[r_2] = w\mu_1 + (1 - w)\mu_2.$$

By applying the rule (3.48), we can compute the variance of the portfolio return as

$$\begin{aligned} \sigma^2(w) &= \text{Var}[r(w)] = \text{Var}[wr_1 + (1 - w)r_2] \\ &= w^2\text{Var}[r_1] + (1 - w)^2\text{Var}[r_2] + 2w(1 - w)\text{Cov}[r_1, r_2] \\ &= w^2\sigma_1^2 + (1 - w)^2\sigma_2^2 + 2w(1 - w)\rho\sigma_1\sigma_2. \end{aligned}$$

As always, the standard deviation is simply the square root of the variance.

By definition, $\rho \leq 1$. If $0 \leq w \leq 1$, we thus have $2w(1-w)\rho\sigma_1\sigma_2 \leq 2w(1-w)\sigma_1\sigma_2$ and therefore

$$\sigma^2(w) \leq w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\sigma_1\sigma_2 = (w\sigma_1 + (1-w)\sigma_2)^2,$$

which leads to Eq. (4.5).¹

The theorem shows that expected return on the portfolio is a weighted average of the expected returns on the assets in the portfolio, but the same relation does generally not hold for the variance or for the standard deviation.

The inequality (4.5) is interesting. Note that when w is between 0 and 1, then $1-w$ is also between 0 and 1. In this case, both portfolio weights are positive, so it is a “long-only portfolio.” The inequality says that for long-only portfolios, the standard deviation is less than the weighted average of the assets’ standard deviations. In that sense, you can reduce risk by forming portfolios, that is by *diversifying* your investment. The proof of the inequality suggests that the smaller the asset correlation ρ , the larger the reduction in risk. In particular, you can diversify away a lot of risk by investing in assets that are negatively correlated. This makes intuitive sense: if the return on asset 1 turns out to be low, the return on asset 2 is typically going to be high, and vice versa. Therefore, you are unlikely to lose a lot—or gain a lot—on the portfolio. Other things equal, the variance reduction lowers both the downside risk and the upside potential.

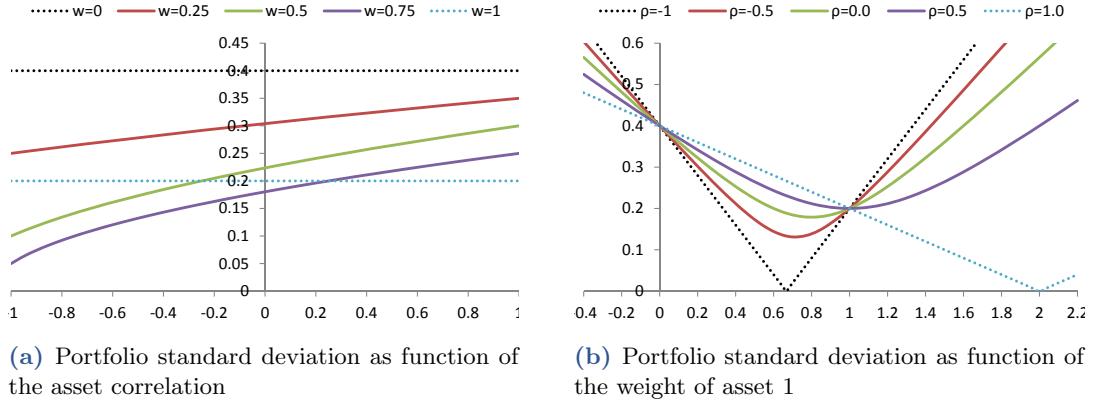
Example 4.1

Let us consider various portfolios of two assets. The return of asset 1 has a standard deviation of $\sigma_1 = 0.2$, whereas the standard deviation of the return of asset 2 is $\sigma_2 = 0.4$. The left panel of Figure 4.1 illustrates how the standard deviation $\sigma(w)$ of the portfolio return varies with the asset return correlation ρ . Each curve corresponds to a given portfolio weight w of asset 1. The black dotted curve labeled $w = 0$ corresponds to a full investment in asset 2 and thus the portfolio standard deviation coincides with the standard deviation of asset 2, irrespective of the correlation with asset 1 as this asset does not enter the portfolio. Likewise, the blue dotted curve labeled $w = 1$ corresponds to a full investment in asset 1 so that the portfolio standard deviation equals the standard deviation of that asset. The colored curves correspond to portfolio weights of 0.25, 0.5, and 0.75, respectively. Note that for some combinations of w and ρ , the portfolio standard deviation falls below the standard deviation of both assets, which again illustrates the benefits of diversification. The right panel of Figure 4.1 is discussed in Example 4.2.

4.1.2 Minimum-variance portfolio

Which portfolio of the two risky assets leads to the lowest possible variance? We can figure that out by minimizing the variance expression (4.3). We summarize the results in

¹In general $\sqrt{x^2} = |x|$, where the right-hand side is the absolute value of x . For example, $\sqrt{(-3)^2} = \sqrt{9} = 3 = |-3|$. If $x > 0$, then $\sqrt{x^2} = x$. However, when $0 \leq w \leq 1$ and $\sigma_1, \sigma_2 \geq 0$, it is clear that $w\sigma_1 + (1-w)\sigma_2$ is non-negative, so we do not need to take the absolute value.



(a) Portfolio standard deviation as function of the asset correlation

(b) Portfolio standard deviation as function of the weight of asset 1

Figure 4.1: The standard deviation of a two-asset portfolio return.

The figures show how the standard deviation of a two-asset portfolio's rate of return depend on the return correlation ρ of the two assets and the portfolio weight w of the first asset. The left panel has the correlation coefficient along the horizontal axis, whereas the right panel has the portfolio weight. The two assets have return standard deviations of $\sigma_1 = 0.2$ and $\sigma_2 = 0.4$, respectively.

the next theorem.

Theorem 4.2

Among the buy-and-hold portfolios of two risky assets, the portfolio with the lowest return variance has a weight of w_{\min} in asset 1 and $1 - w_{\min}$ in asset 2, where

$$w_{\min} = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \quad 1 - w_{\min} = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \quad (4.6)$$

The minimum variance is

$$\sigma^2(w_{\min}) = \frac{(1 - \rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \quad (4.7)$$

and the portfolio has an expected rate of return of

$$\mu(w_{\min}) = \frac{\mu_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + \mu_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \quad (4.8)$$

In the special case of a perfect negative or perfect positive correlation, the minimum-variance portfolio is riskfree. If $\rho = -1$, this is obtained with weights

$$w_{\min} = \frac{\sigma_2}{\sigma_1 + \sigma_2}, \quad 1 - w_{\min} = \frac{\sigma_1}{\sigma_1 + \sigma_2}, \quad (4.9)$$

whereas if $\rho = 1$ (and $\sigma_1 \neq \sigma_2$) the weights are

$$w_{\min} = -\frac{\sigma_2}{\sigma_1 - \sigma_2}, \quad 1 - w_{\min} = \frac{\sigma_1}{\sigma_1 - \sigma_2}. \quad (4.10)$$

Proof

Starting from the variance formula (4.3), the derivative with respect to w is

$$\begin{aligned}\frac{\partial \sigma^2(w)}{\partial w} &= 2w\sigma_1^2 - 2(1-w)\sigma_2^2 + 2(1-2w)\rho\sigma_1\sigma_2 \\ &= 2w(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) - 2\sigma_2^2 + 2\rho\sigma_1\sigma_2.\end{aligned}$$

If we put the derivative equal to zero and solve for w , we find that the minimum-variance portfolio is given by (4.6). The expression (4.7) for the minimum variance follows from a tedious calculation which Exercise 4.1 asks you to provide. The formula (4.8) for the expected return follows from substituting w_{\min} and $1 - w_{\min}$ into the general expression $\mu(w) = w\mu_1 + (1 - w)\mu_2$.

If we substitute $\rho = -1$ into the formula for w_{\min} in (4.6), we get

$$w_{\min} = \frac{\sigma_2^2 + \sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2} = \frac{\sigma_2(\sigma_1 + \sigma_2)}{(\sigma_1 + \sigma_2)^2} = \frac{\sigma_2}{\sigma_1 + \sigma_2}$$

and thus

$$1 - w_{\min} = 1 - \frac{\sigma_2}{\sigma_1 + \sigma_2} = \frac{\sigma_1}{\sigma_1 + \sigma_2}.$$

Similarly, if we substitute $\rho = 1$ into the formula for w_{\min} in (4.6), we get

$$w_{\min} = \frac{\sigma_2^2 - \sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2} = -\frac{\sigma_2(\sigma_1 - \sigma_2)}{(\sigma_1 - \sigma_2)^2} = -\frac{\sigma_2}{\sigma_1 - \sigma_2}$$

and thus

$$1 - w_{\min} = 1 + \frac{\sigma_2}{\sigma_1 - \sigma_2} = \frac{\sigma_1}{\sigma_1 - \sigma_2}.$$

For some values of the correlation coefficient ρ , the minimum variance will be below both σ_1 and σ_2 . This reflects the diversification of risk you can obtain when combining different assets. In fact, the theorem states that if the correlation is either -1 or $+1$, a portfolio with a zero standard deviation exists, i.e. a riskfree portfolio.

Recall that $\rho = -1$ means that the two returns are linearly related in the sense that $r_2 = a - br_1$ for some positive constant b and some other, positive or negative, constant a (see Exercise 3.9). If the return of asset 1 is high, then the return of asset 2 will be low and vice versa. With the right mix of the two assets, as characterized by the weights in (4.9) you get a completely riskfree portfolio. We can verify this as follows: With $r_2 = a - br_1$, we have $\sigma_2 = b\sigma_1$. Substituting this into (4.9), we get $w_{\min} = b/(1+b)$ and $1 - w_{\min} = 1/(1+b)$. Hence the return on the portfolio is

$$\begin{aligned}r(w_{\min}) &= w_{\min}r_1 + (1 - w_{\min})r_2 = w_{\min}r_1 + (1 - w_{\min})(a - br_1) \\ &= \frac{b}{1+b}r_1 + \frac{1}{1+b}(a - br_1) = \frac{a}{1+b},\end{aligned}$$

which indeed is a constant and thus has zero standard deviation. Note that, in this case, both portfolio weights are between zero and one, so this portfolio does not involve short-selling.

The other extreme case $\rho = 1$ means that the returns are related through $r_2 = a + br_1$ with $b > 0$. Again, $\sigma_2 = b\sigma_1$ but, substituting this into (4.10), we get $w_{\min} = -b/(1-b)$ and $1 - w_{\min} = 1/(1-b)$. The portfolio return can be shown to be $r(w_{\min}) = a/(1-b)$ which is a constant, thus representing a riskfree return. If the return of asset 1 is high, then the return of asset 2 is also high. And if the return of asset 1 is low, then the return of asset 2 is also low. To obtain a riskfree position, you need a long position in one asset and a short position in the other asset. This is also satisfied by the weights in (4.10). If $\sigma_1 > \sigma_2$, then $w_{\min} < 0$. If $\sigma_1 < \sigma_2$, then $1 - w_{\min} < 0$. Note that, in this case with $\rho = 1$, if the two assets have identical standard deviations, i.e. $\sigma_1 = \sigma_2$, then any portfolio of the two assets is having a standard deviation equal to that joint value.

In general, if you only consider two-asset portfolios without short-selling, the minimum variance will be increasing in the correlation. The lower the correlation between the assets you invest in, the more risk can be diversified away. By opening up for short-selling, also high positive correlations offer opportunities for significant reductions of risk.

Example 4.2

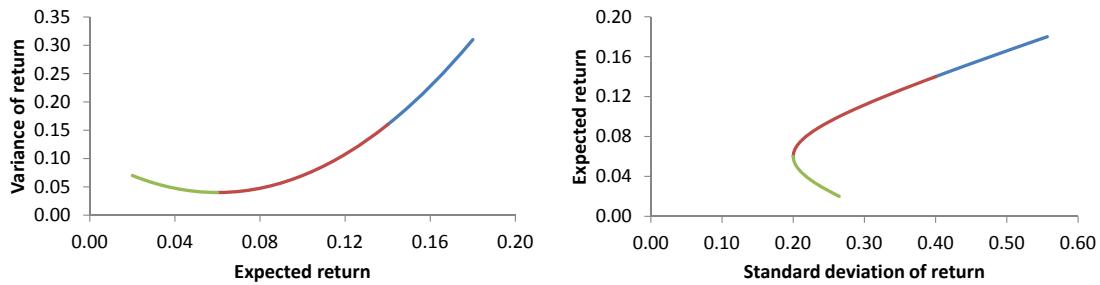
In continuation of Example 4.1, let us consider how the standard deviation of a two-asset portfolio's return varies with the portfolio weight w of the first asset. This is illustrated in the right panel of Figure 4.1. Each curve corresponds to a given asset return correlation. Of course, when the weight of asset 1 is zero, the portfolio equals a full investment in asset 2, so the portfolio standard deviation is identical to $\sigma_2 = 0.4$, no matter what the correlation is. Similarly, when $w = 1$, the portfolio is just a full investment in asset 1 and thus has a standard deviation equal to $\sigma_1 = 0.2$.

With a portfolio weight between 0 and 1, the figure confirms that the portfolio standard deviation is increasing in the correlation. In contrast, if the portfolio weight is above 1 (meaning asset 2 is shorted) or below 0 (asset 1 is shorted), the portfolio standard deviation is *decreasing* in the correlation. We can see this mathematically from (4.3) because either w or $1 - w$ is then negative and the other positive, so the correlation in the last term has a negative multiplier. If your portfolio consists of a long position in one asset and a short position in the other asset, the portfolio risk is lowest if the two assets tend to both provide high returns or both provide low returns.

For each value of the correlation coefficient, the minimum of the portfolio standard deviation and the associated portfolio weights are computed from the formulas preceding this example. The results are:

Correlation coefficient, ρ	-1	-0.5	0	0.5	1
Minimum portfolio variance	0.0000	0.0171	0.0320	0.0400	0.0000
Minimum portfolio standard deviation	0.0000	0.1309	0.1789	0.2000	0.0000
Minimum-variance weight of asset 1	0.6667	0.7143	0.8000	1.0000	2.0000
Minimum-variance weight of asset 2	0.3333	0.2857	0.2000	0.0000	-1.0000

This example confirms the observation that a completely riskfree portfolio can be constructed from two perfectly negatively correlated assets without taking any short positions, whereas shorting is necessary to form a riskfree portfolio of two perfectly positively correlated assets. The table also shows that if $\rho = 0.5$, no combination of the two assets has a standard deviation below the lower of the asset's standard deviations, whereas for the other correlation coefficients a lower portfolio standard deviation is obtainable.



(a) The variance as a function of the expected return

(b) The expected return as a function of the standard deviation

Figure 4.2: Risk and expected return.

The graphs depict the relations between the expectation, variance, and standard deviation of various portfolios given the information in Example 4.3.

4.1.3 The risk-return tradeoff

Of course, an investor should not just care about the risk of her portfolio, but also about its expected return. Probably you would be willing to take a little more risk if you are generously compensated for it by getting a much higher expected return. When searching for the optimal diversification and thus the lowest risk, you might end up with a portfolio with low expected return. Therefore you should really consider the tradeoff between risk and expected return as done in the following example.

Example 4.3

Continuing Examples 4.1 and 4.2, assume that the expected rate of return on asset 1 is $\mu_1 = 0.06 = 6\%$ and the expected rate of return on asset 2 is $\mu_2 = 0.14 = 14\%$. The standard deviations are still $\sigma_1 = 0.2$ and $\sigma_2 = 0.4$. First assume that the correlation between the two returns is $\rho = 0.5$.

Table 4.1 shows the expectation, variance, and standard deviation of the rate of return on various portfolios of the two assets. The left panel of Figure 4.2 shows the variance of the portfolio as a function of the expected rate of return. The right panel shows the expected rate of return as a function of the standard deviation. In both graphs, each point on the curve corresponds to a specific portfolio of the two assets, i.e., a specific portfolio weight w . The blue part of each curve represents portfolios with a short position in asset 1, the green part portfolios with a short position in asset 2, and the red part corresponds to portfolios where both assets have a weight between 0 and 1.

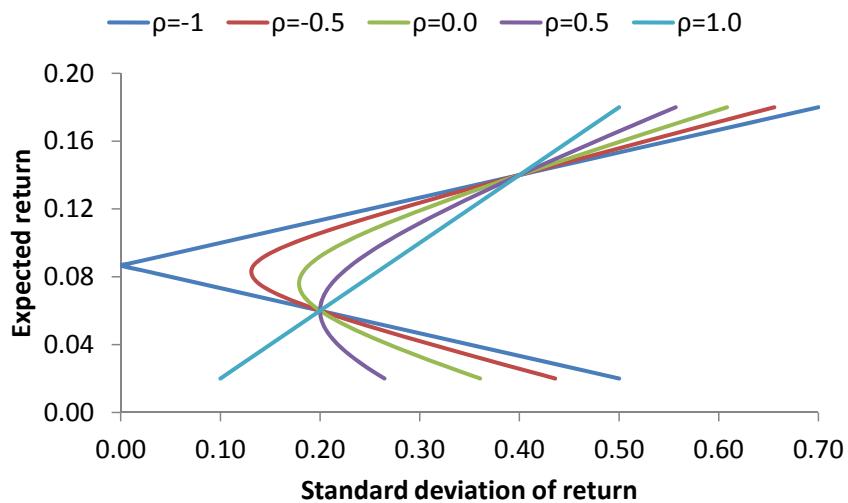
Next, let us see how the tradeoff between risk (standard deviation) and expected return is affected by the correlation coefficient. Figure 4.3 depicts the combinations of the expected return and the standard deviation that can be achieved for each of the five correlation coefficients $-1, -0.5, 0, 0.5$, and 1 . For all the portfolios with no short positions we see again that the standard deviation increases with the correlation, i.e., lower correlation coefficients allow for better diversification of risk.

In the above example, the portfolio variance seems to be a quadratic function of the expected portfolio return. This is no coincidence as the next theorem shows.

Portfolio weights		Expected return	Return variance	Return std dev
Asset 1	Asset 2			
-0.5	1.5	0.1800	0.3100	0.5568
-0.2	1.2	0.1560	0.2128	0.4613
-0.1	1.1	0.1480	0.1852	0.4303
0	1	0.1400	0.1600	0.4000
0.1	0.9	0.1320	0.1372	0.3704
0.2	0.8	0.1240	0.1168	0.3418
0.5	0.5	0.1000	0.0700	0.2646
0.8	0.2	0.0760	0.0448	0.2117
0.9	0.1	0.0680	0.0412	0.2030
1	0	0.0600	0.0400	0.2000
1.1	-0.1	0.0520	0.0412	0.2030
1.2	-0.2	0.0440	0.0448	0.2117
1.5	-0.5	0.0200	0.0700	0.2646

Table 4.1: Risk and expected return of portfolios.

The table lists the expectation, variance, and standard deviation of various portfolios given the information in Example 4.3. The correlation between the two assets is assumed to be 0.5.

**Figure 4.3:** The impact of correlation.

The graph illustrates the relation between the expectation and standard deviation of various portfolios for different correlations. See Example 4.3.

Theorem 4.3

Consider two risky assets with different expected returns, i.e. $\mu_1 \neq \mu_2$. The return variance of a portfolio of the two assets is a quadratic function of the portfolio's expected rate of return, i.e. for every asset 1 weight w , we have

$$\sigma^2(w) = K_0 + K_1 \mu(w) + K_2 \mu(w)^2, \quad (4.11)$$

for some constants K_0 , K_1 , and K_2 stated in the proof. If either the asset correlation satisfies $\rho < 1$ or the two assets have different standard deviations, i.e. $\sigma_1 \neq \sigma_2$, then $K_2 > 0$.

Proof

First, rewrite (4.2) as

$$\mu(w) = \mu_2 + w(\mu_1 - \mu_2). \quad (4.12)$$

Assuming that $\mu_1 \neq \mu_2$, this implies that

$$w = \frac{\mu(w) - \mu_2}{\mu_1 - \mu_2} \quad (4.13)$$

and thus

$$1 - w = 1 - \frac{\mu(w) - \mu_2}{\mu_1 - \mu_2} = \frac{\mu_1 - \mu_2 - (\mu(w) - \mu_2)}{\mu_1 - \mu_2} = \frac{\mu_1 - \mu(w)}{\mu_1 - \mu_2}.$$

Substituting these relations into (4.3), the variance can be rewritten as

$$\begin{aligned} \sigma^2(w) &= \left(\frac{\mu(w) - \mu_2}{\mu_1 - \mu_2} \right)^2 \sigma_1^2 + \left(\frac{\mu_1 - \mu(w)}{\mu_1 - \mu_2} \right)^2 \sigma_2^2 + 2 \frac{\mu(w) - \mu_2}{\mu_1 - \mu_2} \frac{\mu_1 - \mu(w)}{\mu_1 - \mu_2} \rho \sigma_1 \sigma_2 \\ &= K_0 + K_1 \mu(w) + K_2 \mu(w)^2, \end{aligned}$$

where we have skipped some intermediate steps and introduced the constants

$$\begin{aligned} K_0 &= \frac{1}{(\mu_1 - \mu_2)^2} \left(\mu_2^2 \sigma_1^2 + \mu_1^2 \sigma_2^2 - 2\mu_1 \mu_2 \rho \sigma_1 \sigma_2 \right), \\ K_1 &= -\frac{2}{(\mu_1 - \mu_2)^2} \left(\mu_2 \sigma_1^2 + \mu_1 \sigma_2^2 + (\mu_1 + \mu_2) \rho \sigma_1 \sigma_2 \right), \\ K_2 &= \frac{1}{(\mu_1 - \mu_2)^2} \left(\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2 \right). \end{aligned}$$

Note that

$$\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2 \geq \sigma_1^2 + \sigma_2^2 - 2\sigma_1 \sigma_2 = (\sigma_1 - \sigma_2)^2 \geq 0$$

which shows that $K_2 \geq 0$. If $\rho < 1$, the first inequality above is strict; if $\sigma_1 \neq \sigma_2$, the second is strict. In either case, $K_2 > 0$.

Equation (4.11) shows that the return variance is a quadratic function of the return expectation for any two-asset portfolios. In a diagram with expected return along the

horizontal axis and return variance along the vertical axis, the combinations that can be obtained by forming different portfolios of the two assets (that is, by varying w) will trace out a parabola with upward-sloping branches, as we saw it in the left panel of Figure 4.2. The branches are upward-sloping since $K_2 > 0$ (except for the special case of $\rho = 1$ and $\sigma_1 = \sigma_2$).

On the other hand, the relation (4.11) implies that in a diagram with standard deviation along the horizontal axis and expected return along the vertical axis the portfolios will form a hyperbola. This is consistent with the right panel of Figure 4.2.

A popular measure of the risk-return tradeoff of a portfolio is the Sharpe ratio defined in Eq. (3.19), i.e. the ratio of the excess expected return to the standard deviation. For the two-asset portfolio with w denoting the weight of asset 1, the Sharpe ratio is

$$\text{SR}(w) = \frac{\mu(w) - r_f}{\sigma(w)}.$$

Let $\text{SR}_1 = (\mu_1 - r_f)/\sigma_1$ and $\text{SR}_2 = (\mu_2 - r_f)/\sigma_2$ denote the Sharpe ratios of the two assets in the portfolio. Since $\mu_1 - r_f = \sigma_1 \text{SR}_1$ and $\mu_2 - r_f = \sigma_2 \text{SR}_2$, the excess expected portfolio return can be written as

$$\mu(w) - r_f = w\mu_1 + (1-w)\mu_2 - r_f = w(\mu_1 - r_f) + (1-w)(\mu_2 - r_f) = w\sigma_1 \text{SR}_1 + (1-w)\sigma_2 \text{SR}_2,$$

so the Sharpe ratio of the portfolio is linked to the Sharpe ratios of the assets by the relation

$$\text{SR}(w) = \frac{w\sigma_1 \text{SR}_1 + (1-w)\sigma_2 \text{SR}_2}{\sigma(w)}. \quad (4.14)$$

If we focus on long-only portfolios, i.e. $w \in [0,1]$, we can use the inequality (4.5) to conclude that

$$\text{SR}(w) \geq \frac{w\sigma_1 \text{SR}_1 + (1-w)\sigma_2 \text{SR}_2}{w\sigma_1 + (1-w)\sigma_2} = \frac{w\sigma_1}{w\sigma_1 + (1-w)\sigma_2} \text{SR}_1 + \frac{(1-w)\sigma_2}{w\sigma_1 + (1-w)\sigma_2} \text{SR}_2,$$

which shows that the portfolio's Sharpe ratio exceeds a weighted average of the Sharpe ratio of the assets in the portfolio with the weights depending both on the portfolio weights and the standard deviations. Typically, the portfolio's Sharpe ratio is also exceeding the value-weighted average of the asset Sharpe ratios, i.e., $\text{SR}(w) \geq w \text{SR}_1 + (1-w) \text{SR}_2$, but this is violated for certain combinations of input parameters.

It is often assumed in investment theory that many investors prefer to invest in the portfolio with the largest possible Sharpe ratio as this gives the highest compensation for risk. The next theorem characterizes the two-asset portfolio with the largest Sharpe ratio.

Theorem 4.4

Assume two assets satisfy the condition

$$(\sigma_2 - \rho\sigma_1) \text{SR}_1 + (\sigma_1 - \rho\sigma_2) \text{SR}_2 > 0. \quad (4.15)$$

Then the two-asset portfolio with the largest Sharpe ratio is obtained when the portfolio weight of asset 1 is

$$w_{\max\text{SR}} = \frac{\text{SR}_1 - \rho \text{SR}_2}{\text{SR}_1 - \rho \text{SR}_2 + \frac{\sigma_1}{\sigma_2} (\text{SR}_2 - \rho \text{SR}_1)}. \quad (4.16)$$

The maximum value of the Sharpe ratio is

$$\text{SR}_{\max} = \text{SR}(w_{\max\text{SR}}) = \sqrt{\frac{1}{1-\rho^2} (\text{SR}_1^2 + \text{SR}_2^2 - 2\rho \text{SR}_1 \text{SR}_2)}. \quad (4.17)$$

Proof

The proof is in principle straightforward: Solve $\text{SR}'(w) = 0$ for $w = w_{\max\text{SR}}$. The condition (4.15) ensures that this gives a maximum of $\text{SR}(w)$ and not a minimum. Next, substitute that value of w into $\text{SR}(w)$ to find SR_{\max} . However, the details of the calculations are tedious so we only outline the main steps. In Chapter 7 we provide a complete proof of Theorem 7.6 that characterizes the portfolio of N assets that produces the largest Sharpe ratio. Of course, Theorem 4.4 then follows as a special case of Theorem 7.6.

Using the expression (4.14) for the portfolio Sharpe ratio $\text{SR}(w)$, we differentiate with respect to w :

$$\text{SR}'(w) = \frac{(\sigma_1 \text{SR}_1 - \sigma_2 \text{SR}_2)\sigma(w) - (w\sigma_1 \text{SR}_1 + (1-w)\sigma_2 \text{SR}_2)\sigma'(w)}{\sigma(w)^2}.$$

Then we solve the equation $\text{SR}'(w) = 0$ for w , which is the same as solving

$$(\sigma_1 \text{SR}_1 - \sigma_2 \text{SR}_2)\sigma(w) = (w\sigma_1 \text{SR}_1 + (1-w)\sigma_2 \text{SR}_2)\sigma'(w).$$

Differentiating the standard deviation in (4.4), we get

$$\sigma'(w) = \frac{w\sigma_1^2 - (1-w)\sigma_2^2 + (1-2w)\rho\sigma_1\sigma_2}{\sigma(w)}.$$

We plug that into the preceding equation and simplify. After rather tedious calculations, we end up with the expression (4.16) for w . We substitute (4.16) into (4.14) and rewrite considerably to finally reach (4.17).

Of course, the maximum portfolio Sharpe ratio SR_{\max} is always greater than or equal to both SR_1 and SR_2 since a full investment in one of the asset is just a very special portfolio with zero weight in the other asset. If the two assets have the same standard deviation, the first asset will have a larger weight than asset 2 (which means $w_{\max\text{SR}} > 1/2$) if and only if $\text{SR}_1 > \text{SR}_2$. But when the two assets have different standard deviations, the asset with the largest Sharpe ratio may end up having a lower weight in the max-Sharpe portfolio. If the two assets have identical Sharpe ratios, the asset with the lowest standard deviation will have the largest weight. Also note that the square of the maximum Sharpe ratio SR_{\max}^2 equals the sum of the squared Sharpe ratios of the assets, $\text{SR}_1^2 + \text{SR}_2^2$, if the two assets are uncorrelated, but not when they are correlated. If both SR_1 and SR_2 are positive, we have $\text{SR}_{\max} \leq \text{SR}_1 + \text{SR}_2$ no matter what the correlation is. Next we provide an example of how the portfolio Sharpe ratio depends on the portfolio weights of the assets.

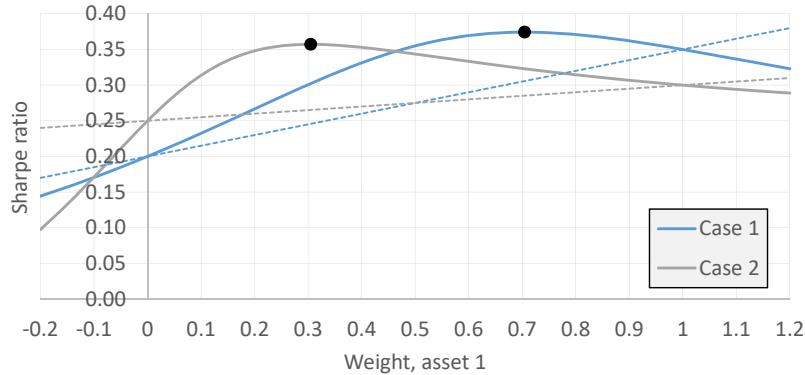


Figure 4.4: Sharpe ratios and portfolio weights.

The solid curves show the Sharpe ratio $\text{SR}(w)$ of a two-asset portfolio as a function of the portfolio weight w of the first asset. The black dot indicates the maximum Sharpe ratio. The dotted lines represent the weighted average of the Sharpe ratios of the assets, $w \text{SR}_1 + (1 - w) \text{SR}_2$. In Case 1, the parameter values are $\sigma_1 = \sigma_2 = 0.4$, $\text{SR}_1 = 0.35$, and $\text{SR}_2 = 0.2$. In Case 2, the values are $\sigma_1 = 0.6$, $\sigma_2 = 0.2$, $\text{SR}_1 = 0.3$, and $\text{SR}_2 = 0.25$. In both cases, $\rho = 0.2$.

Example 4.4

Consider two risky assets with a return correlation of $\rho = 0.2$. First, as Case 1, assume the two assets have identical standard deviations, $\sigma_1 = \sigma_2 = 0.4$, and Sharpe ratios of $\text{SR}_1 = 0.35$ and $\text{SR}_2 = 0.2$. The solid blue curve in Figure 4.4 shows how the Sharpe ratio $\text{SR}(w)$ of the portfolio depends on the portfolio weight w of asset 1. The dashed blue line reflects the weighted average of the Sharpe ratios of the assets, $w \text{SR}_1 + (1 - w) \text{SR}_2$. As explained above, the portfolio's Sharpe ratio exceeds the weighted average of the assets' Sharpe ratios as long as $0 < w < 1$. The maximum Sharpe ratio of 0.3743 is obtained for a weight $w \approx 0.705$, i.e. the portfolio has a larger weight on the asset with the largest Sharpe ratio. Note that the maximum portfolio Sharpe ratio is only slightly larger than the Sharpe ratio of asset 1.

In Case 2, asset 1 is assumed to have both a higher Sharpe ratio and a higher standard deviation than asset 2. More specifically, the Sharpe ratios are $\text{SR}_1 = 0.3$ and $\text{SR}_2 = 0.25$, and the standard deviations are $\sigma_1 = 0.6$ and $\sigma_2 = 0.2$. The grey curve shows the portfolio's Sharpe ratio in this case, where the maximum is obtained for $w \approx 0.305$. Although asset 1 has a higher Sharpe ratio than asset 2, asset 1 has a lower weight in the portfolio maximizing the Sharpe ratio due to its larger standard deviation. In Case 2 the difference between the maximum portfolio Sharpe ratio of 0.3572 and the largest asset Sharpe ratio of 0.3 is larger than the similar difference in Case 1. Also note that in both cases, portfolios that are equal-weighted or close to equal-weighted have Sharpe ratios only slightly below the maximum Sharpe ratio.

The correlation between the two assets is important for how much the Sharpe ratio can be improved by forming a portfolio. The next example illustrates that the improvement is largest when the assets are highly negatively or highly positively correlated. However, with a high positive correlation, the maximum portfolio Sharpe ratio is achieved by a portfolio

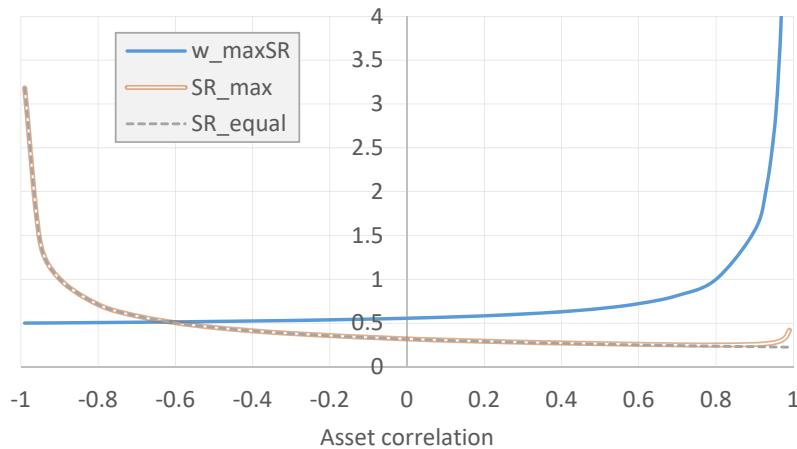


Figure 4.5: Sharpe ratios and asset correlation.

The figure considers two-asset portfolios. The blue curve shows the weight of the first asset that leads to the maximum portfolio Sharpe ratio, which is shown by the orange-band curve. The grey-dotted curve shows the Sharpe ratio of an equal-weighted portfolio of the two assets. The assets are assumed to have identical standard deviations of $\sigma_1 = \sigma_2 = 0.2$, asset 1 has a Sharpe ratio of $SR_1 = 0.25$, and asset 2 a Sharpe ratio of $SR_2 = 0.2$.

with a large negative position in one asset and, consequently, a large positive position in the other asset. Unless the correlation is close to +1, an equal-weighted portfolio generates a Sharpe ratio close to the maximum Sharpe ratio.

Example 4.5

Assume the two risky assets have identical standard deviations, $\sigma_1 = \sigma_2 = 0.2$, but that the Sharpe ratio is slightly larger for asset 1 than asset 2, $SR_1 = 0.25$ and $SR_2 = 0.2$. The orange-band curve in Figure 4.5 shows how the maximum portfolio Sharpe ratio SR_{\max} depends on the correlation between the two assets. The blue curve shows the portfolio weight of asset 1 in the portfolio maximizing the Sharpe ratio. Of course, we expect the optimal portfolio to be tilted towards asset 1 that has the largest Sharpe ratio of the two assets. But for negative correlations, the tilt is very small, so the optimal portfolio is almost equal-weighted. As the correlation is increased, the tilt becomes more distinct, and for correlations above 0.8, the portfolio involves shorting asset 2. For correlations close to 1, the portfolio is highly unbalanced. The grey-dotted curve shows the Sharpe ratio of an equal-weighted portfolio. As long as the correlation is below, say, 0.8, the Sharpe ratio of the equal-weighted portfolio is only little below the maximum Sharpe ratio. For a correlation of 0.95, the difference is somewhat larger, 0.228 compared to 0.277, and for a correlation of 0.99 the maximum Sharpe ratio of 0.419 is far above the 0.226 Sharpe ratio of the equal-weighted portfolio.

4.1.4 Portfolios of a riskfree and a risky asset

Now suppose you combine a risky asset and a riskfree asset into a portfolio. Let \hat{r} denote the rate of return on the risky asset with an expectation of $\hat{\mu} = E[\hat{r}]$ and a variance of

$\hat{\sigma}^2 = \text{Var}[\hat{r}]$. Let r_f denote the riskfree rate of return. The Sharpe ratio of the risky asset is $\widehat{\text{SR}} = (\hat{\mu} - r_f)/\hat{\sigma}$.

If we invest a fraction w of wealth in the risky asset and the remaining fraction $1 - w$ of wealth in the riskfree asset, the rate of return on the combined portfolio is

$$r(w) = w\hat{r} + (1 - w)r_f. \quad (4.18)$$

Since \hat{r} is the only random variable in this expression, we can easily calculate the expectation, variance, and standard deviation by using the rules in Eq. (3.50) directly on (4.18). Alternatively, we can think of the case where one of the assets is riskfree as a special case of the general two-asset case and apply Theorem 4.1. We can also connect the portfolio's expected return to its variance or standard deviation, as we did for two risky assets in Theorem 4.2. The next theorem summarizes the results.

Theorem 4.5

Consider a buy-and-hold portfolio of a riskfree asset with rate of return r_f and a risky asset with expected rate of return $\hat{\mu}$ and standard deviation $\hat{\sigma}$. Let w denote the portfolio weight of asset 1 so that $1 - w$ is the portfolio weight of asset 2. Then the expectation $\mu(w)$, the variance $\sigma^2(w)$, and the standard deviation $\sigma(w)$ of the portfolio's rate of return are

$$\mu(w) = r_f + w(\hat{\mu} - r_f), \quad (4.19)$$

$$\sigma^2(w) = w^2\hat{\sigma}^2, \quad (4.20)$$

$$\sigma(w) = |w|\hat{\sigma}, \quad (4.21)$$

where $|w|$ denotes the absolute value of w .

Assuming $\hat{\mu} \neq r_f$, the portfolio variance is a quadratic function of its expected return,

$$\sigma^2(w) = \frac{(\mu(w) - r_f)^2}{(\hat{\mu} - r_f)^2}\hat{\sigma}^2, \quad (4.22)$$

and the relation between the portfolio's expected return and standard deviation is given by

$$\mu(w) = \begin{cases} r_f + \frac{\hat{\mu} - r_f}{\hat{\sigma}}\sigma(w), & \text{if } w \geq 0, \\ r_f - \frac{\hat{\mu} - r_f}{\hat{\sigma}}\sigma(w), & \text{if } w \leq 0. \end{cases} \quad (4.23)$$

The Sharpe ratio of the portfolio is equal to plus or minus the Sharpe ratio of the risky asset:

$$\text{SR}(w) = \begin{cases} \widehat{\text{SR}}, & \text{if } w \geq 0, \\ -\widehat{\text{SR}}, & \text{if } w \leq 0. \end{cases} \quad (4.24)$$

The standard deviation is the square root of the variance. When the variance is given by $w^2\hat{\sigma}^2$ as in (4.20), you may at first think that the standard deviation would be $w\hat{\sigma}$, and this is also correct provided that $w \geq 0$. However, to cover cases with $w < 0$, we need to take the absolute value. For example, with $w = -0.5$, we have $\sqrt{w^2} = \sqrt{0.25} = 0.5 = |w|$.

Proof

Based on the computational rules in Eq. (3.50), we get

$$\begin{aligned}\mu(w) &= \text{E}[r(w)] = w\hat{\mu} + (1-w)r_f = r_f + w(\hat{\mu} - r_f), \\ \sigma^2(w) &= \text{Var}[r(w)] = w^2\hat{\sigma}^2, \\ \sigma(w) &= \sqrt{w^2\hat{\sigma}^2} = \sqrt{w^2}\sqrt{\hat{\sigma}^2} = |w|\hat{\sigma}.\end{aligned}$$

If we assume that $\hat{\mu} \neq r_f$, Eq. (4.19) implies $w = (\mu(w) - r_f) / (\hat{\mu} - r_f)$, and substituting this into (4.20), we obtain (4.22).

If $w \geq 0$, the equation for the standard deviation in (4.21) implies that $w = \sigma(w)/\hat{\sigma}$, and substituting that into the equation (4.19) for the expected return, we get the first case in (4.23). If $w \leq 0$, we have $|w| = -w$ so the standard deviation of the combined portfolio is $\sigma(w) = -w\hat{\sigma}$. Hence, we get the second case in (4.23).

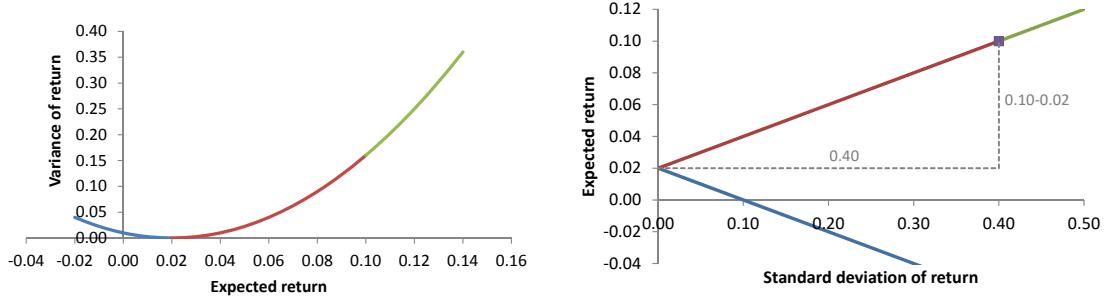
Since the Sharpe ratio of the portfolio is $\text{SR}(w) = (\mu(w) - r_f)/\sigma(w)$, the expression (4.24) follows immediately from (4.23).

Eq. (4.23) shows that, in a diagram with standard deviation along the horizontal axis and expected return along the vertical axis, the set of points $(\sigma(w), \mu(w))$ for $w \geq 0$ forms a straight line with a slope of $(\hat{\mu} - r_f)/\hat{\sigma}$ and with r_f as the intercept on the vertical axis. The slope equals the Sharpe ratio of the risky portfolio, cf. (3.19). The set of points $(\sigma(w), \mu(w))$ for $w \leq 0$ correspond to a straight line with a slope equal to minus the Sharpe ratio. Putting the two lines together, we get a wedge as illustrated below.

Eq. (4.24) shows that the Sharpe ratio of a portfolio of a risky and a riskfree asset equals (plus or minus) the Sharpe ratio of the risky asset in the portfolio. In particular, you cannot improve the Sharpe ratio of a risky asset by combining it with the riskfree asset. Note that if $w > 1$ so that the weight of the riskfree asset is negative, the portfolio corresponds to a leveraged investment in the risky asset, cf. the discussion in Section 2.7. The leveraged position has the same Sharpe ratio as the risky asset itself. This conclusion assumes that the borrowing rate equals the riskfree rate used in the calculation of the Sharpe ratio. If the borrowing rate is larger than that riskfree rate, the Sharpe ratio of the leveraged position will be lower than the Sharpe ratio of the risky asset.

Example 4.6

Suppose the riskfree rate of return is 2% while the risky asset's rate of return has an expectation of 10% and a standard deviation of 40%. The left panel of Figure 4.6 shows the variance of the portfolio as a function of the expected rate of return. The quadratic relation is obvious. The right panel shows the expected rate of return as a function of the standard deviation. Here we see the two straight lines forming a wedge as explained above. The blue part of each curve represents portfolios with a short position in the risky asset, the green part portfolios with a short position in the riskfree asset (a leveraged position in the risky asset), and the red part portfolios where both assets have a weight between 0 and 1. The slope of the red-green line equals the Sharpe ratio of the risky asset, $\frac{\hat{\mu}-r_f}{\hat{\sigma}} = \frac{0.10-0.02}{0.40} = 0.2$.



(a) The variance as a function of the expected return

(b) The expected return as a function of the standard deviation

Figure 4.6: Risk and expected return.

The graphs show the relations between the expectation, variance, and standard deviation of various portfolios given the information in Example 4.6.

4.2 Multi-asset portfolio mathematics

4.2.1 Introduction and summary of main results

Above we considered portfolios of two assets, but let us now generalize the analysis to portfolios of N assets, where $N \geq 1$ is an integer. For each $i = 1, 2, \dots, N$, let r_i denote the rate of return on asset i , and let π_i denote the weight of asset i in the portfolio, i.e. the value of the investment in asset i divided by the total investment in all assets. Since portfolio weights must add up to one, we have $\pi_1 + \pi_2 + \dots + \pi_N = 1$. The rate of return on the portfolio is

$$r_p = \pi_1 r_1 + \pi_2 r_2 + \dots + \pi_N r_N = \sum_{i=1}^N \pi_i r_i, \quad (4.25)$$

cf. Equation (2.22). From Equations (3.43) and (3.44), we see that the expectation and variance of the portfolio return are given by

$$\mathbb{E}[r_p] = \mathbb{E}\left[\sum_{i=1}^N \pi_i r_i\right] = \sum_{i=1}^N \pi_i \mathbb{E}[r_i] \quad (4.26)$$

and

$$\begin{aligned} \text{Var}[r_p] &= \sum_{i=1}^N \pi_i^2 \text{Var}[r_i] + 2 \sum_{i=1}^N \sum_{j=i+1}^N \pi_i \pi_j \text{Cov}[r_i, r_j] \\ &= \sum_{i=1}^N \pi_i^2 \text{Var}[r_i] + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \pi_i \pi_j \text{Cov}[r_i, r_j]. \end{aligned} \quad (4.27)$$

As always, the standard deviation is the square root of the variance:

$$\sigma_p = \text{Std}[r_p] = \sqrt{\sum_{i=1}^N \pi_i^2 \text{Var}[r_i] + 2 \sum_{i=1}^N \sum_{j=i+1}^N \pi_i \pi_j \text{Cov}[r_i, r_j]}. \quad (4.28)$$

If \tilde{r}_p is the rate of return on another portfolio characterized by the weights $\tilde{\pi}_1, \dots, \tilde{\pi}_N$ in the same assets, the covariance between r_p and \tilde{r}_p is

$$\text{Cov}[r_p, \tilde{r}_p] = \sum_{i=1}^N \pi_i \tilde{\pi}_i \text{Var}[r_i] + \sum_{i=1}^N \sum_{\substack{j \neq i \\ j=1}}^N \pi_i \tilde{\pi}_j \text{Cov}[r_i, r_j]. \quad (4.29)$$

Obviously, the expressions—in particular for the variance, the standard deviation, and the covariance—are considerably more complicated for the multi-asset case than the two-asset case. From N objects, you can form $N(N - 1)/2$ different pairs. Hence, with 5 assets there are $5 \times 4/2 = 10$ pairwise covariances. With 10 assets there are $10 \times 9/2 = 45$. And with 100 assets there are $100 \times 99/2 = 4950$. Substituting manually into expressions with so many terms is clearly impractical. We have to develop expressions that are tractable, at least in a spreadsheet or other computational software, even with many assets. For this purpose it is useful to work with vectors and matrices.

In the following subsections, we introduce vectors and matrices and the associated mathematical manipulations, but let us already summarize how we use them in portfolio calculations. We represent a portfolio by a vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ of the portfolio weights in the N assets. Here the symbol $^\top$ indicates the transpose. The expected rates of return on the N assets are gathered in the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)^\top$. The variances and covariances of the assets are captured in the $N \times N$ variance-covariance matrix $\underline{\Sigma}$, where the diagonal elements are variances and the off-diagonal elements are covariances. More precisely, $\underline{\Sigma}_{ii} = \text{Var}[r_i]$ and $\underline{\Sigma}_{ij} = \text{Cov}[r_i, r_j]$. Since $\text{Cov}[r_j, r_i] = \text{Cov}[r_i, r_j]$, we have $\underline{\Sigma}_{ji} = \underline{\Sigma}_{ij}$, so the variance-covariance matrix is symmetric, i.e. $\underline{\Sigma}^\top = \underline{\Sigma}$.

Theorem 4.6

Let $\boldsymbol{\pi}$ denote a portfolio weight vector representing the investments in N assets with expected rate of return vector $\boldsymbol{\mu}$ and variance-covariance matrix $\underline{\Sigma}$. The expectation, variance, and standard deviation of the portfolio's rate of return $r(\boldsymbol{\pi})$ are given by

$$\text{E}[r(\boldsymbol{\pi})] = \boldsymbol{\pi}^\top \boldsymbol{\mu} = \boldsymbol{\pi} \cdot \boldsymbol{\mu}, \quad (4.30)$$

$$\text{Var}[r(\boldsymbol{\pi})] = \boldsymbol{\pi}^\top \underline{\Sigma} \boldsymbol{\pi} = \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}, \quad (4.31)$$

$$\text{Std}[r(\boldsymbol{\pi})] = \sqrt{\boldsymbol{\pi}^\top \underline{\Sigma} \boldsymbol{\pi}} = \sqrt{\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}}. \quad (4.32)$$

If $\tilde{\boldsymbol{\pi}}$ denotes another portfolio weight vector, then the covariance between the rates of return on the two portfolios is

$$\text{Cov}[r(\boldsymbol{\pi}), r(\tilde{\boldsymbol{\pi}})] = \boldsymbol{\pi}^\top \underline{\Sigma} \tilde{\boldsymbol{\pi}} = \boldsymbol{\pi} \cdot \underline{\Sigma} \tilde{\boldsymbol{\pi}}. \quad (4.33)$$

The dot in the formulas in the theorem represent the so-called dot product (or vector product) of two vectors. For example, $\boldsymbol{\pi} \cdot \boldsymbol{\mu}$ denotes the dot product of the two vectors $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$. The dot product of two vectors produces a number. When two vectors or matrices are placed next to each other without any symbol in between, it is to be understood as a matrix product. For example, $\underline{\Sigma} \boldsymbol{\pi}$ is to be understood as the matrix product of the matrix $\underline{\Sigma}$ and the vector $\boldsymbol{\pi}$. The matrix product is generally producing a matrix, but in the case $\underline{\Sigma} \boldsymbol{\pi}$ the result is a vector. Hence, the variance expression $\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}$ is really the dot product of two vectors. Both the dot product and the matrix product are also explained in the

	A	B	C	D	E	F	G	H	I	J	K
1											
2	Stock	Expect	Variance-covariance				Portfolio				
3	Alphabet	0.10	0.090	0.036	0.012		0.5				
4	Boeing	0.12	0.036	0.160	0.008		0.3				
5	Citigroup	0.06	0.012	0.008	0.040		0.2				
6											
7						Expect	0.098	=SUMPRODUCT(G3:G5;B3:B5)			
8						Variance	0.05266	=SUMPRODUCT(G3:G5;MMULT(C3:E5;G3:G5))			
9						Std dev	0.2294777	=SQRT(G8)			
10											

Figure 4.7: Simple portfolio calculations in Excel.

The figure is an excerpt from an Excel sheet and illustrates how the expected return, the variance, and the standard deviation can be calculated from the expected returns and variance-covariance matrix of the assets in the portfolio.

following subsections.

Not only do the expressions in the theorem clearly look simpler than their counterparts above the theorem, they are also very easy to apply in Excel and similar computational software tools. In Excel, a vector is represented by a vertical array of cells, and a matrix by a block of cells. The dot product of two vectors is calculated using the Excel function **SUMPRODUCT**. Multiplying matrices is done in Excel by using the function **MMULT**. In cases where the result you want to calculate is a vector or a matrix, you have to highlight an area of the appropriate dimensions in your Excel sheet, type in the formula, and then simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’ (this depends on your version of Excel). For example, suppose you want to multiply a 4×5 matrix with values written in cells A1:E4 with a 5×2 matrix with values written in cells F1:G5. The result is a 4×2 matrix so highlight an area covering four rows and two columns, say H1:I4, type in **MMULT(A1:E4;F1:G5)**, and simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’.

Figure 4.7 illustrates how a portfolio’s expected return, variance, and standard deviation can be calculated in Excel. The portfolio consists of investments in the three stocks Alphabet, Boeing, and Citigroup. The vector μ of the expected returns of the three stocks is contained in the array B3:B5. The 3×3 variance-covariance matrix is contained in the block C3:E5. The specific portfolio we consider is defined by the vector π of weights listed in the array G3:G5. Then we can calculate the expected return $\pi \cdot \mu$ of the portfolio with the formula **SUMPRODUCT(G3:G5;B3:B5)**. The variance $\pi \cdot \underline{\Sigma} \pi$ of the portfolio is calculated as **SUMPRODUCT(G3:G5;MMULT(C3:E5;G3:G5))**, where the inner part **MMULT(C3:E5;G3:G5)** produces the vector $\underline{\Sigma} \pi$. The standard deviation of the portfolio is, of course, the square root of the variance.

The following subsections provide a more detailed introduction of vectors, matrices, and typical calculations with vectors and matrices. For additional information, the reader is referred to mathematics textbooks such as [Sydsæter, Hammond, Strøm, and Carvajal \(2021\)](#).

4.2.2 Vectors and their use in portfolio computations

Let N be a positive integer. A **vector** of dimension N —or just an N -vector—is an ordered collection of N elements (also called components, coordinates, or entries). An

N -vector \mathbf{x} with elements x_1, \dots, x_N is typically written as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}.$$

Sometimes the fact that a symbol represents a vector is indicated by underlining it instead of writing it in boldface, i.e., \underline{x} instead of \mathbf{x} . In almost all applications in finance, the elements of vectors are real-valued numbers, i.e., $x_i \in \mathbb{R}$ for all $i = 1, 2, \dots, N$.

We can represent a portfolio by a vector of the portfolio weights. Suppose we own a portfolio of three stocks: 20% in the Coca-Cola Company, 30% in Apple, and 50% in Exxon Mobil. We can represent this by the vector

$$\boldsymbol{\pi} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix},$$

where it is then understood that the first number $\pi_1 = 0.2$ is the portfolio weight of Coca-Cola, the second number $\pi_2 = 0.3$ is the portfolio weight of Apple, and the third number $\pi_3 = 0.5$ is the portfolio weight of Exxon Mobil. Likewise, the vector

$$\tilde{\boldsymbol{\pi}} = \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix}$$

represents a portfolio with 60% invested in Coca-Cola, 10% in Apple, and 30% in Exxon Mobil. We need to fix the ordering of the stocks before making any of the calculations of expected return, variance, etc., which we discuss below.

We can also represent the expected rates of return of the assets in the portfolio by a vector $\boldsymbol{\mu}$ which has a dimension equal to the number of assets in the portfolio. Suppose that the expected rates of return of the stocks of Coca-Cola, Apple, and Exxon Mobil are $\mu_1 = 0.12 = 12\%$, $\mu_2 = 0.16 = 16\%$, and $\mu_3 = 0.1 = 10\%$, respectively. In vector-notation, this is written as

$$\boldsymbol{\mu} = \begin{pmatrix} 0.12 \\ 0.16 \\ 0.10 \end{pmatrix}.$$

Given two vectors \mathbf{x} and \mathbf{y} of the same dimension, say N , and a constant scalar $a \in \mathbb{R}$, the vectors $\mathbf{x} + \mathbf{y}$ and $a\mathbf{x}$ are defined as

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{pmatrix}, \quad a\mathbf{x} = a \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_N \end{pmatrix}.$$

For example,

$$\begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} + \begin{pmatrix} 4 \\ -2 \\ -4 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ -5 \end{pmatrix}, \quad 3 \times \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} = \begin{pmatrix} 6 \\ 15 \\ -3 \end{pmatrix}.$$

The *dot product* (or vector product or inner product) of \mathbf{x} and \mathbf{y} is defined as the number

$$\mathbf{x} \cdot \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = x_1y_1 + x_2y_2 + \cdots + x_Ny_N = \sum_{i=1}^N x_iy_i.$$

For example,

$$\begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ -2 \\ -4 \end{pmatrix} = 2 \times 4 + 5 \times (-2) + (-1) \times (-4) = 2.$$

The vectors \mathbf{x} and \mathbf{y} are said to be orthogonal if $\mathbf{x} \cdot \mathbf{y} = 0$. Note that

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}, \quad \mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}, \quad (a\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (a\mathbf{y}) = a(\mathbf{x} \cdot \mathbf{y}). \quad (4.34)$$

The length of a vector \mathbf{x} is defined by

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^N x_i^2}.$$

For example,

$$\left\| \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} \right\| = \sqrt{2^2 + 5^2 + (-1)^2} = \sqrt{30} \approx 5.4772.$$

The Cauchy-Schwartz Inequality says that $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \times \|\mathbf{y}\|$, where the left-hand side is the absolute value of the dot product of \mathbf{x} and \mathbf{y} .

Such vector manipulations are easy in Excel. Typically, a vector in Excel is written in a column array of cells. For example, suppose that \mathbf{x} and \mathbf{y} are four-dimensional vectors and you have vector \mathbf{x} written in cells A1:A4 and vector \mathbf{y} written in cells B1:B4. The sum of the two vectors is also a four-dimensional vector. First, highlight a column area of four cells, say cells C1:C4. Now simply type =A1:A4+B1:B4 and then simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’ (always do this when you type in a formula where the result is a vector or matrix). The result now appears in cells C1:C4. Suppose you want to multiply the \mathbf{x} vector with a number and that number is written in cell D1. Since the result is a vector of the same dimension as \mathbf{x} , first highlight a column area of four cells, say cells E1:E4, type in =D1*A1:A4, and simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’. The dot product of \mathbf{x} and \mathbf{y} is just a number. If you want that result in cell F1, go to that cell and type in =SUMPRODUCT(A1:A4;B1:B4) and press ‘Enter’.

Let $\mathbf{1}$ denote a vector with all elements equal to one (the dimension of $\mathbf{1}$ will be clear from the context). If we have an N -vector \mathbf{x} , then

$$\mathbf{x} \cdot \mathbf{1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = x_1 + x_2 + \cdots + x_N = \sum_{i=1}^N x_i,$$

which is simply the sum of the elements in the vector \mathbf{x} . We generally require portfolio weights to sum up to one. If $\boldsymbol{\pi}$ denotes the vector of portfolio weights, we therefore

require that the constraint $\boldsymbol{\pi} \cdot \mathbf{1} = 1$ holds. Note the difference between the vector $\mathbf{1}$ and the number 1.

The vector computations introduced above are useful when handling portfolios. If we let \mathbf{r} denote a vector of the rates of return on the assets we invest in, then the rate of return on the portfolio $\boldsymbol{\pi}$ is

$$r(\boldsymbol{\pi}) = \pi_1 r_1 + \pi_2 r_2 + \cdots + \pi_N r_N = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} \cdot \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix} = \boldsymbol{\pi} \cdot \mathbf{r}. \quad (4.35)$$

The portfolio return is simply the dot product of the portfolio vector and the return vector. In the three-asset example above, suppose the returns turn out to be 20% on Coca-Cola, -10% on Apple, and 8% on Exxon Mobil. Then the return on the portfolio $\boldsymbol{\pi}$ is

$$\begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.2 \\ -0.1 \\ 0.08 \end{pmatrix} = 0.2 \times 0.2 + 0.3 \times (-0.1) + 0.5 \times 0.08 = 0.05 = 5\%.$$

The same goes for the expected rate of return on the portfolio:

$$\begin{aligned} E[r(\boldsymbol{\pi})] &= E[\pi_1 r_1 + \pi_2 r_2 + \cdots + \pi_N r_N] = \pi_1 E[r_1] + \pi_2 E[r_2] + \cdots + \pi_N E[r_N] \\ &= \pi_1 \mu_1 + \pi_2 \mu_2 + \cdots + \pi_N \mu_N = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix} = \boldsymbol{\pi} \cdot \boldsymbol{\mu}, \end{aligned}$$

which confirms Eq. (4.30). Here μ_i is the expected rate of return on asset i , and $\boldsymbol{\mu}$ is the N -vector stacking up all these expected rates of return. In short, we have $E[r(\boldsymbol{\pi})] = E[\boldsymbol{\pi} \cdot \mathbf{r}] = \boldsymbol{\pi} \cdot E[\mathbf{r}]$. Here $E[\mathbf{r}]$ is the expectation of the return vector, which just means the vector of expected returns. In the three-asset example above, we can compute the expected rate of return on the portfolio $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} \cdot \boldsymbol{\mu} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.12 \\ 0.16 \\ 0.10 \end{pmatrix} = 0.2 \times 0.12 + 0.3 \times 0.16 + 0.5 \times 0.1 = 0.122 = 12.2\%.$$

4.2.3 Matrices

To compute the variance of the rate of return on a portfolio it is useful to introduce matrices. Let M and N be positive integers. A **matrix** of dimension (M, N) —or just an $M \times N$ matrix—is a collection of $M \times N$ elements (or entries) ordered in a rectangular array with M rows and N columns. An $M \times N$ matrix $\underline{\underline{A}}$ is typically written as

$$\underline{\underline{A}} = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{pmatrix}$$

or more shortly as $\underline{\underline{A}} = [A_{ij}]$, where it is then understood that A_{ij} is the element in the position where row i crosses column j , where $i = 1, \dots, M$ and $j = 1, \dots, N$. Sometimes A_{ij} is called element or entry (i, j) of $\underline{\underline{A}}$. In almost all applications in finance, each element is a real-valued number, i.e., $A_{ij} \in \mathbb{R}$ for all $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, N$. A matrix with equally many rows and columns is called a square matrix. Note that an M -dimensional (column) vector \mathbf{x} with elements x_1, \dots, x_M can be seen as a matrix of dimension $M \times 1$. A matrix of dimension $1 \times N$ for some integer N consists of a single row and is often referred to as a *row vector*. We first present some general matrix concepts and computational rules, then we turn to applications related to portfolios.

Adding and scaling matrices. Given two matrices $\underline{\underline{A}} = [A_{ij}]$ and $\underline{\underline{B}} = [B_{ij}]$ of the same dimension (M, N) and a scalar $a \in \mathbb{R}$, the matrices $\underline{\underline{A}} + \underline{\underline{B}}$ and $a\underline{\underline{A}}$ are defined as

$$\begin{aligned}\underline{\underline{A}} + \underline{\underline{B}} &= \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & \dots & A_{1N} + B_{1N} \\ A_{21} + B_{21} & A_{22} + B_{22} & \dots & A_{2N} + B_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} + B_{M1} & A_{M2} + B_{M2} & \dots & A_{MN} + B_{MN} \end{pmatrix}, \\ a\underline{\underline{A}} &= \begin{pmatrix} aA_{11} & aA_{12} & \dots & aA_{1N} \\ aA_{21} & aA_{22} & \dots & aA_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ aA_{M1} & aA_{M2} & \dots & aA_{MN} \end{pmatrix}.\end{aligned}$$

For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} + \begin{pmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} = \begin{pmatrix} 8 & 10 & 12 \\ 14 & 16 & 18 \end{pmatrix}$$

and

$$5 \times \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 5 & 10 & 15 \\ 20 & 25 & 30 \end{pmatrix}.$$

Note that you cannot add matrices of different dimensions.

Such manipulations can be done in Excel as explained for vectors. Suppose you want to add the 3×4 matrix written in cells A1:D3 and the 3×4 matrix written in cells E1:H3. Since the results is also a 3×4 matrix, highlight an area of three rows and four columns, say I1:L3, type in =A1:D3+E1:H3 and simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’.

Transposing matrices. The transpose of an $M \times N$ matrix $\underline{\underline{A}} = [A_{ij}]$ is the $N \times M$ matrix $\underline{\underline{A}}^\top = [A_{ji}]$ with columns and rows interchanged:

$$\underline{\underline{A}} = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{pmatrix} \Rightarrow \underline{\underline{A}}^\top = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{M1} \\ A_{12} & A_{22} & \dots & A_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1N} & A_{2N} & \dots & A_{MN} \end{pmatrix}.$$

A simple example:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}^\top = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}.$$

An N -dimensional column vector \mathbf{x} with elements x_1, \dots, x_N can then be written as $\mathbf{x} = (x_1, \dots, x_N)^\top$. A matrix $\underline{\underline{A}}$ is called *symmetric* if $\underline{\underline{A}}^\top = \underline{\underline{A}}$, which is only possible if $\underline{\underline{A}}$ is a square matrix. Note that for any matrix $\underline{\underline{A}}$, we have $(\underline{\underline{A}}^\top)^\top = \underline{\underline{A}}$, i.e., if we interchange columns and rows twice, we are back to where we started.

Transposing matrices are easy in Excel using the function **TRANSPOSE**. If you have a 5×3 matrix in cells A1:C5, first highlight an area covering 3 rows and 5 columns, say D1:H3, type in =**TRANSPOSE(A1:C5)** and simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’.

Multiplying matrices. We can multiply matrices together if they have appropriate dimensions. Given an $M \times N$ matrix $\underline{\underline{A}} = [A_{ij}]$ and an $N \times K$ matrix $\underline{\underline{B}} = [B_{jk}]$, the matrix product $\underline{\underline{A}}\underline{\underline{B}}$ is the $M \times K$ matrix with (i, k) ’th entry given by

$$(\underline{\underline{A}}\underline{\underline{B}})_{ik} = \sum_{j=1}^N A_{ij} B_{jk},$$

which can be seen as the dot product of the i ’th row of $\underline{\underline{A}}$ and the k ’th column of $\underline{\underline{B}}$. Note that the multiplication is only possible if the number of columns in the first matrix equals the number of rows in the second matrix.

Here is an example:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} = \begin{pmatrix} 66 & 72 & 78 \\ 156 & 171 & 186 \end{pmatrix}$$

Multiplying a 2×3 matrix and a 3×3 matrix is possible and results in a 2×3 matrix. Here the entry in place $(1, 1)$ is computed as the dot product of row 1 in the left matrix and column 1 in the right matrix:

$$1 \times 7 + 2 \times 10 + 3 \times 13 = 66.$$

The entry in place $(1, 2)$ is the dot product of row 1 in the left matrix and column 2 in the right matrix:

$$1 \times 8 + 2 \times 11 + 3 \times 14 = 72.$$

The entry in place $(2, 3)$ is the dot product of row 2 in the left matrix and column 3 in the right matrix:

$$4 \times 9 + 5 \times 12 + 6 \times 15 = 186.$$

The other entries are computed similarly. Please check that you can compute these numbers correctly.

Note that, in general, $\underline{\underline{A}}\underline{\underline{B}} \neq \underline{\underline{B}}\underline{\underline{A}}$. In fact, in some cases where you can calculate $\underline{\underline{AB}}$, the expression $\underline{\underline{BA}}$ does not even make sense, for example when $\underline{\underline{A}}$ is a 3×5 matrix and $\underline{\underline{B}}$ is a 5×4 matrix. And even if both $\underline{\underline{AB}}$ and $\underline{\underline{BA}}$ make sense, they are generally different.

For three matrices $\underline{\underline{A}}$, $\underline{\underline{B}}$, and $\underline{\underline{C}}$ of appropriate dimensions, we have

$$\begin{aligned} (\underline{\underline{A}} + \underline{\underline{B}})\underline{\underline{C}} &= \underline{\underline{AC}} + \underline{\underline{BC}}, \\ \underline{\underline{A}}(\underline{\underline{B}} + \underline{\underline{C}}) &= \underline{\underline{AB}} + \underline{\underline{AC}}, \\ (\underline{\underline{A}}\underline{\underline{B}})\underline{\underline{C}} &= \underline{\underline{A}}(\underline{\underline{BC}}), \\ (\underline{\underline{AB}})^\top &= \underline{\underline{B}}^\top \underline{\underline{A}}^\top. \end{aligned} \tag{4.36}$$

By iterating the latter expression, we also see that $(\underline{ABC})^\top = \underline{C}^\top \underline{B}^\top \underline{A}^\top$.

The matrix product generalizes the dot product for vectors. To see this, look at the matrix product when $M = K = 1$ so that the matrix \underline{A} is really a row vector and the matrix \underline{B} is a column vector. It follows that for two vectors $\mathbf{x} = (x_1, \dots, x_N)^\top$ and $\mathbf{y} = (y_1, \dots, y_N)^\top$, we have

$$\mathbf{x}^\top \mathbf{y} = \mathbf{x} \cdot \mathbf{y}. \quad (4.37)$$

This implies the following result which is sometimes useful:

$$\underline{\underline{A}} \text{ symmetric} \Rightarrow \mathbf{x} \cdot \underline{\underline{A}} \mathbf{y} = \mathbf{y} \cdot \underline{\underline{A}} \mathbf{x} \quad (4.38)$$

since $\mathbf{x} \cdot \underline{\underline{A}} \mathbf{y} = \mathbf{x}^\top \underline{\underline{A}} \mathbf{y} = (\mathbf{x}^\top \underline{\underline{A}} \mathbf{y})^\top = \mathbf{y}^\top \underline{\underline{A}}^\top \mathbf{x} = \mathbf{y} \cdot \underline{\underline{A}}^\top \mathbf{x} = \mathbf{y} \cdot \underline{\underline{A}} \mathbf{x}$, where the second equality is correct since $\mathbf{x}^\top \underline{\underline{A}} \mathbf{y}$ is just a scalar and the transpose of a scalar gives the scalar itself as there are no rows and columns to interchange.

Let $\underline{\underline{1}}$ denote a square matrix with one in the main diagonal and zeroes in all other entries, i.e.,

$$\underline{\underline{1}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

where the dimension of the matrix will be clear from the context. This matrix is called the *identity matrix*. Then we have $\underline{\underline{1}} \underline{\underline{A}} = \underline{\underline{A}}$ and $\underline{\underline{A}} \underline{\underline{1}} = \underline{\underline{A}}$. For any vector \mathbf{x} , we have $\underline{\underline{1}} \mathbf{x} = \mathbf{x}$.

Multiplying matrices is done in Excel by using the function MMULT. Suppose you want to multiply a 4×5 matrix with values written in cells A1:E4 with a 5×2 matrix with values written in cells F1:G5. The result is a 4×2 matrix so highlight an area covering four rows and two columns, say H1:I4, type in MMULT(A1:E4;F1:G5) and simultaneously press ‘Ctrl’, ‘Shift’, and ‘Enter’.

Inverting matrices. An $N \times N$ matrix $\underline{\underline{A}}$ is said to be non-singular or invertible if there exists a matrix $\underline{\underline{A}}^{-1}$ so that $\underline{\underline{A}} \underline{\underline{A}}^{-1} = \underline{\underline{1}}$, where $\underline{\underline{1}}$ is the $N \times N$ identity matrix, and then $\underline{\underline{A}}^{-1}$ is called the inverse of $\underline{\underline{A}}$. If $\underline{\underline{A}}^{-1}$ exists, we also have $\underline{\underline{A}}^{-1} \underline{\underline{A}} = \underline{\underline{1}}$. Conversely, if $\underline{\underline{A}}^{-1}$ satisfies $\underline{\underline{A}}^{-1} \underline{\underline{A}} = \underline{\underline{1}}$, it will also satisfy $\underline{\underline{A}} \underline{\underline{A}}^{-1} = \underline{\underline{1}}$.

We have a simple explicit formula for the inverse of a 2×2 matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

which obviously requires that $ad - bc \neq 0$. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} -2 & 1 \\ 1.5 & -0.5 \end{pmatrix}$$

because

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ 1.5 & -0.5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

If $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are non-singular matrices of appropriate dimensions,

$$(\underline{\underline{A}} \underline{\underline{B}})^{-1} = \underline{\underline{B}}^{-1} \underline{\underline{A}}^{-1}. \quad (4.39)$$

Furthermore, $(\underline{\underline{A}}^{-1})^{-1} = \underline{\underline{A}}$. If $\underline{\underline{A}}$ is non-singular, then the transposed matrix $\underline{\underline{A}}^\top$ is also non-singular and $(\underline{\underline{A}}^\top)^{-1} = (\underline{\underline{A}}^{-1})^\top$. If $\underline{\underline{A}}$ is symmetric and non-singular, then the inverse $\underline{\underline{A}}^{-1}$ is also symmetric. In Excel, matrices can be inverted by using the function MINVERSE. This is applied in the same way as the other matrix-related functions explained above.

Solving equation system using matrix manipulations. A system of N linear equations

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + \dots + A_{1N}x_N &= b_1, \\ A_{21}x_1 + A_{22}x_2 + \dots + A_{2N}x_N &= b_2, \\ &\vdots && \vdots \\ A_{N1}x_1 + A_{N2}x_2 + \dots + A_{NN}x_N &= b_N \end{aligned}$$

in N unknowns x_1, \dots, x_N can be written in matrix-vector form as

$$\underline{\underline{A}}\underline{\underline{x}} = \underline{\underline{b}},$$

where $\underline{\underline{A}} = [A_{ij}]$ is an $N \times N$ matrix, $\underline{\underline{x}} = (x_1, \dots, x_N)^\top \in \mathbb{R}^N$, and $\underline{\underline{b}} = (b_1, \dots, b_N)^\top \in \mathbb{R}^N$. If $\underline{\underline{A}}$ is non-singular, then the system has the unique solution $\underline{\underline{x}} = \underline{\underline{A}}^{-1}\underline{\underline{b}}$.

Derivatives of vector and matrix expressions. The derivative of a real-valued function $f(\underline{\underline{x}})$ of a vector $\underline{\underline{x}} = (x_1, \dots, x_n)$ is the n -vector of partial derivatives:

$$\frac{\partial f}{\partial \underline{\underline{x}}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)^\top.$$

It can be shown that when $\underline{\underline{a}} = (a_1, \dots, a_n)^\top$ is a vector of constants, then

$$\frac{\partial(\underline{\underline{a}} \cdot \underline{\underline{x}})}{\partial \underline{\underline{x}}} = \frac{\partial(\underline{\underline{x}} \cdot \underline{\underline{a}})}{\partial \underline{\underline{x}}} = \underline{\underline{a}}. \quad (4.40)$$

For example, with $\underline{\underline{a}} = (a_1, a_2)^\top$ and $\underline{\underline{x}} = (x_1, x_2)^\top$, we have $\underline{\underline{a}} \cdot \underline{\underline{x}} = a_1x_1 + a_2x_2$, so

$$\frac{\partial(\underline{\underline{a}} \cdot \underline{\underline{x}})}{\partial x_1} = a_1, \quad \frac{\partial(\underline{\underline{a}} \cdot \underline{\underline{x}})}{\partial x_2} = a_2 \quad \Rightarrow \quad \frac{\partial(\underline{\underline{a}} \cdot \underline{\underline{x}})}{\partial \underline{\underline{x}}} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \underline{\underline{a}}.$$

It can also be shown that for a vector $\underline{\underline{x}}$ of dimension n and an $n \times n$ matrix $\underline{\underline{A}}$,

$$\frac{\partial(\underline{\underline{x}} \cdot \underline{\underline{A}}\underline{\underline{x}})}{\partial \underline{\underline{x}}} = (\underline{\underline{A}} + \underline{\underline{A}}^\top)\underline{\underline{x}}.$$

For example, with dimension $n = 2$, we have

$$\underline{\underline{x}} \cdot \underline{\underline{A}}\underline{\underline{x}} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + A_{22}x_2^2$$

so

$$\frac{\partial(\underline{\underline{x}} \cdot \underline{\underline{A}}\underline{\underline{x}})}{\partial x_1} = 2A_{11}x_1 + (A_{12} + A_{21})x_2, \quad \frac{\partial(\underline{\underline{x}} \cdot \underline{\underline{A}}\underline{\underline{x}})}{\partial x_2} = (A_{12} + A_{21})x_1 + 2A_{22}x_2,$$

which means

$$\begin{aligned}\frac{\partial(\mathbf{x} \cdot \underline{\underline{A}}\mathbf{x})}{\partial \mathbf{x}} &= \begin{pmatrix} 2A_{11}x_1 + (A_{12} + A_{21})x_2 \\ (A_{12} + A_{21})x_1 + 2A_{22}x_2 \end{pmatrix} = \begin{pmatrix} 2A_{11} & A_{12} + A_{21} \\ A_{12} + A_{21} & 2A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \left[\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix} \right] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = [\underline{\underline{A}} + \underline{\underline{A}}^\top] \mathbf{x}.\end{aligned}$$

In particular, if $\underline{\underline{A}}$ is symmetric so that $\underline{\underline{A}}^\top = \underline{\underline{A}}$, we get

$$\frac{\partial(\mathbf{x} \cdot \underline{\underline{A}}\mathbf{x})}{\partial \mathbf{x}} = 2\underline{\underline{A}}\mathbf{x}. \quad (4.41)$$

We shall use these derivatives when determining the optimal portfolio in the mean-variance framework in Chapter 7.

4.2.4 Portfolio variance and standard deviation

Suppose we consider investing in a portfolio of N assets. The rates of return of these assets is written as a vector $\mathbf{r} = (r_1, r_2, \dots, r_N)^\top$. The variance-covariance matrix $\underline{\Sigma} = \text{Var}[\mathbf{r}]$ of \mathbf{r} is defined as the $N \times N$ matrix

$$\underline{\Sigma} = \begin{pmatrix} \text{Var}[r_1] & \text{Cov}[r_1, r_2] & \text{Cov}[r_1, r_3] & \dots & \text{Cov}[r_1, r_N] \\ \text{Cov}[r_2, r_1] & \text{Var}[r_2] & \text{Cov}[r_2, r_3] & \dots & \text{Cov}[r_2, r_N] \\ \text{Cov}[r_3, r_1] & \text{Cov}[r_3, r_2] & \text{Var}[r_3] & \dots & \text{Cov}[r_3, r_N] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[r_N, r_1] & \text{Cov}[r_N, r_2] & \text{Cov}[r_N, r_3] & \dots & \text{Var}[r_N] \end{pmatrix}.$$

Since $\text{Var}[r_i] = \text{Cov}[r_i, r_i]$, we can write any element of the matrix as $\Sigma_{ij} = \text{Cov}[r_i, r_j]$ even for $j = i$. And since $\text{Cov}[r_i, r_j] = \text{Cov}[r_j, r_i]$, the variance-covariance matrix is symmetric in the sense that $\underline{\Sigma}^\top = \underline{\Sigma}$.

Let us now verify Eq. (4.32), i.e. that the variance of a portfolio return is given by $\boldsymbol{\pi}^\top \underline{\Sigma} \boldsymbol{\pi} = \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}$. Note that the two expressions are identical since both $\boldsymbol{\pi}$ and $\underline{\Sigma} \boldsymbol{\pi}$ are column vectors (of dimension N) and then the rule (4.37) applies. First compute

$$\begin{aligned}\underline{\Sigma} \boldsymbol{\pi} &= \begin{pmatrix} \text{Var}[r_1] & \text{Cov}[r_1, r_2] & \text{Cov}[r_1, r_3] & \dots & \text{Cov}[r_1, r_N] \\ \text{Cov}[r_2, r_1] & \text{Var}[r_2] & \text{Cov}[r_2, r_3] & \dots & \text{Cov}[r_2, r_N] \\ \text{Cov}[r_3, r_1] & \text{Cov}[r_3, r_2] & \text{Var}[r_3] & \dots & \text{Cov}[r_3, r_N] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[r_N, r_1] & \text{Cov}[r_N, r_2] & \text{Cov}[r_N, r_3] & \dots & \text{Var}[r_N] \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \vdots \\ \pi_N \end{pmatrix} \\ &= \begin{pmatrix} \pi_1 \text{Var}[r_1] + \pi_2 \text{Cov}[r_1, r_2] + \pi_3 \text{Cov}[r_1, r_3] + \dots + \pi_N \text{Cov}[r_1, r_N] \\ \pi_1 \text{Cov}[r_2, r_1] + \pi_2 \text{Var}[r_2] + \pi_3 \text{Cov}[r_2, r_3] + \dots + \pi_N \text{Cov}[r_2, r_N] \\ \pi_1 \text{Cov}[r_3, r_1] + \pi_2 \text{Cov}[r_3, r_2] + \pi_3 \text{Var}[r_3] + \dots + \pi_N \text{Cov}[r_3, r_N] \\ \vdots \\ \pi_1 \text{Cov}[r_N, r_1] + \pi_2 \text{Cov}[r_N, r_2] + \pi_3 \text{Cov}[r_N, r_3] + \dots + \pi_N \text{Var}[r_N] \end{pmatrix}.\end{aligned}$$

Since $\text{Var}[r_1] = \text{Cov}[r_1, r_1]$, we can rewrite the first element of the vector on the right-hand

side as

$$\begin{aligned}\pi_1 \text{Cov}[r_1, r_1] + \pi_2 \text{Cov}[r_1, r_2] + \pi_3 \text{Cov}[r_1, r_3] + \cdots + \pi_N \text{Cov}[r_1, r_N] \\ = \text{Cov}[r_1, \pi_1 r_1 + \pi_2 r_2 + \pi_3 r_3 + \cdots + \pi_N r_N] \\ = \text{Cov}[r_1, r(\boldsymbol{\pi})],\end{aligned}$$

where we have used (3.49) and (4.35). Similarly, the second element of the vector is $\text{Cov}[r_2, r(\boldsymbol{\pi})]$, and so forth. In sum,

$$\underline{\Sigma} \boldsymbol{\pi} = \begin{pmatrix} \text{Cov}[r_1, r(\boldsymbol{\pi})] \\ \text{Cov}[r_2, r(\boldsymbol{\pi})] \\ \vdots \\ \text{Cov}[r_N, r(\boldsymbol{\pi})] \end{pmatrix} \quad (4.42)$$

By multiplying this vector by $\boldsymbol{\pi}$, we obtain

$$\begin{aligned}\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} &= \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} \cdot \begin{pmatrix} \text{Cov}[r_1, r(\boldsymbol{\pi})] \\ \text{Cov}[r_2, r(\boldsymbol{\pi})] \\ \vdots \\ \text{Cov}[r_N, r(\boldsymbol{\pi})] \end{pmatrix} \\ &= \pi_1 \text{Cov}[r_1, r(\boldsymbol{\pi})] + \pi_2 \text{Cov}[r_2, r(\boldsymbol{\pi})] + \cdots + \pi_N \text{Cov}[r_N, r(\boldsymbol{\pi})] \\ &= \text{Cov}[\pi_1 r_1 + \pi_2 r_2 + \cdots + \pi_N r_N, r(\boldsymbol{\pi})] \\ &= \text{Cov}[r(\boldsymbol{\pi}), r(\boldsymbol{\pi})] \\ &= \text{Var}[r(\boldsymbol{\pi})],\end{aligned}$$

which confirms (4.32).

Let us return to the example with a portfolio of the stocks of Coca-Cola, Apple, and Exxon Mobil. Suppose the variance-covariance matrix of the three stocks (in the order just listed) is

$$\underline{\Sigma} = \begin{pmatrix} 0.0625 & 0.03 & 0.018 \\ 0.03 & 0.16 & 0.0144 \\ 0.018 & 0.0144 & 0.1296 \end{pmatrix}.$$

In particular, this means that the standard deviations of the three stocks are $\sqrt{0.0625} = 0.25$, $\sqrt{0.16} = 0.4$, and $\sqrt{0.1296} = 0.36$, respectively. The covariance between Coca-Cola and Apple is 0.03 corresponding to a correlation of $0.03/(0.25 \times 0.4) = 0.3$. Similarly, the correlation between Coca-Cola and Exxon Mobil is 0.2, and the correlation between Apple and Exxon Mobil is 0.1. Now the variance of the rate of return on the portfolio $\boldsymbol{\pi} = (0.2, 0.3, 0.5)^\top$ is computed as

$$\begin{aligned}\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} &= \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.0625 & 0.03 & 0.018 \\ 0.03 & 0.16 & 0.0144 \\ 0.018 & 0.0144 & 0.1296 \end{pmatrix} \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.03050 \\ 0.06120 \\ 0.07272 \end{pmatrix} \\ &= 0.2 \times 0.03050 + 0.3 \times 0.06120 + 0.5 \times 0.07272 = 0.06082\end{aligned}$$

and the standard deviation is $\sqrt{0.06082} \approx 0.24662$. Note that the standard deviation of the portfolio return is smaller than the standard deviation of any of the three assets in the portfolio. This is again the diversification effect at play: you can reduce the risk by spreading your investments over several assets.

The computation of the portfolio variance in Eq. (4.31) is easy in Excel. You need to have the variance-covariance matrix in a square block of cells. Then you can compute the portfolio variance in one step by using the Excel functions `MMULT` and `TRANSPOSE`. For example, if you have ten assets and the variance-covariance matrix is in the square block A1:J10 and the portfolio weight vector is in the column K1:K10, you can compute the variance of the portfolio in a single cell by typing

```
=MMULT(TRANSPOSE(K1:K10);MMULT(A1:J10;K1:K10))
```

Here, `TRANSPOSE(K1:K10)` corresponds to $\boldsymbol{\pi}^\top$ and `MMULT(A1:J10;K1:K10)` corresponds to $\underline{\Sigma}\boldsymbol{\pi}$, and the outer `MMULT` then delivers $\boldsymbol{\pi}^\top \underline{\Sigma}\boldsymbol{\pi}$, which is the portfolio variance. Alternatively, if you write the portfolio variance as $\boldsymbol{\pi} \cdot \underline{\Sigma}\boldsymbol{\pi}$, it is clear that you can compute this in Excel as

```
SUMPRODUCT(K1:K10;MMULT(A1:J10;K1:K10))
```

because the `SUMPRODUCT` is the way to compute the dot product.

Sometimes you may prefer to have the portfolios written in a row block in your Excel sheet. Then this corresponds to $\boldsymbol{\pi}^\top$ and you need to transpose the row to get $\boldsymbol{\pi}$. If in the ten-asset example your portfolio weights are in the row A11:J11, the portfolio variance is obtained by

```
MMULT(A11:J11,MMULT(A1:J10;TRANSPOSE(A11:J11)))
```

Alternatively, you can use

```
SUMPRODUCT(TRANSPOSE(A11:J11);MMULT(A1:J10;TRANSPOSE(A11:J11)))
```

or even the simpler

```
SUMPRODUCT(MMULT(A11:J11;A1:J10);A11:J11)
```

that produces the dot product between the two row vectors $\boldsymbol{\pi}^\top \underline{\Sigma}$ and $\boldsymbol{\pi}^\top$ which is another way of expressing $\boldsymbol{\pi}^\top \underline{\Sigma}\boldsymbol{\pi}$.

Section 4.1 showed that two-asset portfolios involve a clear tradeoff between risk and expected return. With two given assets, any fixed target expected return is obtained by a unique portfolio and thus a certain variance and a certain standard deviation. This is the key focus of the so-called mean-variance theory, which is a central model in financial economics and is explored in detail in Chapter 7. With many assets, the link between mean and variance (or standard deviation) is less clear. Any fixed target expected return can be obtained by many different portfolios of three or more assets. As investors generally shy away from risk if they are not compensated by higher expected returns, you would expect investors to search for the portfolio with lowest variance among all the portfolios providing the targeted expected return. This boils down to a constrained minimization problem with many choice variables, a type of problem that can only be solved with some mathematical efforts as we shall see in Chapter 7.

4.2.5 The covariance of the returns of two portfolios

Eq. (4.33) claims that the covariance between the returns on two portfolios $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\pi}}$ equals $\boldsymbol{\pi} \cdot \underline{\Sigma} \tilde{\boldsymbol{\pi}}$. Let us verify this. From Eq. (4.42), we have

$$\underline{\Sigma} \tilde{\boldsymbol{\pi}} = \begin{pmatrix} \text{Cov}[r_1, r(\tilde{\boldsymbol{\pi}})] \\ \text{Cov}[r_2, r(\tilde{\boldsymbol{\pi}})] \\ \vdots \\ \text{Cov}[r_N, r(\tilde{\boldsymbol{\pi}})] \end{pmatrix}.$$

By multiplying this vector by $\boldsymbol{\pi}$, we obtain

$$\begin{aligned} \boldsymbol{\pi} \cdot \underline{\Sigma} \tilde{\boldsymbol{\pi}} &= \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{pmatrix} \cdot \begin{pmatrix} \text{Cov}[r_1, r(\tilde{\boldsymbol{\pi}})] \\ \text{Cov}[r_2, r(\tilde{\boldsymbol{\pi}})] \\ \vdots \\ \text{Cov}[r_N, r(\tilde{\boldsymbol{\pi}})] \end{pmatrix} \\ &= \pi_1 \text{Cov}[r_1, r(\tilde{\boldsymbol{\pi}})] + \pi_2 \text{Cov}[r_2, r(\tilde{\boldsymbol{\pi}})] + \dots + \pi_N \text{Cov}[r_N, r(\tilde{\boldsymbol{\pi}})] \\ &= \text{Cov}[\pi_1 r_1 + \pi_2 r_2 + \dots + \pi_N r_N, r(\tilde{\boldsymbol{\pi}})] \\ &= \text{Cov}[r(\boldsymbol{\pi}), r(\tilde{\boldsymbol{\pi}})], \end{aligned}$$

which confirms (4.33).

In our three-asset example from earlier the covariance between the return on the portfolio $\boldsymbol{\pi} = (0.2, 0.3, 0.5)^\top$ and the return on the portfolio $\tilde{\boldsymbol{\pi}} = (0.6, 0.1, 0.3)^\top$ is therefore

$$\boldsymbol{\pi} \cdot \underline{\Sigma} \tilde{\boldsymbol{\pi}} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.0625 & 0.03 & 0.018 \\ 0.03 & 0.16 & 0.0144 \\ 0.018 & 0.0144 & 0.1296 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.1 \\ 0.3 \end{pmatrix} = 0.046236.$$

Since the standard deviations of the return on the portfolios are 0.24662 and 0.21612, respectively, the correlation between the two portfolio returns is $0.046236 / (0.24662 \times 0.21612) \approx 0.86748$.

4.3 Higher-order moments of portfolio returns

As the above sections have shown, the expected portfolio return follows quite easily from the portfolio weights and the expected returns on the assets in the portfolio, whereas the portfolio variance is a more complicated expression involving the portfolio weights, the asset variances, as well as the pairwise asset covariances. Proceeding to higher-order moments leads to even more complicated relations. The skewness of the portfolio return involves the so-called coskewnesses. For every three assets i , j , and k , the **coskewness** S_{ijk} is defined as

$$S_{ijk} = \frac{\mathbb{E}[(r_i - \mu_i)(r_j - \mu_j)(r_k - \mu_k)]}{\sigma_i \sigma_j \sigma_k},$$

where $\mu_i = \mathbb{E}[r_i]$, $\sigma_i = \text{Std}[r_i]$ and similarly for j and k . Note that with $i = j = k$, the coskewness is identical to the skewness, cf. the definition in Eq. (3.17). This is similar to the relation between the variance and the covariance. Also note that i , j , and k can be interchanged so that $S_{ijk} = S_{ikj} = S_{jik} = S_{jki} = S_{kij} = S_{kji}$.

Intuitively, the coskewness S_{ijk} is a measure of how asset i varies together with the

covariance between assets j and k . With $k = j$, the coskewness is

$$S_{ijj} = \frac{\mathbb{E}[(r_i - \mu_i)(r_j - \mu_j)^2]}{\sigma_i \sigma_j^2},$$

which measures how the return on asset i covaries with extreme (positive or negative) returns of asset j .

For a two-asset portfolio with rate of return $r_p = wr_1 + (1-w)r_2$, the skewness can be written as

$$S_p = w^3 S_1^3 \frac{\sigma_1^3}{\sigma_p^3} + 3w^2(1-w) S_{112} \frac{\sigma_1^2 \sigma_2}{\sigma_p^3} + 3w(1-w)^2 S_{122} \frac{\sigma_1 \sigma_2^2}{\sigma_p^3} + (1-w)^3 S_2^3 \frac{\sigma_2^3}{\sigma_p^3}, \quad (4.43)$$

where S_1 and S_2 are the skewnesses of the two asset returns, σ_1 and σ_2 are their standard deviations, and σ_p is the portfolio standard deviation given by (4.4). In particular, we can see that even if the two asset returns are non-skewed, i.e. $S_1 = S_2 = 0$, the portfolio return can be positively or negatively skewed depending on the signs and magnitudes of the coskewnesses S_{112} and S_{122} . In general, it seems impossible to say whether the portfolio skewness S_p is larger or smaller than the weighted average of the asset skewnesses, $wS_1 + (1-w)S_2$.

More generally, for a multi-asset portfolio return $r_p = \sum_i \pi_i r_i$, the skewness is

$$S_p = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \pi_i \pi_j \pi_k S_{ijk} \frac{\sigma_i \sigma_j \sigma_k}{\sigma_p^3}, \quad (4.44)$$

which again involves the asset-specific skewnesses $S_i = S_{iii}$ and all the $n^3 - n = n(n^2 - 1)$ coskewnesses. When the number of assets n increases, the number of terms in the triple-sum increases rapidly and the expression quickly gets intractable.

Similarly, the kurtosis of the portfolio return $r_p = \sum_i \pi_i r_i$ is

$$K_p = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \pi_i \pi_j \pi_k \pi_\ell K_{ijkl} \frac{\sigma_i \sigma_j \sigma_k \sigma_\ell}{\sigma_p^4} - 3, \quad (4.45)$$

where

$$K_{ijkl} = \frac{\mathbb{E}[(r_i - \mu_i)(r_j - \mu_j)(r_k - \mu_k)(r_\ell - \mu_\ell)]}{\sigma_i \sigma_j \sigma_k \sigma_\ell}$$

is the cokurtosis between assets i, j, k , and ℓ . Again, the expression for K_p is intractable unless n is very low. Also note that the kurtosis of the portfolio return does not have to be of the same sign or magnitude as the kurtosis of the individual assets, and we cannot generally determine whether the portfolio kurtosis K_p is larger or smaller than the weighted average of the asset kurtoses, $\sum_i \pi_i K_i$.

4.4 Risk reduction through diversification

Section 4.1 illustrates how investors can reduce their risk substantially by investing in two assets. What if the portfolio consists of more, say N , assets? What happens to the risk if we combine more than two assets in a portfolio? This section first presents an example with portfolios of three stocks. Afterwards, we focus on the case of equally-weighted portfolios and explore how the portfolio risk changes when we increase the number of

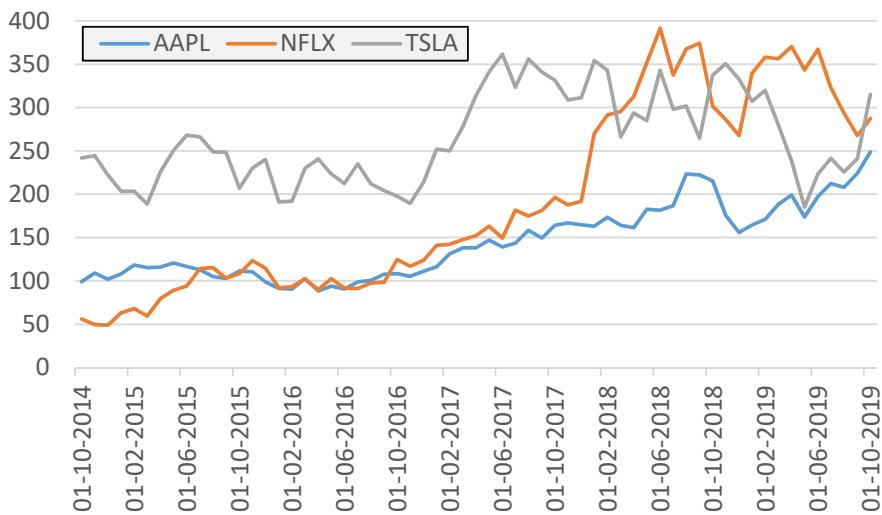


Figure 4.8: Stock prices of Apple, Netflix, and Tesla.

Adjusted monthly closing prices of the stocks of Apple (AAPL), Netflix (NFLX), and Tesla (TSLA) from October 2014 to October 2019. The prices were downloaded from Yahoo Finance in November 2019.

assets in the portfolio.

4.4.1 An example with three stocks

Let us consider portfolios of three stocks, namely Apple (AAPL), Netflix (NFLX), and Tesla (TSLA). Figure 4.8 shows adjusted monthly closing prices of the three stocks from October 2014 to October 2019. The prices were downloaded from Yahoo Finance in November 2019. From the prices, time series of 60 monthly rates of return are derived. The left part of Table 4.2 shows summary statistics for each of the three stocks. For example, Netflix delivered an impressive average monthly return of 3.48% over the period, but also a relatively high standard deviation of 12.56% per month, and both a substantial positive skewness and kurtosis. The estimated correlations were 0.29 (Apple-Netflix), 0.22 (Apple-Tesla), and 0.18 (Netflix-Tesla). By combining the estimated standard deviations and correlations, we can form an estimated variance-covariance matrix.

The right part of Table 4.2 shows summary statistics for five selected portfolios. The first has equal weights on the three stocks, whereas the next three has 60% in one stock and 20% in each of the two others. The final portfolio is an example of a more extreme portfolio with 200% in Apple, 100% in Netflix, and -200% in Tesla. The mean of each portfolio can be calculated from the portfolio weights and the estimated means of the three stocks by applying either Eq. (4.26) or the vector-version (4.30). Alternatively, we can calculate what the rate of return on the portfolio would have been in each month in the sample and then take the average of the 60 monthly portfolio returns. Similarly, the standard deviation of each portfolio can be calculated from the portfolio weights and the estimated variance-covariance matrix by applying either Eq. (4.28) or the vector-matrix-version (4.32). Or, alternatively, calculate the sample standard deviation of the 60 monthly portfolio returns. The median, minimum, and maximum portfolio return cannot be calculated from the same numbers for the individual stocks so we have to derive them from the time series of monthly portfolio returns.

As discussed in Section 4.3, a portfolio's skewness and kurtosis could, in principle,

	Individual stocks			Portfolios				
	AAPL	NFLX	TSLA	Equal	0.6,0.2,0.2	0.2,0.6,0.2	0.2,0.2,0.6	2,1,-2
Mean	1.83%	3.48%	1.12%	2.15%	2.02%	2.68%	1.74%	4.90%
Median	1.74%	3.39%	-0.20%	2.41%	3.63%	2.13%	1.38%	-0.09%
Min	-18.40%	-19.71%	-22.43%	-15.86%	-13.61%	-17.40%	-17.65%	-80.24%
Max	19.62%	40.81%	30.74%	17.96%	14.27%	27.03%	22.14%	58.53%
Std dev	7.63%	12.56%	11.87%	7.49%	6.82%	8.91%	8.53%	28.29%
- weighted				10.69%	9.46%	11.44%	11.16%	
Skew	-0.280	0.691	0.271	-0.040	-0.354	0.343	0.116	-0.244
- weighted				0.227	0.024	0.413	0.245	
Kurt	-0.009	0.722	-0.174	-0.244	-0.448	0.058	0.014	0.363
- weighted				0.180	0.104	0.396	0.038	

Table 4.2: Summary statistics for three stocks and some portfolios.

The statistics are derived from monthly rates of return on the stocks of Apple (AAPL), Netflix (NFLX), and Tesla (TSLA) from November 2014 to October 2019. The returns are calculated from adjusted monthly closing prices downloaded from Yahoo Finance in November 2019. The heading of each portfolio column shows the portfolio weights for Apple, Netflix, and Tesla, respectively.

be calculated from Eqs. (4.44) and (4.45), but then we would have to estimate all the coskewnesses and cokurtoses. Instead, we calculate the portfolio skewness and kurtosis as sample estimates based on the 60 monthly portfolio returns. For the portfolios with only positive weights, the table also shows the weighted average of the standard deviations of the stocks, i.e. $\sum_i \pi_i \sigma_i$, and similarly the weighted average of the skewnesses and kurtoses; such calculations make little sense with negative portfolio weights.

We see that the non-extreme portfolios all have a standard deviation significantly lower than the weighted average of the stocks' standard deviations. This is a clear sign of the diversification of risk, which is also reflected by the fact that the minimum and maximum returns are mostly less extreme for the portfolios than for the individual assets. All four non-extreme portfolios have a skewness lower than the weighted average skewness of the stocks. The same relation holds for the kurtosis. This illustrates that portfolios typically have lower skewness and kurtosis than individual stocks. This is backed by extensive empirical research. For example, [Albuquerque \(2012\)](#) reports that individual stock returns typically have positive skewness, whereas broad stock portfolios have negative skewness. The extreme portfolio in the right-most column has a much higher standard deviation and kurtosis than the other portfolios considered, and also a more significant minimum and maximum return observation.

Figure 4.9 shows some attainable combinations of standard deviations and expected returns (means). The black squares represent the three individual stocks, whereas all the circles represent different portfolios. The red circles are for the five portfolios included in Table 4.2. The green circles represent six portfolios that all have the same expected return as the equally-weighted portfolio, but clearly different levels of standard deviation. The blue circles show various other portfolios of the three stocks. The figure suggests that there is some boundary for how low a standard deviation we can get for any fixed level of expected return. We explore this phenomenon in much more detail in Chapter 7. In that chapter, we shall, among many other things, identify the so-called minimum-variance portfolio that has the lowest possible variance or standard deviation among all portfolios of the available stocks. The large gray circle in the figure corresponds to this minimum-variance portfolio in our three-stock case. We shall also introduce the maximum-slope

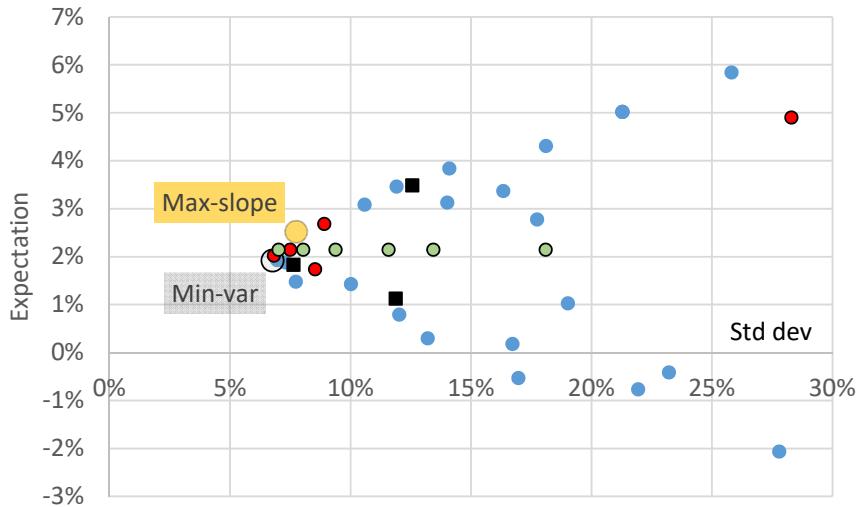


Figure 4.9: Risk-return combinations of three stocks.

The three black squares show the standard deviation and the expected return on the stocks of Apple, Netflix, and Tesla estimated from monthly returns from November 2014 to October 2019. The returns are calculated from adjusted monthly closing prices downloaded from Yahoo Finance in November 2019. Each circle represents a certain portfolio of the three stocks. The red circles are for the five portfolios included in Table 4.2; the green circles show portfolios with the same expected return as the equally-weighted portfolio; the blue circles show various other portfolios. The large gray circle represents the minimum-variance portfolio and the large yellow circle the maximum-slope portfolio.

portfolio that has an attractive risk-return tradeoff. In our case this portfolio corresponds to the yellow circle in the diagram.

4.4.2 An analysis of equal-weighted portfolios

This section focuses on equally weighted portfolios. Of course, with N assets all the portfolio weights are then $\pi_1 = \pi_2 = \dots = \pi_N = 1/N$. In this case the portfolio variance formula (4.27) implies that

$$\text{Var}[r_p] = \sum_{i=1}^N \frac{1}{N^2} \text{Var}[r_i] + \sum_{i=1}^N \sum_{j \neq i}^N \frac{1}{N^2} \text{Cov}[r_i, r_j] = \frac{1}{N^2} \sum_{i=1}^N \text{Var}[r_i] + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{Cov}[r_i, r_j].$$

Define the average variance across the N assets as

$$\overline{\text{Var}} = \frac{1}{N} \sum_{i=1}^N \text{Var}[r_i].$$

With N assets, there are $N(N - 1)$ covariances (which are pairwise identical). The average covariance is therefore

$$\overline{\text{Cov}} = \frac{1}{N(N - 1)} \sum_{i=1}^N \sum_{j=1}^N \text{Cov}[r_i, r_j].$$

Now we can rewrite the portfolio variance stated above as

$$\text{Var}[r_p] = \frac{1}{N^2} N \overline{\text{Var}} + \frac{1}{N^2} N(N-1) \overline{\text{Cov}} = \frac{1}{N} \overline{\text{Var}} + \left(1 - \frac{1}{N}\right) \overline{\text{Cov}}.$$

As we increase N , the number of assets in the portfolio, we can see that the first term becomes smaller and smaller and eventually goes to zero. The second term on the other hand will approach $\overline{\text{Cov}}$ as we increase N . The limit is

$$\text{Var}[r_p] \rightarrow \overline{\text{Cov}} \quad \text{as } N \rightarrow \infty. \quad (4.46)$$

The average covariance is thus a lower bound on the portfolio variance, at least for equally-weighted portfolios. Intuitively, you can think of each asset as being sensitive both to some overall market movements and to some asset-specific events. The risk of an asset can then be split into a market risk component and an asset-specific component. By forming large portfolios with small weights in each asset, you can diversify away the asset-specific risk, but you cannot diversify away the market-wide risk which is captured by the covariances across assets. We will be much more specific about this risk decomposition in later chapters.

As we noted in the two-asset case, the diversification benefits are highly dependent on the correlations between the assets. To illustrate this assume that all assets have identical return standard deviations equal to $\sigma = \text{Std}[r_i]$ and that all pairwise return correlations are equal to $\rho = \text{Corr}[r_i, r_j]$, which is assumed to be non-negative. Then the variance of an equally-weighted portfolio is²

$$\text{Var}[r_p] = \frac{1}{N} \sigma^2 + \left(1 - \frac{1}{N}\right) \rho \sigma^2 = \sigma^2 \left(\rho + \frac{1-\rho}{N}\right).$$

Increasing N towards infinity, the variance approaches a lower limit of $\rho\sigma^2$, i.e., the standard deviation has a lower limit of $\sigma\sqrt{\rho}$.

Table 4.3 and Figure 4.10 illustrate how the standard deviation $\sigma_p = \sqrt{\text{Var}[r_p]}$ declines as more assets are added to the portfolio. Each asset is assumed to have a return standard deviation of $\sigma = 0.4$. First note that for a portfolio of few assets, the addition of another asset reduces the variance by more when the correlation is low. For example, going from 10 to 11 assets reduces the portfolio standard deviation by 4.7% if $\rho = 0$, by 0.6% if $\rho = 0.4$, and only by 0.1% if $\rho = 0.8$. Secondly, the portfolio standard deviation is bounded from below by $\sqrt{\rho\sigma^2} = \sqrt{\rho}\sigma$. This bound is 0 if $\rho = 0$, it is 0.2530 if $\rho = 0.4$, and 0.3578 if $\rho = 0.8$. Note that by combining just two uncorrelated assets, you obtain a lower standard deviation than can be achieved by combining infinitely many assets with correlation $\rho = 0.8$. And by combining just three uncorrelated assets, you get below the lower bound for $\rho = 0.4$. These results highlight the importance of the correlation for diversification purposes. Generally, stocks are only modestly correlated so a lot of risk can be eliminated by investing in many stocks at the same time. In contrast, different government bonds tend to be highly correlated with each other so that less is gained by spreading your investment over many bonds. We return to these considerations in subsequent chapters.

Table 4.3 and Figure 4.10 further show that the marginal benefits from diversification

²Of course, the variance must stay non-negative, which implies that $\rho \geq -1/(N-1)$, i.e., there is a bound on how negatively correlated many assets can be. In the limit when N approaches infinity, the bound goes to zero.

N	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
	σ_p	change in σ_p	σ_p	change in σ_p	σ_p	change in σ_p
1	0.4000		0.4000		0.4000	
2	0.2828	-29.3%	0.3347	-16.3%	0.3795	-5.1%
3	0.2309	-18.4%	0.3098	-7.4%	0.3724	-1.9%
4	0.2000	-13.4%	0.2966	-4.3%	0.3688	-1.0%
5	0.1789	-10.6%	0.2884	-2.8%	0.3666	-0.6%
10	0.1265		0.2713		0.3622	
11	0.1206	-4.7%	0.2697	-0.6%	0.3618	-0.1%
15	0.1033		0.2653		0.3607	
16	0.1000	-3.2%	0.2646	-0.3%	0.3606	-0.1%
20	0.0894		0.2623		0.3600	
21	0.0873	-2.4%	0.2619	-0.2%	0.3599	-0.0%
30	0.0730		0.2592		0.3593	
31	0.0718	-1.6%	0.2590	-0.1%	0.3592	-0.0%
40	0.0632		0.2577		0.3589	
41	0.0625	-1.2%	0.2576	-0.0%	0.3589	-0.0%
∞	0.0000		0.2530		0.3578	

Table 4.3: Standard deviation of equally-weighted portfolios.

The table shows how the standard deviation of the return on an equally-weighted portfolio varies with the number of assets N in the portfolio. All assets are assumed to have a return standard deviation of $\sigma = 0.4$. All pairs of assets are assumed to have a return correlation of ρ as indicated in the column headings. The table also shows the percentage reduction in the standard deviation by going from $N - 1$ to N assets in the portfolio. For example, for $\rho = 0.0$ the reduction in standard deviation when going from one asset ($\sigma_p = 0.4000$) to two assets ($\sigma_p = 0.2828$) is $(0.2828 - 0.4000)/0.4000 = -0.293 = -29.3\%$.

are decreasing with the number of assets in the portfolio. By adding an extra asset to your portfolio, the reduction in risk is largest if the existing portfolio contains few assets, as reflected by the convex shape of the curves in Figure 4.10. The sensitivity of the marginal risk reduction to the number of assets is particularly large when the assets have low correlations with each other.

Above, we considered equally weighted portfolios. With a given set of assets we can typically obtain a better diversification (i.e., a lower return variance or standard deviation) with a different set of portfolio weights as we already saw in the two-asset case. Chapter 7 explains how we can find the best diversified multi-asset portfolio, where ‘best’ is in the sense of having the lowest possible return standard deviation.

4.4.3 Diversification with portfolios having non-equal weights

The previous subsection studied the diversification benefits of equally-weighted portfolios. How big are such benefits for portfolios with non-equal weights? Of course, this depends on how different the weights are. For a portfolio with 99% invested in a single stock, the return standard deviation is going to be very close to that of the stock. Nevertheless, the analysis below indicated that even quite unbalanced portfolios of many stocks generate huge diversification benefits.

As in the previous subsection, suppose that all assets have the same return standard deviation σ and all pairs of assets have the same return correlation ρ . For a portfolio

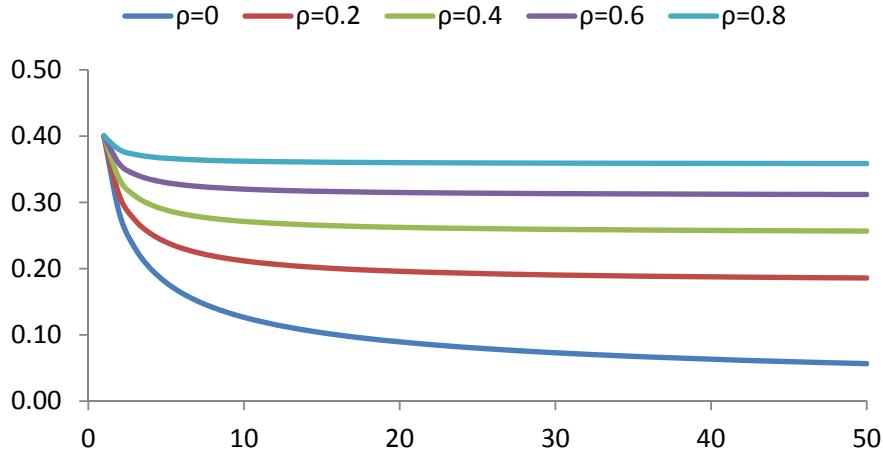


Figure 4.10: Portfolio standard deviation as a function of the number of assets.

The curves show how the standard deviation of the return on an equally weighted portfolio varies with the number of assets in the portfolio. All assets are assumed to have a return standard deviation of $\sigma = 0.4$. All pairs of assets are assumed to have a return correlation of ρ . Each curve corresponds to the indicated value of ρ .

consisting of N assets with weights π_1, \dots, π_N summing to one, the return variance from Eq. (4.27) then becomes

$$\text{Var}[r_p] = \sigma^2 \sum_{i=1}^N \pi_i^2 + 2\rho\sigma^2 \sum_{i=1}^N \sum_{j=i+1}^N \pi_i \pi_j. \quad (4.47)$$

The lowest portfolio variance is obtained with equal weights, $\pi_i = 1/N$ for all $i = 1, \dots, N$. But how much bigger is the variance with non-equal weights? Of course, there are infinitely many ways in which we can choose N weights that sum to one, so we need to impose structure on the weights to obtain some insights.

A relevant and tractable case is where the weights are related to a power series. Let α be a positive real number smaller than or equal to one. Then we can define portfolio weights $\pi_1, \pi_2, \dots, \pi_N$ by two conditions: $\pi_{i+1} = \alpha \pi_i$ for all $i = 1, 2, \dots, N - 1$ and, of course, $\pi_1 + \pi_2 + \dots + \pi_N = 1$. With $\alpha = 1$, all weights are equal to $1/N$. With $\alpha < 1$, the weights π_i will be decreasing. In fact, it can be shown that the sum K of the terms $\alpha, \alpha^2, \dots, \alpha^N$ is given by

$$K = \sum_{i=1}^N \alpha^i = \alpha + \alpha^2 + \dots + \alpha^N = \begin{cases} N, & \text{if } \alpha = 1, \\ \frac{\alpha}{1-\alpha}(1 - \alpha^N), & \text{if } \alpha \neq 1. \end{cases} \quad (4.48)$$

Therefore the weights can be written as

$$\pi_i = \frac{\alpha^i}{K} = \begin{cases} \frac{1}{N}, & \text{if } \alpha = 1, \\ \alpha^{i-1} \frac{1-\alpha}{1-\alpha^N}, & \text{if } \alpha \neq 1. \end{cases} \quad (4.49)$$

When the portfolio weights have this structure, the portfolio variance in Eq. (4.47) turns

out to be

$$\text{Var}[r_p] = \sigma^2 \frac{(1-\alpha)(1-\alpha^{2N}) + 2\rho\alpha(1-(1+\alpha)\alpha^{N-1} + \alpha^{2N-1})}{(1+\alpha)(1-\alpha^N)^2}. \quad (4.50)$$

For completeness, here is a proof of this result that relies on several applications of (4.48):

$$\begin{aligned} \text{Var}[r_p] &= \sigma^2 \sum_{i=1}^N \pi_i^2 + 2\rho\sigma^2 \sum_{i=1}^N \sum_{j=i+1}^N \pi_i \pi_j = \frac{\sigma^2}{K^2} \sum_{i=1}^N \alpha^{2i} + 2\frac{\rho\sigma^2}{K^2} \sum_{i=1}^N \sum_{j=i+1}^N \alpha^{i+j} \\ &= \frac{\sigma^2}{K^2} \sum_{i=1}^N (\alpha^2)^i + 2\frac{\rho\sigma^2}{K^2} \sum_{i=1}^N \alpha^{2i} \sum_{u=1}^{N-i} \alpha^u \\ &= \frac{\sigma^2}{K^2} \frac{\alpha^2}{1-\alpha^2} \left(1 - (\alpha^2)^N\right) + 2\frac{\rho\sigma^2}{K^2} \frac{\alpha}{1-\alpha} \sum_{i=1}^N \alpha^{2i} (1 - \alpha^{N-i}) \\ &= \frac{\sigma^2}{K^2} \frac{\alpha^2}{1-\alpha^2} \left(1 - \alpha^{2N}\right) + 2\frac{\rho\sigma^2}{K^2} \frac{\alpha}{1-\alpha} \left(\sum_{i=1}^N \alpha^{2i} - \alpha^N \sum_{i=1}^N \alpha^i\right) \\ &= \frac{\sigma^2}{K^2} \frac{\alpha^2}{1-\alpha^2} \left(1 - \alpha^{2N}\right) + 2\frac{\rho\sigma^2}{K^2} \frac{\alpha}{1-\alpha} \left(\frac{\alpha^2}{1-\alpha^2} (1 - \alpha^{2N}) - \alpha^N \frac{\alpha}{1-\alpha} (1 - \alpha^N)\right) \\ &= \frac{\sigma^2}{K^2} \frac{\alpha^2}{1-\alpha^2} \left(1 - \alpha^{2N}\right) + 2\frac{\rho\sigma^2}{K^2} \frac{\alpha^3}{(1-\alpha)^2} \left(\frac{1}{1+\alpha} (1 - \alpha^{2N}) - \alpha^{N-1} (1 - \alpha^N)\right), \end{aligned}$$

and after substitution of K and further simplifications, this leads to (4.50).

As a measure of how unbalanced or concentrated a portfolio is, we can calculate the total weight of the 10% assets with the largest weights. Of course, the top 10% assets have a total weight of at least 10%. Assuming that $\alpha < 1$ and N is a multiple of 10, this total weight is

$$\begin{aligned} W &= w_1 + w_2 + \cdots + w_{N/10} = \frac{1}{K} \left(\alpha + \alpha^2 + \cdots + \alpha^{N/10} \right) \\ &= \frac{1}{K} \frac{\alpha}{1-\alpha} (1 - \alpha^{N/10}) = \frac{1 - \alpha^{N/10}}{1 - \alpha^N}, \end{aligned}$$

where the third equality applies (4.48) with $N/10$ replacing N . For a given N , we choose α so that the total weight of the top 10% equals a given number W bigger than 10%.

Figure 4.11 illustrates different portfolios of 100 stocks. The flat black line represents an equal-weighted portfolio with a 1% weight on each stock. The yellow curve corresponds to a portfolio where the 10% largest stocks have a total weight of 20% which is obtained for $\alpha \approx 0.9817$. Similarly the red dashed, the solid green, and the dotted grey curves represent portfolios where the top 10% stocks have a total weight of 40%, 60%, or 80%, respectively, corresponding to α -values of approximately 0.9506, 0.9125, and 0.8513.

Let us assume that all stocks have a return standard deviation of $\sigma = 0.4$ and that all pairwise correlations are $\rho = 0.4$. The upper part of Table 4.4 shows the standard deviation for different portfolios of 10, 100, or 500 stocks, where the portfolio weights have the described structure and α is set so that 10% largest stocks have a total weight of 10%, 20%, ..., 80%. The standard deviation for equal-weighted portfolios is thus under the heading ‘10%’. The table confirms the intuitive relation that the more unbalanced the portfolio weights are, the larger is the portfolio standard deviation. The table further

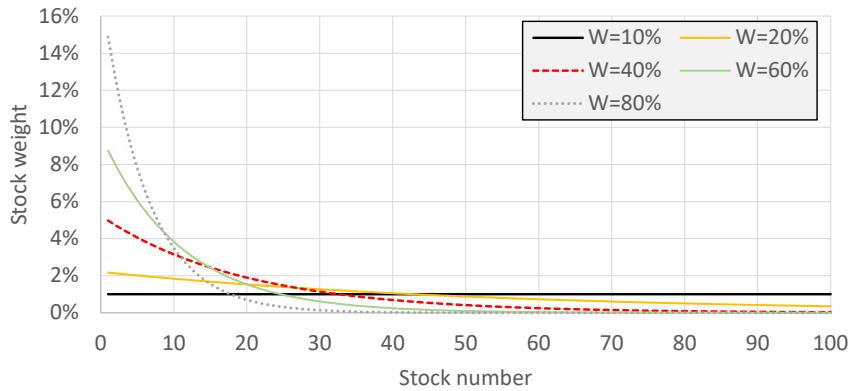


Figure 4.11: Unbalanced portfolio weights.

Each curve represents a given portfolio of 100 stocks and shows the weights of the individual stocks. Each portfolio is characterized by the total weight W of the ten (equal to 10%) stocks with the largest weights. See the text for details on how the weights are defined.

shows that a certain degree of portfolio concentration has a lower impact on the standard deviation when there are more stocks in the portfolio. For example, if we go from an equal-weighted portfolio to a portfolio where the top 10% stocks have a total weight of 50%, the standard deviation increases by 14.2% (from 0.2713 to 0.3099) with 10 stocks in the portfolio, by 1.8% with 100 stocks, and only by 0.4% with 500 stocks. For highly concentrated portfolios where the top 10% stocks have a total weight of 80%, the relative increase in standard deviation compared to an equally-weighted portfolio is 31.9% with 10 stocks, 5.1% with 100 stocks, and only 1.0% with 500 stocks.

The lower part of Table 4.4 compares the portfolios' standard deviations to the 0.4 standard deviation per stock. For example, for portfolios with 100 stocks, a portfolio with 50% in the largest 10% of stocks achieves 96.8% of the maximum reduction in standard deviation (reduction of $0.4 - 0.2595$ relative to the $0.4 - 0.2549$ for the equally-weighted portfolio). The portfolio with 80% in the largest 10% of stocks achieves 91.1% of the maximum reduction. With 10 stocks, the corresponding numbers are only 70.0% and 32.8%. With 500 stocks, the numbers are 99.4% and 98.2%.

These findings support the view that even quite unbalanced portfolios of many stocks generate huge diversification benefits. In practice, not all stocks have the same standard deviation as we have assumed. In fact, the tendency is that stocks in larger companies have lower return standard deviation than stocks in smaller companies. Since most investors must have higher weights on the larger stocks, this means that the standard deviation of typical unbalanced portfolios of many stocks tends to be quite low.

4.5 Special portfolios: arbitrage, replication, and tracking

An **arbitrage** is a portfolio offering a riskfree profit. In its most strict definition, an arbitrage is a portfolio satisfying one of two conditions:

- (1) the portfolio can be purchased at a negative cost—i.e., you receive money when obtaining the portfolio—and in the future the portfolio will always give you a non-negative cash flow, no matter what happens;
- (2) the portfolio can be purchased at zero or even negative cost, and its future cash flow

Weight W of 10% largest stocks								
	10%	20%	30%	40%	50%	60%	70%	80%
<i>Portfolio standard deviation</i>								
$N = 10$	0.2713	0.2759	0.2853	0.2968	0.3099	0.3243	0.3401	0.3578
$N = 100$	0.2549	0.2554	0.2564	0.2578	0.2595	0.2615	0.2641	0.2678
$N = 500$	0.2534	0.2535	0.2537	0.2540	0.2543	0.2547	0.2553	0.2560
<i>Percentage diversification</i>								
$N = 10$	100.0%	96.4%	89.1%	80.2%	70.0%	58.8%	46.5%	32.8%
$N = 100$	100.0%	99.7%	98.9%	98.0%	96.8%	95.4%	93.6%	91.1%
$N = 500$	100.0%	99.9%	99.8%	99.6%	99.4%	99.1%	98.7%	98.2%

Table 4.4: Diversification with unequally-weighted portfolios.

The upper part of the table shows the return standard deviation of different portfolios of N stocks. The lower part of the table shows the achieved percentage of diversification calculated as the reduction in standard deviation from the single-stock case relative to the maximum reduction possible (achieved by the equally-weighted portfolio). All stocks have a return standard deviation of $\sigma = 0.4$, and all pairs of stocks have a return correlation of $\rho = 0.4$. The column heading is the total weight of the 10% largest stocks in the portfolio and thus indicates how different the weights in the portfolio are. The specific weights are given by (4.49) where α is chosen so that the total weight of the 10% largest stocks in the portfolio indeed matches the column heading.

is non-negative for sure and with a strictly positive probability of getting a strictly positive payment at some date (this is like a free lottery ticket).

An arbitrage offers something for nothing and is attractive to any investor. If one or more investors identify an arbitrage, they would purchase that portfolio immediately (if they can), and the increasing demand would drive up the price of the portfolio until it no longer constitutes an arbitrage. When the portfolio only involves assets that are traded in public markets at very low transaction costs, we should not expect the arbitrage to exist, at least not for more than a very short time period (seconds or maybe only milliseconds). Generally, we should thus expect that prices in financial markets are set so that arbitrage opportunities are non-existing. This is the *absence of arbitrage pricing principle*.

An arbitrage is typically constructed as a position in a certain asset and an off-setting position in a portfolio replicating that asset. A **replicating portfolio** for a given asset is a portfolio that provides exactly the same cash flow as the reference asset, no matter what happens.

Suppose the price of a replicating portfolio for an asset exceeds the price of the asset itself. Then you can construct an arbitrage by purchasing the asset and selling the replicating portfolio. Since you buy cheap and sell expensive, you make a profit right now. Having the short position in the portfolio, you have to deliver the future cash flow of the portfolio, but this is exactly offset by the cash flow of the asset. Hence, the net future cash flow is zero. You get money now, and you never have to pay anything: a clear arbitrage.

Conversely, if the price of the asset exceeds the price of a portfolio replicating the asset, the arbitrage consists of selling the asset and purchasing the replicating portfolio. Again you make a profit now and the net future cash flow is zero. This is an arbitrage.

In other words, if an asset can be replicated by a portfolio, the only no-arbitrage price of the asset is the price of the replicating portfolio. This observation is often useful to price rather complicated assets that can be decomposed into a portfolio of simpler and easier-to-price assets. We shall use this idea in Chapter 5 to link the prices of coupon

bonds to the prices of the simpler so-called zero-coupon bonds. Here is a basic example:

Example 4.7

A riskfree bond gives a cash flow of \$5,000 one year from now, \$105,000 two years from now, and no other payments. Let us refer to this bond as the bullet bond (the bond market terminology is explained in Chapter 5). What is a fair price of the bullet bond?

Suppose that in the bond market, at a price of \$980, you can buy a riskfree bond whose only payment is \$1,000 one year from now. Let us refer to this bond as the one-year zero-coupon bond. And suppose that, at a price of \$960, you can buy a riskfree bond whose only payment is \$1,000 two years from now. Call this the two-year zero-coupon bond.

Form a portfolio of 5 units of the one-year zero-coupon bond and 105 units of the two-year zero-coupon bond. This portfolio offers a total payment of $5 \times \$1,000 = \$5,000$ one year from now and $105 \times \$1,000 = \$105,000$ two years from now, so the portfolio replicates the bullet bond. The price of the portfolio is

$$5 \times \$980 + 105 \times \$960 = \$105,700.$$

This is then the fair price of the bullet bond in the sense that, if the price would be different, an arbitrage could easily be constructed.

In the above example there is no uncertainty about the future cash flows, but replication can also be across possible outcomes. Here is an example:

Example 4.8

Suppose Apple stocks currently trade at \$116. You are confident that one year from now the stock price has either increased or decreased by 25% so that it is either \$145 or \$87. Your broker agrees with you and offers you a contract that pays you \$29 if the stock price goes up to \$145 (the \$29 is exactly the increase in the price) and nothing if the stock price goes down to \$87. You get the upside, but not the downside. This contract is a so-called call option on the stock. How much should you be willing to pay for the option?

Suppose the riskfree rate over the year is zero. Consider a portfolio consisting of half a unit of the stock and a loan of \$43.50. The current price of the portfolio is

$$0.5 \times \$116 - \$43.50 = \$14.50.$$

If the stock price goes up to \$145, you can sell the half unit of the stock, repay the loan, and cash in

$$0.5 \times \$145 - \$43.50 = \$29.$$

If the stock price goes down to \$87, you sell the half unit of the stock and repay the loan, generating a net payment of

$$0.5 \times \$87 - \$43.50 = \$0.$$

Note that whether the stock price goes up or down, the value of the portfolio is exactly equal to the payoff of the option. Hence, the portfolio is replicating the option. Since the price of the portfolio is \$14.50, this is the maximum amount you should be willing to pay

for the option. The pricing of options and other derivatives is discussed in much more detail in Chapters 14 and 15.

Investors focusing on identifying and exploiting arbitrage opportunities are called arbitrageurs. Practitioners often use the term arbitrage in a less strict sense, for example to represent an investment they believe is going to be profitable with a very large probability, although strictly smaller than one. If two assets are traded at significantly different prices, an investor expecting the assets to provide very similar future cash flows might be tempted to take a long position in the cheaper asset and a short position in the more expensive asset. Some hedge funds frequently engage in such transactions to profit from what they believe is a relative mispricing in the markets, but generally they acknowledge that the strategy involves some risk.

The term **tracking portfolio** is sometimes used for a portfolio that tracks a certain asset or other quantity as well as possible. The tracking portfolio for an asset could be a portfolio having a value that stays as close as possible to the price of the benchmark asset. The tracking error measures the difference between the tracking portfolio and the benchmark. For example, some portfolio managers try to track the S&P 500 stock index with a portfolio of relatively few selected stocks and maybe other financial assets. Of course, the S&P 500 index can, in principle, be perfectly tracked—which really means replicated—by a portfolio of all the 500 stocks the index consists of. But the portfolio managers hope to find a cheaper portfolio that still moves around very much like the index.

Other investors might want to design a portfolio of traded financial assets that tracks the movements in some macroeconomic variables like the GDP growth rate or the inflation rate. Such tracking portfolios can be derived mathematically, for example, by minimizing the variance of the difference between the value of the portfolio and the value of the benchmark variable.

4.6 Exercises

Exercise 4.1. Show the derivations leading to the Equation (4.7).

Exercise 4.2. Suppose the annual return on stocks in the company ABC is normally distributed with a mean of 10% and a standard deviation of 40%.

- What is the probability that the return on ABC stocks over the next year is negative?
- What is the 5% value-at-risk of ABC stocks over the next year?
- What is the 1% value-at-risk of ABC stocks over the next year?

Suppose that the annual return on stocks in the company XYZ is also normally distributed with a mean of 10% and a standard deviation of 40%. Let ρ denote the correlation between the returns on the two stocks and consider an equally-weighted portfolio of the two stocks.

- Answer the following questions for $\rho = -0.8$, $\rho = 0$, and $\rho = 0.8$:
 - What is the expectation and the standard deviation of the return on the portfolio?
 - What is the probability that the portfolio return over the next year is negative?
 - What is the 5% value-at-risk of the portfolio over the next year?
 - What is the 1% value-at-risk of the portfolio over the next year?
- Explain and discuss the impact of the correlation on the results found above.

Exercise 4.3. Suppose that the annual return on stocks in the company *Sorensen Soups* is normally distributed with a mean of 20% and a standard deviation of 60%.

- (a) What is the probability of a negative return on the stocks of *Sorensen Soups* over the next year? What is the 5% value-at-risk of *Sorensen Soups* stocks over the next year? (In other words: find x so that the return is less than x with a 5% probability.)

Suppose that the annual return on stocks in the company *Lando Lasers* is also normally distributed with a mean of 20% and a standard deviation of 60%. Consider an equally-weighted portfolio of stocks in *Sorensen Soups* and stocks in *Lando Lasers*.

- (b) What is the probability of a negative return and the 5% value-at-risk of the portfolio over the next year if the returns on the two stocks have a correlation of $\rho = 0$? What if the correlation is $\rho = 0.5$? Comment on the role of the correlation for portfolio risk.

Now suppose that you invest in an equally-weighted portfolio of 10 stocks where each of the stocks has a mean of 20% and a standard deviation of 60%.

- (c) What is the probability of a negative return and the 5% value-at-risk of the portfolio over the next year if the returns on all 10 stocks have a pairwise correlation of $\rho = 0$? What if the correlation is $\rho = 0.5$? Comment on your findings.

Exercise 4.4. Let

$$\underline{\underline{A}} = \begin{pmatrix} 1 & -1 & 3 \\ 2 & 3 & 4 \end{pmatrix}, \quad \underline{\underline{B}} = \begin{pmatrix} 1 & 3 \\ 0 & -1 \\ 1 & 2 \end{pmatrix}.$$

- (a) Can you compute $\underline{\underline{A}} + \underline{\underline{B}}$?
- (b) Compute $\underline{\underline{A}}\underline{\underline{B}}$ and $\underline{\underline{B}}\underline{\underline{A}}$ both by hand and in Excel.
- (c) Determine $\underline{\underline{A}}^\top$ and $\underline{\underline{B}}^\top$.
- (d) Compute $(\underline{\underline{A}}\underline{\underline{B}})^\top$ and $\underline{\underline{B}}^\top\underline{\underline{A}}^\top$ and compare.

Exercise 4.5. Let

$$\underline{\underline{A}} = \begin{pmatrix} 1 & -1 & 3 \\ 2 & 3 & 4 \\ 2 & 1 & 2 \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Compute $\underline{\underline{A}}\mathbf{1}$ and $\mathbf{1}^\top\underline{\underline{A}}$ both by hand and in Excel. Here $\mathbf{1}^\top$ denotes the transpose of the column vector $\mathbf{1}$.

Exercise 4.6. Suppose you want to solve the following three equations in the three unknowns x_1, x_2, x_3 :

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 31, \\ 3x_1 &\quad + 4x_3 = -31, \\ x_2 + 5x_3 &= 62. \end{aligned}$$

- (a) Explain why you can write this system of equations as

$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & 0 & 4 \\ 0 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 31 \\ -31 \\ 62 \end{pmatrix}.$$

- (b) Use Excel to show that the solution to the system of equations is

$$x_1 = -21, \quad x_2 = 22, \quad x_3 = 8.$$

Exercise 4.7. In Exercise 3.5, you should have estimated the averages, the variances, and covariances of the monthly returns on three stocks, namely Coca-Cola, Pepsico, and The Home Depot.

- (a) Based on your estimates, what is the estimate of the expected monthly return, the variance, and the standard deviation of an equally-weighted portfolio of the three stocks? Do these computations both by hand and by using Excel. Annualize your results.

- (b) Consider at least 10 different portfolios of the three stocks and compute for each (using the vector and matrix functions in Excel) the expectation and the standard deviation of the monthly return and annualize these values appropriately. You decide on the precise portfolios, but let the portfolio weights differ a lot from case to case to get an impression of the possible combinations of standard deviation and expected return.
- (c) Draw a scatter diagram with standard deviation along the horizontal axis and expected return along the vertical axis and show the points corresponding to the portfolios considered. Also show the points corresponding to the three individual stocks. As always: Comment on your findings!
- (d) Compute (in Excel) the covariance and the correlation between the following two portfolios:
 1. 10% in Coca-Cola, 10% in Pepsico, 80% in The Home Depot
 2. 80% in Coca-Cola, 10% in Pepsico, 10% in The Home Depot

Exercise 4.8. Suppose you can trade in infinitely many stocks whose returns are normally distributed with the same expected rate of return μ and the same standard deviation σ . Furthermore, for any pair of stocks the correlation is the same, namely ρ . In the following, consider portfolios of these stocks. Let the positive integer N denote the number of stocks in the portfolio.

- (a) For any fixed value of N , explain in words why a risk-averse investor would prefer the equally-weighted portfolio to any other portfolio.
- (b) Show that the standard deviation of the return of an equally-weighted portfolio of N stocks can be written as

$$\text{Std}[r_p] = \sigma \sqrt{\rho + \frac{1 - \rho}{N}}.$$

- (c) Suppose $\sigma = 0.4$ and $\rho = 0.25$. What is the largest lower bound on the portfolio standard deviation, i.e., the largest value of σ_{low} such that $\text{Std}[r_p] \geq \sigma_{\text{low}}$ for all values of N ? How many stocks do you need to invest in to get the portfolio standard deviation below 0.24?

Exercise 4.9. Suppose you can trade in infinitely many stocks whose annual returns are normally distributed with the same expected rate of return μ and the same standard deviation σ . Furthermore, for any pair of stocks the correlation is the same, namely ρ . In the following, consider equally-weighted portfolios of these stocks. Let the positive integer N denote the number of stocks in the portfolio.

- (a) Explain why the 5% value at risk for the portfolio's rate of return over a one-year investment horizon is given by

$$\text{VaR} = \mu - 1.645 \times \sigma \sqrt{\frac{1}{N} + \left(1 - \frac{1}{N}\right)\rho}.$$

- (b) Assume that $\mu = 0.1$, $\sigma = 0.4$, and $\rho = 0.5$. What is the 5% value at risk derived in (a) for $N = 10$? Explain carefully what this number means. Illustrate graphically how this value at risk depends on N .
- (c) In the graph constructed in (b), add a curve showing how the value at risk depends on N when $\rho = 0.2$, whereas $\mu = 0.1$ and $\sigma = 0.4$ as before. Explain why the value of ρ has the effect on the value at risk that can be seen from your graph.
- (d) Find a formula (in terms of μ , σ , and ρ) for the limit of the 5% value at risk when $N \rightarrow \infty$. Given that $\mu = 0.1$ and $\sigma = 0.4$, what is the value of that limit for $\rho = 0.2$ and for $\rho = 0.5$?

Exercise 4.10. The supplementary material for these lecture notes includes the Excel file named `Exercise4_10_data.xlsx` that contains the adjusted monthly closing prices from Yahoo Finance between December 2015 and December 2020 on the following six exchange-traded funds (ETFs):

Ticker symbol	Asset class	Specific ETF name
GLD	Gold	SPDR Gold Shares
VNQ	Commercial real estate	Vanguard Real Estate Index Fund ETF Shares
SPTL	Long-term Treasury bonds	SPDR Portfolio Long Term Treasury ETF
SPY	Stock market	SPDR S&P 500 ETF Trust
USO	Oil	United States Oil Fund, LP
LQD	Corporate bonds	iShares iBoxx \$ Invest Grade Corp Bond ETF

In addition the Excel document contains the one-month Treasury rate at the end of each month, which represents the riskfree rate.

- (a) For each of the six ETFs, calculate the returns each month.
- (b) For each of the six ETFs, compute the average and standard deviation of the monthly returns in this dataset. Annualize these numbers.
- (c) For each of the six ETFs, draw a histogram of the observed monthly returns, and estimate the skewness and kurtosis.
- (d) For each of the six ETFs, compute the monthly and the annual Sharpe ratio.
- (e) Compute the variance-covariance matrix and the correlation matrix. Which pairs of ETFs have the highest correlation? Which pairs of ETFs have the lowest correlation?
- (f) What is the average and the standard deviation of the monthly return on an equally-weighted portfolio of the six ETFs? Discuss the diversification benefits in this case. Without using optimization techniques as those introduced in subsequent chapters, suggest a change in the portfolio weights away from the equal weights which would reduce the portfolio risk even more.

Exercise 4.11. In this problem, all returns considered are measured over the next year. The expected rate of return is 0.12 or 12% on stock A and 0.08 or 8% on stock B. For both stocks, the standard deviation of the rate of return is 0.25 or 25%. The correlation between the rates of return on the two stocks is 0.5.

- (a) Suppose that the rate of return on stock A is normally distributed. What is the probability that the rate of return on A is lower than -30% ?
- (b) Consider a portfolio of the two stocks with 20% invested in A and 80% in B. What is the expected rate of return and the return standard deviation on this portfolio?
- (c) Determine the portfolio of the two stocks that has the lowest return standard deviation. What is the standard deviation of this portfolio?
- (d) Suppose the riskfree rate of return over the next year is 0.02 or 2%. What is the Sharpe ratio of each of the stocks? Determine the portfolio of the two stocks that has the largest Sharpe ratio. What is the Sharpe ratio of this portfolio?
- (e) Suppose that the rates of return on the two stocks are jointly normally distributed. What is the probability that the rate of return on stock A will exceed the rate of return on the portfolio found in (c), i.e. the portfolio with the lowest standard deviation?

CHAPTER 5

Bonds

The markets for fixed income securities—meaning bonds and related debt securities—were introduced in Section 1.3. This chapter takes a closer look at the different types of bonds and how they are valued. We also discuss the risks of bond investments, and how bonds can be used by investors to manage interest rate risk.

Section 5.1 reviews the characteristics of the main types of bonds, including bullet bonds, zero-coupon bonds, and annuity bonds. Section 5.2 explains how bonds are priced when the same discount rate is applied to all the bond's payments. Section 5.3 defines the yield of a bond as the discount rate that leads to a bond price identical to the current market price. The concept of yield curves is also introduced and some empirical findings on yield curves are presented. In particular, we focus on the zero-coupon yield curve that shows how the yields of zero-coupon bonds vary with the time to maturity of the bonds.

Section 5.4 discusses how the no-arbitrage pricing principle induces ties between the prices of bonds with overlapping payment dates. We explain how this idea can allow us to derive zero-coupon yields from prices of coupon bonds. Next, Section 5.5 introduces forward rates and links them to zero-coupon yields. Section 5.6 briefly explains how the level and shape of the yield curve are influenced by investors' expectations and uncertainty about the future state of the economy, but also by some investors having preferences for certain bond maturities. Stylized empirical facts on bond returns and interest rates are presented in Section 5.7.

Bond prices vary with interest rates, and Section 5.8 introduces the duration and the convexity and show how these variables quantify the interest rate sensitivity of bonds. Section 5.9 applies the duration and convexity in forming so-called immunization strategies that aim at protecting a position against interest rate risk. Section 5.10 discusses bonds for which the future payments are not fully known due to a floating coupon rate, the potential default of the issuer, or the issuer's option to buy back the bond before maturity. Such features are relevant for many corporate bonds and mortgage-backed bonds. Finally, Section 5.11 briefly discusses the role of bonds in broad portfolios.

5.1 Bond types and characteristics

A bond is nothing but a tradable loan contract. The issuer borrows a certain amount of money—equal to the face value of the bond—and promises to repay the amount with interest according to a certain schedule stipulated in the contract. The final scheduled

payment date is called the maturity date of the bond.

Bonds are issued by governments, private and public corporations, and financial institutions. The initial buyer of a bond is the original lender and if he holds on to the bond until maturity, he has a legal claim to the entire promised payment schedule. Most bonds are subsequently traded at organized exchanges or in semi-organized over-the-counter (OTC) markets so that the claim to the future payments can change hands. The main bond investors are pension funds and other financial institutions, central banks, corporations, and households. Bonds are traded with various maturities and with various types of payment schedule. Moreover, in the so-called money markets large financial institutions offer various bond-like loan agreements of a maturity of less than one year. Below, we will introduce some basic concepts and terminology.

We distinguish between zero-coupon bonds and coupon bonds. A **zero-coupon bond** is the simplest possible bond. It promises a single payment equal to the face value at a single future date, the maturity date of the bond. Bonds promising more than one payment when issued are referred to as **coupon bonds**. A coupon bond has a sequence of payment dates occurring at regular intervals, typically annually, semi-annually, quarterly, or monthly. Let us denote the payment dates by $i = 1, 2, \dots, n$ so that time is measured in units of the payment interval, e.g., in quarters if the bond has quarterly payments. The payment at time i (i.e., after i periods) is denoted by M_i .

The bond payments are determined by the **face value**, the **coupon rate**, and the **amortization principle** of the bond. At each payment date i , the total bond payment M_i equals the sum of an interest payment I_i and a repayment of debt X_i (sometimes referred to as an installment), i.e.,

$$M_i = I_i + X_i.$$

The outstanding debt or face value immediately after payment date i is denoted by F_i . Immediately after the bond is issued, the outstanding debt equals the initial face value F_0 . Afterwards, the outstanding debt is reduced by any repayments:

$$F_i = F_{i-1} - X_i.$$

After the final payment date, the outstanding debt must equal zero, $F_n = 0$. The interest payment at any date i equals the product of the coupon rate q and the outstanding debt after the previous payment date,

$$I_i = qF_{i-1}.$$

The face value is also known as the par value or principal of the bond, and the coupon rate is also called the nominal rate or stated interest rate. In many cases, the coupon rate is quoted as an annual rate even when payments occur more frequently. If a bond with a payment frequency of δ years has a quoted coupon rate of Q , this means that the periodic coupon rate is $q = \delta Q$. A bond with a quoted coupon rate of 8% and quarterly payments has a periodic coupon rate of $q = 2\%$. Here are the standard coupon bond types:

Bullet bonds (or straight-coupon bonds). Most government bonds and corporate bonds are so-called bullet bonds, typically with one or two annual payment dates. For example, the vast majority of the many bonds issued by the United States Department of the Treasury—or simply Treasury bonds—are bullet bonds with semi-annual payments. All the payments before the final payment are equal to the product of the coupon rate and the face value. The final payment at the maturity date is the sum of the same interest rate payment and the face value. If q denotes the periodic coupon rate and F_0 the initial

face value of the bond, the payments are therefore

$$M_i = \begin{cases} qF_0, & i = 1, \dots, n-1 \\ (1+q)F_0, & i = n \end{cases} \quad (5.1)$$

The face value thus remains constant throughout the life of a bullet bond. Of course, for $q = 0$ or $n = 1$, the bullet bond is really a zero-coupon bond.

Annuity bonds. Many mortgage-backed bonds and a few other bonds are so-called annuity bonds. The total payment is the same for all payment dates. Each payment is the sum of an interest payment and a partial repayment of the loan amount. The outstanding debt and the interest payment are gradually decreasing over the life of an annuity, so that the repayment increases over time. Let again $q > 0$ denote the periodic coupon rate and F_0 the initial face value. Then the constant periodic payment is

$$M_i = M = \frac{F_0}{A(q,n)}, \quad i = 1, \dots, n. \quad (5.2)$$

where

$$A(q,n) = \sum_{j=1}^n (1+q)^{-j} = \frac{1 - (1+q)^{-n}}{q} \quad (5.3)$$

is the so-called *annuity factor*, which equals the present value of a periodic payment of 1 dollar over n periods when a discount rate of q is applied. Here, the last equality applies the mathematical result

$$\sum_{j=1}^n \alpha^j = \alpha + \alpha^2 + \dots + \alpha^n = \frac{\alpha}{1-\alpha} (1 - \alpha^n), \quad \alpha \neq 1,$$

with $\alpha = (1+q)^{-1} = \frac{1}{1+q}$ so that $\frac{\alpha}{1-\alpha} = \frac{1}{q}$.

It can be shown that the outstanding debt of the annuity immediately after the i 'th payment is

$$F_i = MA(q,n-i) = M \frac{1 - (1+q)^{-(n-i)}}{q}.$$

The interest part and the repayment part of the i 'th payment are then

$$I_i = qF_{i-1} = qF_0 \frac{1 - (1+q)^{-(n-i+1)}}{1 - (1+q)^{-n}},$$

$$X_i = M - I_i = M - qF_{i-1} = M - M \left(1 - (1+q)^{-(n-(i-1))} \right) = M (1+q)^{-(n-i+1)},$$

which ensures that $X_i + I_i = M$.

Serial bonds. A serial bond pays back the face value in equal instalments. The payment at a given payment date is then the sum of the instalment and the interest rate on the outstanding debt. The interest rate payments, and hence the total payments, will therefore decrease over the life of the bond. With an initial face value of F_0 , each instalment or repayment is $X_i = F_0/n$ for any $i = 1, \dots, n$. Immediately after the i 'th payment date, the outstanding debt must be $F_i = F_0(n-i)/n = F_0[1 - (i/n)]$. The interest payment at

i	Bullet bond				Annuity bond				Serial bond			
	I_i	X_i	M_i	F_i	I_i	X_i	M_i	F_i	I_i	X_i	M_i	F_i
1	6.00	0.00	6.00	100.00	6.00	7.59	13.59	92.41	6.00	10.00	16.00	90.00
2	6.00	0.00	6.00	100.00	5.54	8.04	13.59	84.37	5.40	10.00	15.40	80.00
3	6.00	0.00	6.00	100.00	5.06	8.52	13.59	75.85	4.80	10.00	14.80	70.00
4	6.00	0.00	6.00	100.00	4.55	9.04	13.59	66.81	4.20	10.00	14.20	60.00
5	6.00	0.00	6.00	100.00	4.01	9.58	13.59	57.23	3.60	10.00	13.60	50.00
6	6.00	0.00	6.00	100.00	3.43	10.15	13.59	47.08	3.00	10.00	13.00	40.00
7	6.00	0.00	6.00	100.00	2.82	10.76	13.59	36.32	2.40	10.00	12.40	30.00
8	6.00	0.00	6.00	100.00	2.18	11.41	13.59	24.91	1.80	10.00	11.80	20.00
9	6.00	0.00	6.00	100.00	1.49	12.09	13.59	12.82	1.20	10.00	11.20	10.00
10	6.00	100.00	106.00	0.00	0.77	12.82	13.59	0.00	0.60	10.00	10.60	0.00

Table 5.1: Bond payment schedules.

The table shows the payment schedule of a bullet bond, an annuity bond, and a serial bond. All bonds have $F_0 = 100$, $q = 0.06$, and $n = 10$.

date i is therefore $I_i = qF_{i-1} = qF_0(1 - [(i-1)/n])$. Consequently, the total payment is

$$M_i = X_i + I_i = \frac{F_0}{n} + qF_0 \left(1 - \frac{i-1}{n}\right).$$

Example 5.1

Consider a bond having an initial face value of $F_0 = 100$, periodic coupon rate $q = 6\% = 0.06$, and maturing in 10 periods. The payment schedule of the bond depends on the amortization principle. Table 5.1 shows the payment schedule for a bullet bond, an annuity bond, and a serial bond. Figure 5.1 illustrates how the total payments of each bond are divided into an interest payment and a repayment of debt.

In addition to the three above-listed bond types, so-called **perpetuities** (also known as perpetual bonds or consols) are also traded, although they are quite rare. These bonds never mature but last forever and only pay interest, i.e., $M_i = qF_0$, $i = 1, 2, \dots$. The face value of a perpetuity is never repaid. Some of these bonds are callable (see below) so that they may eventually cease to exist.

Most coupon bonds have a fixed coupon rate, but for some bonds the coupon rate is reset periodically over the life of the bond. Such bonds are called **floating rate bonds**. Typically, the coupon rate effective for the payment at the end of one period is set at the beginning of the period at the current market interest rate for that period, e.g., to the 6-month interest rate for a floating rate bond with semi-annual payments.

Some governments issue bonds having a face value which is adjusted during the life of the bond by the inflation rate. In the U.S., these **inflation-indexed bonds** are often called TIPS which is short for Treasury Inflation-Protected Securities. While a standard non-indexed bond promises certain dollar payments in the future, an inflation-indexed bond really promises payments with a certain future purchasing power.

Some bonds have embedded options so that they may end up having a payment stream different from the scheduled one. For example, for a **callable bond** the issuer has the right to call the bond, i.e., to buy back the bond which amounts to prepaying the remaining debt. The price the issuer has to pay is typically the outstanding debt at the time plus a

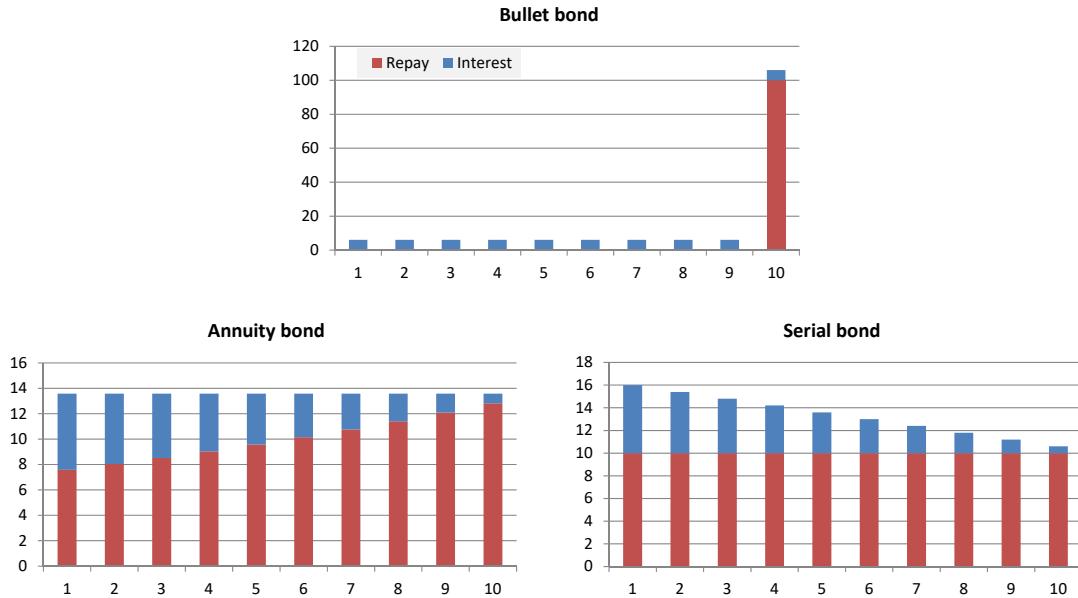


Figure 5.1: Bond cash flows.

The diagrams show the payment schedule of a bullet bond, an annuity bond, and a serial bond. All bonds have $F_0 = 100$, $q = 0.06$, and $n = 10$.

premium of a few percent, but the exact call price structure is pre-specified in the bond contract. Many mortgage-backed bonds and corporate bonds are callable.

Note that while the bond issuer promises a given payment stream, there is a risk that the issuer defaults and cannot make the promised payments. Any payments to the bond holders at default or after default depend on the value of the issuer's assets. Because of the **default risk**, the cash flow from the bond is uncertain.

5.2 Bond prices

The price of a bond reflects the present value of its future payments. To begin with, assume that we are currently at time 0 and the bond will make sure payments of M_1 at time 1 (after one period), M_2 at time 2 (after two periods), and so on until a final payment of M_n at time n (after n periods). Given a constant discount rate of r per period, the theoretical price of the bond is then

$$B_0 = \sum_{i=1}^n M_i (1+r)^{-i}. \quad (5.4)$$

Notice that the first- and second-order derivatives with respect to the discount rate are

$$\begin{aligned} \frac{\partial B_0}{\partial r} &= - \sum_{i=1}^n i M_i (1+r)^{-i-1} < 0, \\ \frac{\partial^2 B_0}{\partial r^2} &= \sum_{i=1}^n i(i+1) M_i (1+r)^{-i-2} > 0, \end{aligned}$$

so that the bond price is a decreasing, convex function of the discount rate.

For the standard bond types the sum in the above pricing equation can be expressed as a relatively simple function of F_0 , q , r , and n as summarized in the following theorem. Note that, in Excel, the built-in function PRICE can be used to compute prices of bullet bonds.

Theorem 5.1

Suppose that the discount rate of r per period applies to all payment dates of the bond.

- (a) For a zero-coupon bond maturing at time n with a face value of F , the price at time $t < n$ is

$$Z_{t,n} = F(1+r)^{-(n-t)}. \quad (5.5)$$

- (b) Let time 0 denote a time point with a full period until the next payment date and n full periods until maturity. Let F_0 denote the face value and q the coupon rate. Then the time 0 prices of different coupon bonds are

$$\text{Bullet bond: } B_0 = F_0 \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) (1+r)^{-n} \right), \quad (5.6)$$

$$\text{Perpetuity: } B_0 = F_0 \frac{q}{r}, \quad (\text{provided } r > 0) \quad (5.7)$$

$$\text{Annuity bond: } B_0 = F_0 \frac{q}{r} \frac{1 - (1+r)^{-n}}{1 - (1+q)^{-n}}, \quad (5.8)$$

$$\text{Serial bond: } B_0 = F_0 \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) \frac{A(r,n)}{n} \right). \quad (5.9)$$

- (c) If there is only a fraction $1 - t \in (0, 1)$ of a period until the next payment date, then the coupon bond price is given by

$$B_t = (1+r)^t B_0, \quad (5.10)$$

where the above formulas for B_0 can be used for standard coupon bond types.

Proof

- (a) The formula for the zero-coupon bond is straightforward.
(b) For the bullet bond we have $M_i = qF_0$ for $i = 1, 2, \dots, n-1$ and $M_n = (1+q)F$, so the price is

$$\begin{aligned} B_0 &= \sum_{i=1}^n qF_0(1+r)^{-i} + F_0(1+r)^{-n} \\ &= F_0 \left(qA(r,n) + (1+r)^{-n} \right) = F_0 \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) (1+r)^{-n} \right). \end{aligned}$$

The formula for the perpetuity follows by letting $n \rightarrow \infty$ in the formula for the bullet bond price since $(1+r)^{-n} \rightarrow 0$ if $r > 0$.

For the annuity bond all payments equal $M = F_0/A(q,n)$ so the price is

$$\begin{aligned} B_0 &= \sum_{i=1}^n M(1+r)^{-i} = M \sum_{i=1}^n (1+r)^{-i} \\ &= MA(r,n) = F_0 \frac{A(r,n)}{A(q,n)} = F_0 \frac{q}{r} \frac{1 - (1+r)^{-n}}{1 - (1+q)^{-n}}. \end{aligned}$$

A serial bond with n remaining payments can be seen as a portfolio of n bullet bonds each having a face value of F_0/n . The first bullet bond matures after one period, the second after two periods, etc. From (5.6), the price of the bullet bond maturing after j periods is

$$B_0^{(j)} = \frac{F_0}{n} \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) (1+r)^{-j} \right).$$

Hence, the price of the serial bond must be

$$\begin{aligned} B_0 &= \sum_{j=1}^n B_0^{(j)} = \sum_{j=1}^n \frac{F_0}{n} \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) (1+r)^{-j} \right) \\ &= \frac{F_0}{n} \left(\sum_{j=1}^n \frac{q}{r} + \left(1 - \frac{q}{r}\right) \sum_{j=1}^n (1+r)^{-j} \right) \\ &= \frac{F_0}{n} \left(n \frac{q}{r} + \left(1 - \frac{q}{r}\right) A(r,n) \right) = F_0 \left(\frac{q}{r} + \left(1 - \frac{q}{r}\right) \frac{A(r,n)}{n} \right). \end{aligned}$$

(c) Note that for any coupon bond, the face value F_t at time t is identical to the face value F_0 immediately after time 0 as no payments have been made since then. Hence, the scheduled future payments have not changed since time 0. Eq. (5.10) now follows from the calculation

$$B_t = \sum_{i=1}^n M_i (1+r)^{-(i-t)} = (1+r)^t \sum_{i=1}^n M_i (1+r)^{-i} = (1+r)^t B_0.$$

The formula (5.10) for the bond price at a general point in time is based on the idea to first discount all payments back to the most recent past payment date and then discount forward to the current date.

As long as $r > 0$, zero-coupon bonds trade at a discount relative to the face value, i.e., the price is lower than the face value. Also note that when $r > 0$, the zero-coupon bond price is a decreasing function of the time-to-maturity,

$$m < n \Rightarrow Z_{t,m} > Z_{t,n}. \quad (5.11)$$

This relation simply reflects the fact the investors would prefer to receive the face value as soon as possible and are therefore willing to pay more for the zero-coupon bond with the shorter maturity.

Figure 5.2 shows how the bond price depends on the discount rate for the three 10-year bonds considered in Example 5.1. The graphs confirm that the bond price is a decreasing, convex function of the discount rate. For all three bond types, the price equals the face

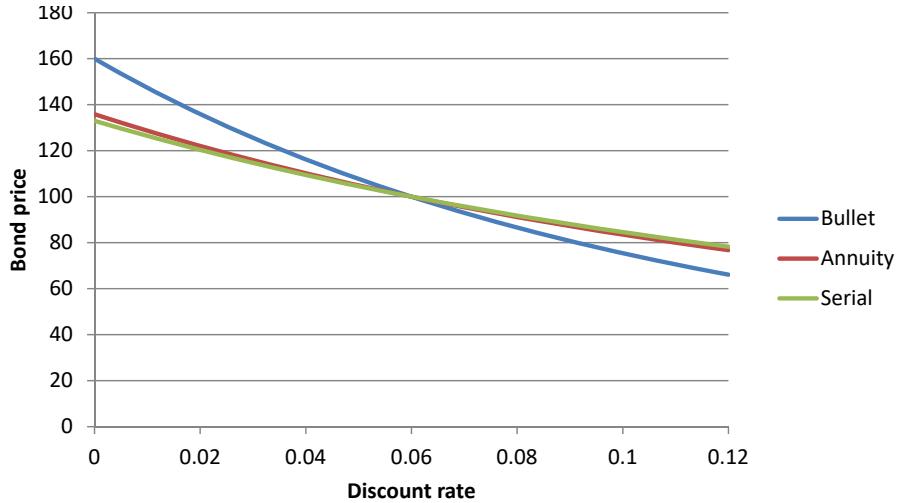


Figure 5.2: Bond prices and discount rates.

The graphs show the price as a function of the discount rate for a bullet bond, an annuity bond, and a serial bond. All bonds have $F_0 = 100$, $q = 0.06$, and $n = 10$.

value when the discount rate equals the coupon rate. The price of the bullet bond is much more sensitive to the discount rate than the annuity and the serial bond. This is natural because the present value of payments far into the future is more sensitive to the discount rate than payments in the near future, and by far the largest payment of the bullet bond is at the maturity date in 10 years. The payments of the annuity bond and the serial bond are more evenly spread out over time and thus includes some sizable payments in the near future as well. Section 5.8 has more on the interest rate sensitivity of bond prices.

Note that for all the standard bond types, the following holds:

1. the price equals the face value (“the bond trades at par”) if the discount rate equals the coupon rate: $r = q \Rightarrow B_0 = F_0$;
2. the price exceeds the face value (“the bond trades at a premium”) if the discount rate is smaller than the coupon rate: $r < q \Rightarrow B_0 > F_0$;
3. the price is below the face value (“the bond trades at a discount”) if the discount rate exceeds the coupon rate: $r > q \Rightarrow B_0 < F_0$.

The interest part of the next bond payment equals the product of the coupon rate and the current outstanding debt, which is meant as a compensation for lending the bond issuer the money over the period that has passed since the previous payment date. If you purchase a bond between two payment dates, say at time $t \in (0,1)$ as above, and hold on to it until the next payment date 1, you have only acted as the lender over the fraction $1-t$ of the period. Nevertheless, you will receive the entire interest payment qF_0 at time 1. The seller of the bond will not receive his fair share of this interest payment, which is tqF_0 . Therefore, when you purchase the bond at time t , you will have to pay $Q_t^{\text{acc}} = tqF_0$ to the seller on top of the listed bond price. This is the so-called **accrued interest**. If we let B_t^{list} denote the listed bond price at time t , the total or “true” purchase price is the sum $B_t^{\text{list}} + Q_t^{\text{acc}}$. Since the total purchase price must equal the present value of all future payments, the theoretical listed bond price at time $t \in (0,1)$ must be

$$B_t^{\text{list}} = B_t - Q_t^{\text{acc}}, \quad (5.12)$$

where B_t is given in Theorem 5.1.¹ At time 0—that is at any payment date—the accrued interest is zero so

$$B_0^{\text{list}} = B_0. \quad (5.13)$$

In most bond markets, bond prices are quoted and listed as a percentage of the face value and therefore tend to be in the neighborhood of 100. Long before maturity the listed price of a bond can deviate substantially from 100 (i.e., 100 percent of the face value or outstanding debt) if the coupon rate is very different from the appropriate discount rate. But as time passes and the maturity date of the bond comes close, the listed bond price will approach 100. First think of a zero-coupon bond. If T denotes the remaining time-to-maturity, the present value of the face value is $F(1+r)^{-T}$ which will approach F (from below) as $T \rightarrow 0$. Secondly, when a coupon bond only has one payment date left, it is effectively a zero-coupon bond. For example, a bullet bond with face value F and coupon rate q that has a single payment left is equivalent to a zero-coupon bond with a face value of $(1+q)F$. The present value of this future payment is $(1+q)F(1+r)^{-T}$ and as the maturity date approaches, the present value reaches $(1+q)F$. Since the accrued interest approaches qF when the maturity date is near, the listed bond price goes to $(1+q)F - qF = F$. The same argument works for an annuity bond or a serial bond if we let F denote the outstanding debt before the final payment date.

Figure 5.3 illustrates how the price of a 6% bullet bond with annual payments changes over time using a fixed discount rate of either 4% or 8%. The path of the total price (dotted curves) has a zigzag pattern with an annual frequency. This is due to the fact that when the next coupon payment date is approaching, the present value of the future payments naturally increases. If you buy the bond immediately before the coupon payment, you get all the same payments as if you buy it immediately after *plus* the immediate coupon payment. Therefore, just around a coupon payment date, the present value drops by an amount equal to the coupon payment. The listed price (solid curves) controls for this mechanical effect by subtracting the accrued interest and is thus smooth. For both discount rates the listed price converges to 100 as the maturity date approaches.

In reality, the appropriate discount rate for a bond does not stay constant over the life of the bond, but can change significantly as discussed in subsequent sections. Therefore, the realized path of even the listed price of a bond will be much less smooth than illustrated in Figure 5.3. However, the listed bond price will nevertheless approach the face value as the maturity date comes close, because the applied discount rate becomes unimportant as the discounting period goes to zero.

5.3 Bond yields and yield curves

5.3.1 Definition of the yield of a bond

The yield-to-maturity or just **yield** of a bond is the discount rate which ensures that the present value of the future payments discounted at that rate is equal to the current market price of the bond. We denote the yield of a bond by y . Immediately after a payment date and thus with a full period until the next payment, the yield of a bond is such that the equation

$$B^{\text{mkt}} = \sum_{i=1}^n M_i(1+y)^{-i} \quad (5.14)$$

¹The listed price is sometimes referred to as the *clean* price and the true price is sometimes called the *dirty* price.

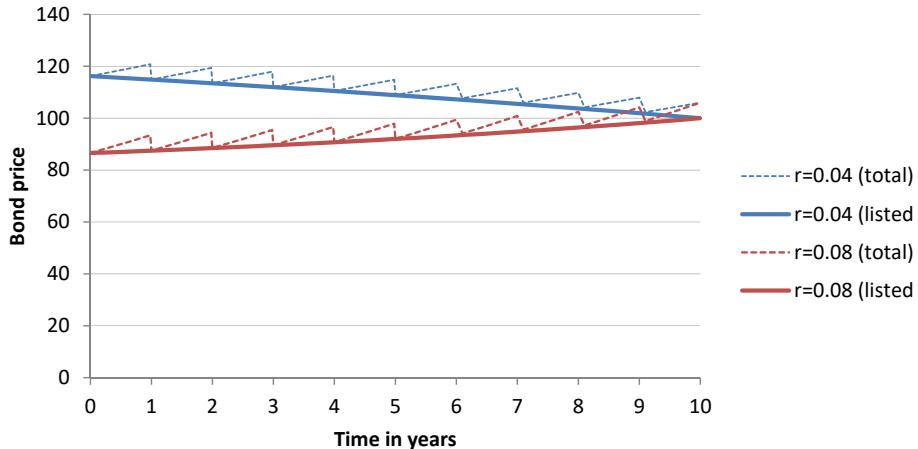


Figure 5.3: Bond price paths for a fixed discount rate.

The graph shows how the price of a bullet bond will change over time until its maturity date assuming that the appropriate discount rate remains unchanged. The bond has face value $F_0 = 100$, coupon rate $q = 6\%$, one annual payment date, and matures after 10 years. The blue curves are based on a discount rate of $r = 4\%$, the red curves on a discount rate of $r = 8\%$. The dotted jagged curves show the total price of the bond, whereas the solid smooth curves show the listed price, i.e., the total price less the accrued interest.

holds, where B^{mkt} is the market price of the bond. If we stand at time $t \in (0,1)$ with the fraction $1 - t$ of a period until the next payment date and B_t^{mkt} is the current market price of the bond (including any accrued interest), the yield is such that

$$B_t^{\text{mkt}} = \sum_{i=1}^n M_i (1+y)^{-(i-t)} = (1+y)^t B_0(y),$$

where $B_0(y)$ is the time 0 price using the discount rate y . Since $B_0(y)$ is an n 'th order polynomial, the yield y can generally only be found by a more or less advanced trial-and-error approach or by using the Solver in Excel. In Excel, you can alternatively use the built-in function YIELD to compute the yield of a bullet bond.

For zero-coupon bonds the yield is easy to calculate. For a zero-coupon bond with a face value of F that matures in n periods, the market price and the yield are related via

$$Z_{0,n}^{\text{mkt}} = F(1+y_n)^{-n} \Leftrightarrow y_n = \left(\frac{F}{Z_{0,n}^{\text{mkt}}} \right)^{1/n} - 1. \quad (5.15)$$

Here the subscript n on the yield indicates the maturity of the zero-coupon bond.

Note that there is a one-to-one relation between the price and the yield of a bond, so the two quantities carry the same information. Often it is easier to relate to the magnitude of the yield than the bond price, especially if you want to compare bonds of different maturities. If we measure time in years when applying the above formulas (so that n represents n years), the yield in those formulas is an effective, annualized yield. Sometimes continuously compounded yields are used instead. The continuously compounded yield y^c

on a coupon bond is implicitly defined by the relation

$$B_t^{\text{mkt}} = \sum_{i=1}^n M_i e^{-y^c(i-t)}, \quad (5.16)$$

and the continuously compounded zero-coupon yield y_n^c of an n -year bond is defined by

$$Z_{0,n}^{\text{mkt}} = F e^{-y_n^c \times n} \Leftrightarrow y_n^c = \frac{1}{n} \ln \left(\frac{F}{Z_{0,n}^{\text{mkt}}} \right), \quad (5.17)$$

cf. the discussion of continuously compounded returns in Section 2.2. The two yield measures are related through

$$y_n = e^{y_n^c} - 1 \Leftrightarrow y_n^c = \ln(1 + y_n). \quad (5.18)$$

For low levels of the yields, the two measures are very close, but they differ more for larger levels. For example, an effective annual yield of $y_n = 0.02 = 2\%$ corresponds to a continuously compounded yield of $y_n^c = \ln(1.02) \approx 0.01980 = 1.980\%$. But an effective annual yield of $y_n = 20\%$ corresponds to a continuously compounded yield of $y_n^c \approx 18.232\%$, which is significantly lower. In any case, if you want to compare different yields, they should be computed in the same way. Unless otherwise mentioned, this book applies the effective annual yields.

From the relations between bond prices and discount rates derived in the preceding section, we know that the bond price is a decreasing, convex function of the yield. Moreover, when the yield equals the coupon rate, the market bond price equals the face value. When the yield exceeds the coupon rate, the market price is below the face value. When the yield is smaller than the coupon rate, the market price exceeds the face value.

5.3.2 Yields versus returns

The yield can be interpreted as the average rate of return to an investor who holds the bond until maturity. An investor reselling the bond before maturity will experience a holding-period return equal to the yield at the time of purchase provided that the yield of the bond is the same at the selling date as it was at purchase. Assume, for example, that you buy a bond at time 0 and sell it again at time 1 immediately after the next payment of the bond. The purchase price is $B_0 = \sum_{i=1}^n M_i (1 + y_0)^{-i}$ so that y_0 is the bond's yield at purchase. The selling price at time 1 is $B_1 = \sum_{i=2}^n M_i (1 + y_1)^{-(i-1)}$, where y_1 is the bond's yield when selling. If we add the payment M_1 received at time 1, we get

$$M_1 + B_1 = M_1 + \sum_{i=2}^n M_i (1 + y_1)^{-(i-1)} = \sum_{i=1}^n M_i (1 + y_1)^{-(i-1)} = (1 + y_1) \sum_{i=1}^n M_i (1 + y_1)^{-i}.$$

Therefore, the rate of return over the period is

$$r = \frac{M_1 + B_1}{B_0} - 1 = (1 + y_1) \frac{\sum_{i=1}^n M_i (1 + y_1)^{-i}}{\sum_{i=1}^n M_i (1 + y_0)^{-i}} - 1.$$

In particular, if the yield is unchanged meaning $y_1 = y_0$, the ratio in the last expression equals one, and hence the return is simply $r = (1 + y_1) - 1 = y_1 = y_0$ and thus equal to the yield at purchase. If the yield decreases over the holding period, the selling price is

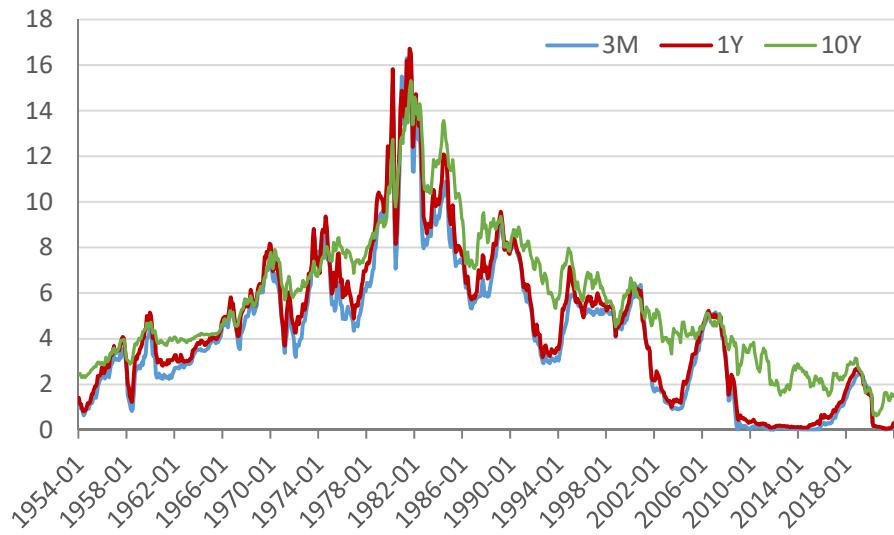


Figure 5.4: Time series of yields.

The graphs show yields of nominal U.S. Treasury bills and bonds of maturities of three months, one year, and ten years from January 1954 to December 2021. The yields are in percent. Source: <http://www.federalreserve.gov/releases/h15/data.htm>, data retrieved on July 4, 2022. 3-month yields from January 1954 to August 1981 are estimates downloaded on September 7, 2015, but they are apparently not included in the data series that could be downloaded in July 2022.

larger and thus the return is higher than in the case of an unchanged yield, and therefore the return is higher than the yield at purchase. Conversely, if the yield increases over the holding period, the return is lower than the yield at purchase.

If bond yields remain constant, the yield level shows the rate of return a bond investor obtains. If the yield level moves over a certain period of time, say, from a high level to a low level, also bond returns will eventually move to the lower level. However, in the transition period where the yield level is declining, bond prices increase so the bond investor realizes high returns as the above argument shows. Conversely, if yields move from one level to a higher level, bond investors will temporarily realize low returns, but eventually higher returns as yields stabilize at the new, higher level.

5.3.3 Yield curves

Figure 5.4 shows yields on U.S. Treasury bills and bonds of maturities of three months, one year, and ten years over the period from January 1954 to December 2021. The bonds are bullet bonds with semi-annual payments. Clearly, the yield for a given time-to-maturity varies over time and yields for different maturities are not the same. Most of the time, the three-month yield is less than the one-year yield which again is less than the ten-year yield. Note the high variability of the level of interest rates over the full period, whereas over shorter periods yields are quite persistent. Furthermore, short-maturity yields are more volatile than long-maturity yields.

The **yield curve** at a given point in time shows how the yields of bonds depend on the time-to-maturity. The yield curve is often referred to as the **term structure of interest rates**. The three yields illustrated in Figure 5.4 suggest that the yield curve is most often increasing: higher yields on longer-maturity bonds. This is also reflected by

Figure 5.5 where the solid curves show the yield curve from January in the years 1985, 1990, 1995, 2000, 2005, 2010, 2015, and 2020. Each yield curve is based on the prices of U.S. Treasury bonds of various maturities. The yield curves from January 1990 and 2020 are almost flat, but the other yield curves are upward-sloping at least up to maturities of 5-10 years after which the curves are relatively flat. However, in some rather short time periods, short-maturity yields have been higher than long-maturity yields indicating a downward-sloping or *inverted* yield curve. Sometimes the yield curve is non-monotonic and may exhibit a “hump” (first increasing to a maximum, then decreasing) or a “trough” (first decreasing to a minimum, then increasing) or have some even more complex shape. The two dashed curves in Figure 5.5, taken from November 2000 and March 2007, are examples of non-increasing yield curves.

Before economic expansions the yield curve tends to be steeply upward-sloping, whereas it is often flat or even downward-sloping before recessions, cf. [Chen \(1991\)](#) and [Estrella and Hardouvelis \(1991\)](#). In other words, the slope of the yield curve forecasts economic growth.

When the yield curve is upward-sloping, it is typically concave, that is relatively steep for short maturities and almost flat for long maturities. The yield curves in Figure 5.5 are consistent with this observation. [Campbell \(2000\)](#) reports that the average historical yield difference (spread) to the one-month yield is 33 basis points (0.33 percentage points) for the three-month yield, 77 basis points for the one-year yield, and 96 basis points for the two-year yield, whereas there is only little difference between the average two-year and 10-year yields. In periods where the yield curve is downward-sloping, it is typically convex, that is steeply decreasing for short maturities and almost flat for long maturities.

Two bonds of the same maturity can have different yields if they have different coupon rates or different payment schedules. For example, a 10-year annuity bond may have a different yield than a 10-year bullet bond. The yield reflects how the market values the entire payment stream. If bond investors are willing to pay a higher price for the 10-year annuity bond than the 10-year bullet bond, the annuity bond will have a smaller yield than the bullet bond. The default risk of the bond issuer also affects the bond price and therefore the yield. If two bonds promise identical future payment streams, the bond with the lowest default risk has the highest price and therefore the lowest yield of the two. When forming a yield curve, the bonds used should ideally only differ with respect to their time-to-maturity, and at least they should be similar in terms of amortization principle, the default risk of the issuer, etc.

The yield of a coupon bond maturing in n periods is a measure of the interest rate which the market thinks is fair on a loan over n periods paid back according to the promised payment schedule. The yield y_n of a zero-coupon bond maturing in n periods gives a direct measure of the fair interest rate to use when discounting payments to be received in n periods back to today. It is not “polluted” by any intermediate interest payments or partial repayments of the face value. Hence, market participants and analysts often focus on the **zero-coupon yield curve**, i.e., the relation at a given point in time between yields on zero-coupon bonds and their time-to-maturity. In most bond markets, few zero-coupon bonds are traded so most of these zero-coupon yields have to be derived or estimated from prices of coupon bonds. This is based on the relation between the prices of coupon bonds and prices of zero-coupon bonds discussed in the next section.

5.3.4 Real yields and interest rates

In some countries, governments issue inflation-protected bonds, where the face value in the relevant currency is regularly adjusted by the realized inflation rate. Therefore such

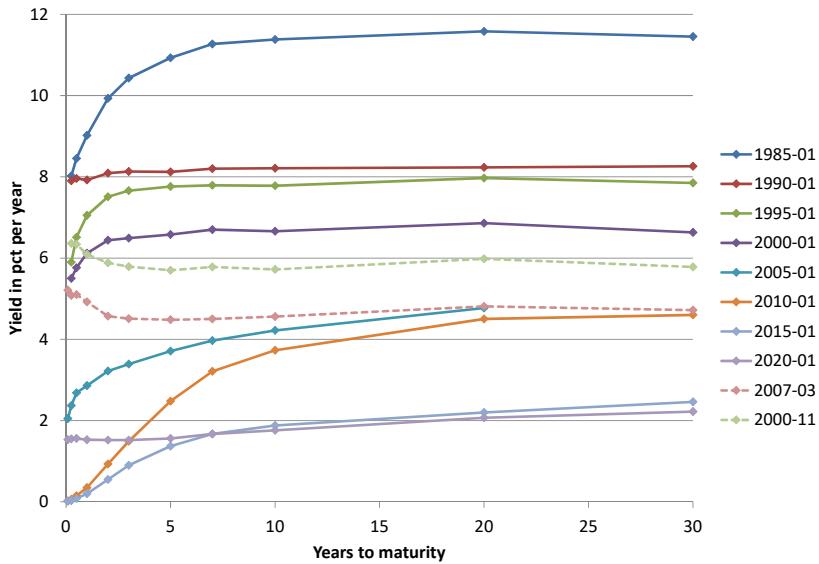


Figure 5.5: Historical yield curves.

The solid curves are yield curves determined from nominal U.S. Treasury bonds in January of 1985, 1990, 1995, 2000, 2005, 2010, 2015, and 2020. The dashed curves are for November 2000 and March 2007. The curves are drawn by connecting reported yields for maturities of 1, 3, and 6 months as well as 1, 2, 3, 5, 7, 10, 20, and 30 years. For 2000 and earlier, there is no data for the one-month yield. For the 1990 curve the 20-year yield is obtained by interpolating the reported 10-year and 30-year yields. For the 2005 curve there is no data for the 30-year yield. Source: <http://www.federalreserve.gov/releases/h15/data.htm>, data retrieved on July 4, 2022.

bonds deliver a certain future purchasing power to investors. If, for example, the yield of such a bond is 2% when you purchase it, and you keep it until maturity, you will experience a *real* rate of return equal to 2% per year. We refer to the yield of an inflation-indexed bond of a given maturity as the *real yield* for that maturity.

In 1997, the U.S. Treasury introduced the so-called TIPS (Treasury Inflation-Protected Securities) which are inflation-indexed bullet bonds. The Treasury regularly issues new series of these bonds of a few selected maturities up to 30 years. The left panel of Figure 5.6 illustrates how the real yields for certain maturities have varied from January 2003 to July 2021. At almost all dates, we see that the real yield is increasing in the maturity of the bond with exceptions, e.g., in late 2008. The real yields were negative for all maturities in large parts of 2020 and 2021 and for maturities of 5 or 10 years in most of the period 2010-2013.

In the United Kingdom, government bonds are often called gilts, short for gilt-edged securities. Index-linked gilts of various maturities have been issued since 1981. The right panel of Figure 5.6 shows the variation in the yields for selected maturities from January 1985 to July 2021. Again, the real yield is increasing in maturity most of the time but with exceptions, e.g., in 1998-2001 and 2007-2009. The real U.K. yields were all positive until mid-2009, where the shorter-maturity yields turned negative, and from March 2014 to July 2021 real yields of all maturities have remained negative.

Provided that the issuing government never defaults, the yield of an index-linked government bond is reflecting a riskfree real return until the maturity of the bond. Basically, an index-linked government bond is as close as you can get to an asset which is riskfree in

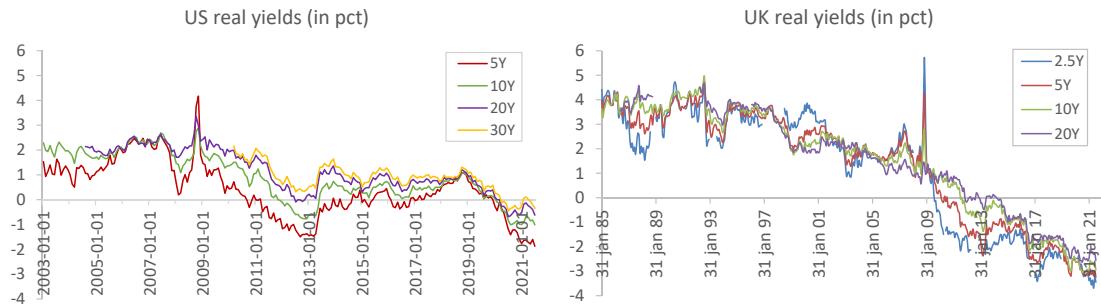


Figure 5.6: Historical real yields.

The figure shows the variation over time in real yields of inflation-indexed U.S. and U.K. government bonds of selected maturities. The U.S. data were downloaded from the homepage of the St. Louis Fed. The U.K. data were downloaded from the homepage of the Bank of England. Download date was August 12, 2021.

real terms. As the indexation follows the official inflation rates in the country of the issuer, the yield of an inflation-indexed bond might not be completely riskfree in real terms to you, if your consumption basket differs from the basket underlying the published consumer price index, but we ignore any such discrepancy here.

If inflation-indexed bonds are not available for the country, time period, or maturity you are interested in, how can you then estimate a real riskfree rate? Of course, we can calculate the realized real return on a nominal (i.e., not inflation-indexed) government bond at the end of a given holding period, as explained in Section 2.6. However, investors would often like to have an estimate of the real return you can obtain on a bond over a future period. The real interest rate for a given maturity can then be estimated by taking the yield of a nominal government bond of that maturity and subtracting the expected inflation until the maturity of the bond. You should generally also divide the difference by one plus the expected inflation rate in line with Eq. (2.15). As the realized inflation might differ from the expected inflation, the real rate estimated in this way is not a truly riskfree rate.

Numerous researchers have produced estimates of real interest rates going back in time. For example, based on various historical records, Schmelzing (2020) provides an estimate of a global real interest rate back to 1311, which is illustrated in Figure 5.7. The real interest rate has a clear downward trend of roughly one percentage point per century, but with substantial variation around the trend.

5.4 Price relations across bonds

Prices of bonds are generally much more closely related than prices of stocks. If we disregard any uncertainty about the future payments, an investment in a bond is simply a way to move money (and thus capital for consumption or other investments) from today to later. Unlike a stock investor, the bond investor is not making a bet on the financial performance of the issuer over the life of the bond, at least not if we assume the issuer in any case will make the promised payments to the bond holders. Conversely, the issuer of the bond simply moves money (capital) from later to today. The prices of default-free, government bonds are therefore driven by the investor's desire to postpone consumption opportunities (determining the demand for bonds) and the government's funding demand (determining the supply of bonds). Both factors depend on the current and the expected

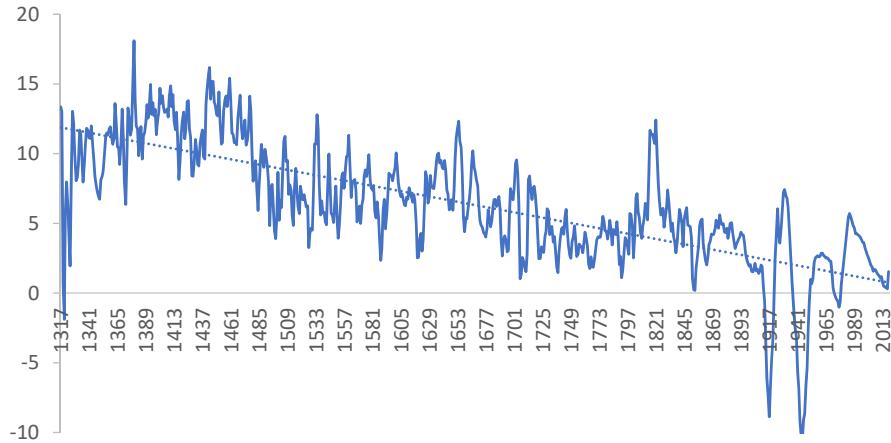


Figure 5.7: Real rates over 708 years.

The figure shows an estimate of a global real interest rate from 1311 to 2018. The graph was constructed from the data made available by Paul Schmelzing and the Bank of England at <https://www.bankofengland.co.uk/working-paper/2020/eight-centuries-of-global-real-interest-rates-r-g-and-the-suprasecular-decline-1311-2018>. Download date was August 12, 2021. See the paper by Schmelzing (2020) for a detailed explanation of how the global real rates are determined.

future macroeconomic conditions which therefore drive the prices of all government bonds. There is generally no bond-specific risk, at least not for government bonds, in contrast to stocks that typically have a significant stock-specific or unsystematic return component.

In fact, some bond prices must be related in a very specific way, otherwise there would be an arbitrage opportunity, cf. the definition and discussion in Section 4.5. Consider a coupon bond that pays M_1 at time 1 (in one period from now), M_2 at time 2, etc., until a final payment of M_n at time n . We can see this coupon bond as a portfolio of zero-coupon bonds, namely a portfolio of M_1 zero-coupon bonds maturing at time 1 and having a face value of 1, M_2 zero-coupon bonds maturing at time 2 and having a face value of 1, etc. If all these zero-coupon bonds are traded in the market, the price of the coupon bond at any time t must be

$$B_0 = \sum_{i=1}^n M_i Z_{0,i}. \quad (5.19)$$

If this relation does not hold, there is a clear arbitrage opportunity in the market. If $B_0 > \sum_{i=1}^n M_i Z_{0,i}$, an arbitrage profit can be locked in by selling the coupon bond and buying the portfolio of zero-coupon bonds. This leads to an immediate profit of $B_0 - \sum_{i=1}^n M_i Z_{0,i}$, whereas all net future payments are zero. Conversely if $B_0 < \sum_{i=1}^n M_i Z_{0,i}$. When traders start exploiting the arbitrage by implementing the appropriate strategy, prices on both the coupon bond and the zero-coupon bonds change until the arbitrage opportunity has disappeared. Note that the relevant arbitrage strategy involves selling bonds. If you do not already own these bonds, you would have to short-sell them. While some investors may not be allowed to do so, there are investors for which it is possible so they can surely implement the arbitrage strategy.

Example 5.2

Consider a bullet bond with a face value of 100 dollars, a coupon rate of 7%, annual payments, and exactly three years to maturity. Suppose zero-coupon bonds are traded with face values of 1 dollar and time-to-maturity of 1, 2, and 3 years, respectively. Assume that the prices of these zero-coupon bonds are $Z_{0,1} = 0.94$, $Z_{0,2} = 0.90$, and $Z_{0,3} = 0.87$. According to (5.19), the price of the bullet bond must then be

$$B_0 = 7 \times 0.94 + 7 \times 0.90 + 107 \times 0.87 = 105.97.$$

If the price is lower than 105.97, riskfree profits can be locked in by buying the bullet bond and selling 7 one-year, 7 two-year, and 107 three-year zero-coupon bonds. If the price of the bullet bond is higher than 105.97, sell the bullet bond and buy 7 one-year, 7 two-year, and 107 three-year zero-coupon bonds.

Note that if we write the price of each bond appearing in (5.19) in terms of the bond's yield, we obtain the relation

$$\sum_{i=1}^n M_i(1+y)^{-i} = \sum_{i=1}^n M_i(1+y_i)^{-i}, \quad (5.20)$$

which shows that the yield y of the coupon bond is a complicated form of average of the zero-coupon yields y_1, y_2, \dots, y_n associated with the coupon bond's future payment dates. The larger the payment M_i , the larger the relative weight of the associated zero-coupon yield y_i in this average. For example, the yield of a 5-year bullet bond with annual payments is some average of the zero-coupon yields for maturities 1, 2, 3, 4, and 5 years with the largest weight on the 5-year zero-coupon yield as the final payment of the bullet bond clearly exceeds the other. Hence, the yield of an n -year bullet bond is typically close to, but not identical to, the n -year zero-coupon yield.

Maybe not all the zero-coupon bonds relevant for the replication are traded. In that case we cannot justify the relation (5.19) as a result of the no-arbitrage principle. Still it is a valuable relation. Suppose that an investor has determined (from private or macroeconomic information) a discount function $i \mapsto Z_{0,i}$ showing the value *she* attributes to payments at different future points in time. Then she can value all sure cash flows in a consistent way by substituting that discount function into (5.19).

There are also no-arbitrage price relations between some coupon bonds. Here is an example:

Example 5.3

Suppose that two annuity bonds both with n remaining periods and a face value of 100 are traded. The bonds have the same payment dates, but different coupon rates, say q_1 and q_2 . Then the constant periodic payments are

$$M_1 = \frac{100q_1}{1 - (1 + q_1)^{-n}}, \quad M_2 = \frac{100q_2}{1 - (1 + q_2)^{-n}},$$

respectively, cf. Eqs. (5.2) and (5.3). If, for example, $n = 10$, $q_1 = 5\%$, and $q_2 = 10\%$, then $M_1 \approx 12.950$ and $M_2 \approx 16.275$. By purchasing $16.275/12.950 \approx 1.25668$ units of the

5% bond, we get exactly the same payments as by purchasing one unit of the 10% bond. Hence, the price of the 10% bond must equal 1.25668 times the price of the 5% bond, otherwise an arbitrage opportunity exists.

Here is an example where the arbitrage strategy is more involved:

Example 5.4

Suppose three two-year bonds are traded. They have annual payments and thus two remaining payment dates. Bond 1 is an annuity bond paying 550 at both dates. Bond 2 is a 10% bullet bond with a face value of 1000 and thus a payment of 100 after one year and 1100 after two years. Bond 3 is a 4.5% bullet bond with a face value of 1000 and thus payments of 45 and 1045 after one and two years, respectively. Then the payments of bond 3 can be replicated perfectly by a portfolio of bonds 1 and 2. Let N_1 denote the number of units of bond 1 in the portfolio and N_2 the same for bond 2. To replicate bond 3, we must pick N_1 and N_2 so that

$$550N_1 + 100N_2 = 45, \quad 550N_1 + 1100N_2 = 1045.$$

The solution is

$$N_1 = -0.1, \quad N_2 = 1.$$

So, by purchasing one unit of bond 2 and shorting -0.1 units of bond 1, you will receive exactly the same payments as by purchasing one unit of bond 3. The prices B_1, B_2, B_3 of the bonds have to be related via

$$-0.1B_1 + B_2 = B_3,$$

otherwise there would be an arbitrage opportunity. Say the price of bond 1 is 1000 and the price of bond 2 is 1020. Then the price of bond 3 must equal $-0.1 \times 1000 + 1020 = 920$ to exclude arbitrage opportunities.

More generally, if you have a set of bonds with a total of n different payment dates, there are at most n degrees of freedom in their prices. Once you know the prices of n of the bonds, the prices of the other bonds follow from the no-arbitrage pricing principle.

As mentioned above few zero-coupon bonds are traded in most bond markets. However, by forming specific portfolios of traded coupon bonds, you can sometimes obtain the same payments as for a zero-coupon bond. In other words, zero-coupon bonds can be constructed from coupon bonds.

Example 5.5

In the above example with the three different two-year coupon bonds, we can construct a zero-coupon bond maturing in one year with a face value of 1000 by forming a portfolio of N_1 units of the annuity bond and N_2 units of the 10% bullet bond. We need to pick

N_1 and N_2 so that

$$550N_1 + 100N_2 = 1000, \quad 550N_1 + 1100N_2 = 0.$$

The solution is $N_1 = 2$ and $N_2 = -1$. With prices $B_1 = 1000$ and $B_2 = 1020$, this portfolio costs $2 \times 1000 - 1020 = 980$, which is then the implicit price of the one-year zero-coupon bond. Likewise, we can construct a two-year zero-coupon bond by picking N_1 and N_2 so that

$$550N_1 + 100N_2 = 0, \quad 550N_1 + 1100N_2 = 1000.$$

The solution is $N_1 = -100/550 \approx -0.1818$ and $N_2 = 1$. This portfolio costs $-0.1818 \times 1000 + 1020 = 838.18$, which is thus the price of the two-year zero-coupon bond implicit in the prices of the coupon bonds.

Many zero-coupon bonds can be constructed in a market with numerous coupon bonds with gradually increasing maturities. In fact, if coupon bonds maturing in $1, 2, \dots, T$ periods are all traded, we can sequentially construct zero-coupon bonds maturing in $1, 2, \dots, T$ periods. This procedure is often referred to as **bootstrapping** or **stripping** the yield curve. It is routinely applied in the huge market for U.S. Treasury bonds. The zero-coupon bond maturing in T periods can be constructed as a portfolio of all the coupon bonds maturing in $1, 2, \dots, T$ periods.

To see how the portfolio is constructed, let M_{it} denote the payment at time i of the coupon bond maturing at time t . Let N_t denote the number of units in the portfolio of the coupon bond maturing at time t . Then to match a zero-coupon bond maturing at time T with a face value of (say) 100, we need to choose N_1, N_2, \dots, N_T so that

$$\begin{aligned} N_1M_{11} + N_2M_{12} + N_3M_{13} + \cdots + N_{T-1}M_{1,T-1} + N_TM_{1T} &= 0, \\ N_2M_{22} + N_3M_{23} + \cdots + N_{T-1}M_{2,T-1} + N_TM_{2T} &= 0, \\ N_3M_{33} + \cdots + N_{T-1}M_{3,T-1} + N_TM_{3T} &= 0, \\ &\vdots \qquad \vdots \\ N_{T-1}M_{T-1,T-1} + N_TM_{T-1,T} &= 0, \\ N_TM_{TT} &= 100. \end{aligned}$$

We can write this system of equations in vector-matrix form as

$$\begin{pmatrix} M_{11} & M_{12} & M_{13} & \dots & M_{1,T-1} & M_{1T} \\ 0 & M_{22} & M_{23} & \dots & M_{2,T-1} & M_{2T} \\ 0 & 0 & M_{33} & \dots & M_{3,T-1} & M_{3T} \\ \vdots & & \ddots & & \vdots & \\ 0 & 0 & 0 & \dots & M_{T-1,T-1} & M_{T-1,T} \\ 0 & 0 & 0 & \dots & 0 & M_{TT} \end{pmatrix} \begin{pmatrix} N_1 \\ N_2 \\ N_3 \\ \vdots \\ N_{T-1} \\ N_T \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 100 \end{pmatrix}.$$

Here the left-hand sides reflect the payments from the portfolio at the various dates, and the right-hand sides show the payments of the zero-coupon bond that we want to match. We can solve this system of equations directly in Excel as explained in Section 4.2.3. Alternatively, we can solve the equations sequentially starting with the latter equation, which leads to N_T . Then we can substitute this into the preceding equation and solve for N_{T-1} and so forth. The price of each “synthetical” zero-coupon bond equals the price of

the replicating portfolio of coupon bonds.

More generally, suppose we have a collection of T bonds for which all payments are made at time points in the set $\{1, 2, \dots, T\}$, and we let M_{in} denote the payment of bond n at time i . Then we can generate any desired cash flow stream $\mathbf{P} = (P_1, P_2, \dots, P_T)^\top$ at these time points by finding a portfolio $\mathbf{N} = (N_1, N_2, \dots, N_T)^\top$ of the bonds so that the equation system

$$\underline{\underline{M}}\mathbf{N} = \mathbf{P}$$

holds. Solving for \mathbf{N} is possible provided that the $T \times T$ matrix $\underline{\underline{M}}$ of bond payments is invertible. Here, row i of the matrix consists of the payments M_{i1}, \dots, M_{iT} of the different bonds at time i . Column j contains the payments of bond j at different points in time.

Example 5.6

Suppose you can trade in three bonds, all having annual payments and exactly one year to the next payment date. Bond 1 is a two-year annuity bond with an annual payment of 55. Bond 2 is a two-year 10% bullet bond with face value 100, so it pays 10 after one year and 110 after two years. Bond 3 is a 5% bullet bond with face value 100, so it pays 5 after one year and after two years and 105 after three years. The three bonds form the invertible payment matrix

$$\underline{\underline{M}} = \begin{pmatrix} 55 & 10 & 5 \\ 55 & 110 & 5 \\ 0 & 0 & 105 \end{pmatrix}$$

Now we can find a portfolio \mathbf{N} of the three bonds that generates any desired stream of payments \mathbf{P} on the same dates. For example, we can replicate zero-coupon bonds for year 1, 2, and 3. Assuming a face value of 100, the zero-coupon bond for year 3 has payment stream $\mathbf{P} = (0, 0, 100)^\top$ and gives the portfolio $\mathbf{N} = \underline{\underline{M}}^{-1}\mathbf{P} = (-0.0866, 0, 0.9524)^\top$, i.e. a short position in 0.0866 units of the annuity bond and a long position of 0.9524 units of the three-year bullet bond. As another example, you can get a payment stream of $\mathbf{P} = (50, 75, 100)^\top$ with the portfolio $\mathbf{N} = \underline{\underline{M}}^{-1}\mathbf{P} = (0.7771, 0.25, 0.9524)^\top$.

5.5 Forward rates

A **forward rate** is the “fair” interest rate set today for a loan between two future dates. As an example, let time 0 denote the current date, and assume the loan starts at time 1 and is paid back at time 2. Let F be the face value, i.e., the proceeds paid to the borrower at time 1. If $f_{1,2}$ denotes the interest rate on the loan, it means that the amount to be paid back at time 2 is $(1 + f_{1,2})F$. Since the interest rate is fixed already today, both the payment at time 1 and the payment at time 2 are deterministic, at least if we assume that the two parties do not violate the loan agreement. Discounting the two payments with the market yields y_1 and y_2 prevailing at time 0 for one- and two-period zero-coupon bonds, the present value of the payments to the borrower is

$$PV_0 = F(1 + y_1)^{-1} - (1 + f_{1,2})F(1 + y_2)^{-2},$$

whereas the present value to the lender is then $-PV_0$. The “fair” value of the forward rate is the one that ensures a zero present value to both parties, which is satisfied when

$$f_{1,2} = \frac{(1+y_2)^2}{1+y_1} - 1. \quad (5.21)$$

Note that this implies that

$$(1+y_2)^2 = (1+y_1)(1+f_{1,2}),$$

where the left-hand side shows the value of discounting a dollar two years into the future in one step, and the right-hand side shows the same value but obtained by first discounting over the first year and then discounting forward over the second year using the forward rate. This confirms that the forward rate $f_{1,2}$ is the fair interest rate fixed today for discounting between time 1 and time 2.

Alternatively, we can express the present value in terms of the prices $Z_{0,1}$ and $Z_{0,2}$ of zero-coupon bonds with face values of 1:

$$PV_0 = FZ_{0,1} - (1+f_{1,2})FZ_{0,2}.$$

A zero present value requires a forward rate of

$$f_{1,2} = \frac{Z_{0,1}}{Z_{0,2}} - 1. \quad (5.22)$$

The forward rate is implicit in the bond prices. Given the link between the zero-coupon bond prices and yields, it is easy to check that the two expressions for the forward rate are consistent.

Along the same lines we can derive current forward rates for other future periods from the current prices or yields of zero-coupon bonds. For example, the forward rate for the period between time $n-1$ and time n is

$$f_{n-1,n} = \frac{(1+y_n)^n}{(1+y_{n-1})^{n-1}} - 1 = \frac{Z_{0,n-1}}{Z_{0,n}} - 1, \quad (5.23)$$

which is equivalent to the relation

$$(1+y_n)^n = (1+y_{n-1})^{n-1}(1+f_{n-1,n}). \quad (5.24)$$

From a zero-coupon yield curve $n \mapsto y_n$, we can thus derive a one-period **forward rate curve** $n \mapsto f_{n-1,n}$. Conversely, we can derive the zero-coupon yield curve from a one-period forward rate curve. The two curves carry the same information. Note that it follows from (5.23) that

$$\frac{1+f_{n-1,n}}{1+y_{n-1}} = \left(\frac{1+y_n}{1+y_{n-1}} \right)^n. \quad (5.25)$$

From this relation we can conclude that $y_n > y_{n-1}$ if and only if $f_{n-1,n} > y_{n-1}$, and the same statement holds if both inequality signs are changed or both are replaced by equality signs. For example, if the yield curve is increasing from maturity $n-1$ to maturity n , the forward rate $f_{n-1,n}$ must exceed the yield y_{n-1} . This supports the interpretation of the forward rate as capturing the marginal increase in the yield if you extend the maturity by another period.

If we apply Eq. (5.24) recursively, we obtain

$$(1 + y_n)^n = (1 + y_1)(1 + f_{1,2})(1 + f_{2,3}) \dots (1 + f_{n-1,n}). \quad (5.26)$$

The forward rate for a period starting today is simply the current yield over the period, so we can even replace y_1 in this equation by $f_{0,1}$. Then it is evident that the n -period yield can be seen as a geometric average of all the one-period forward rates $f_{0,1}, f_{1,2}, \dots, f_{n-1,n}$.

Forward rates can also be derived for longer time intervals. For example, the forward rate between time m and time n (where $m < n$) is

$$f_{m,n} = \left(\frac{(1 + y_n)^n}{(1 + y_m)^m} \right)^{1/(n-m)} - 1 = \left(\frac{Z_{0,m}}{Z_{0,n}} \right)^{1/(n-m)} - 1. \quad (5.27)$$

This ensures that

$$(1 + y_n)^n = (1 + y_m)^m (1 + f_{m,n})^{n-m},$$

where the left-hand side again reflects direct discounting over n periods, whereas the right-hand side reflects discounting first over the m nearest periods and then, using the forward rate, over the $n - m$ remaining periods.

5.6 Determinants of the shape of the yield curve

The yield on a default-free government bond of a given maturity aligns the demand and the supply of bonds with that maturity and therefore aligns

- 1 the willingness of bond investors to move money from today to the maturity date and thus postponing consumption or productive (non-financial) investments, and
- 2 the government's desire to move money from the maturity date to today and thus finance the current budget deficit, i.e., current expenses or productive investments.

Recall that households also act as suppliers of bonds when they take out mortgages financed by the issuance of bonds. Corporations can be both bond suppliers and investors.

5.6.1 The modern view

The level of the yield curve is clearly affected by the general time preference of investors. Suppose the average investor is very impatient and really wants to consume now rather than later. This would lead to a lot of borrowing (issuance of bonds) and very little demand for bonds. Consequently, prices of bonds drop and market yields and interest rates increase. High impatience leads to high yields.

Both the demand and supply side are affected by expectations of the future economic activity and the degree of uncertainty about that future activity. If the economy is expected to grow at a high rate over, say, the next year, then investors anticipate high consumption in one year from now relative to today. Many investors like to smooth consumption over time and would therefore want to borrow money over that year, since then they can increase their consumption today at the expense of a reduction in the high expected consumption next year. This mechanism decreases the demand for and increases the supply of one-year bonds, leading to a lower price and thus a higher yield on one-year bonds. Formulated differently, one-year bonds have to offer a high yield to convince investors to postpone consumption. Other things equal, if the economic growth rate is expected to increase over a number of years, you should expect to see an increasing yield curve for maturities in that range.

Uncertainty is also central to bond prices and therefore to the yield curve. When the uncertainty about the future growth rate of the economy is high, risk-averse investors find default-free government bonds particularly attractive. The high demand drives up bond prices and therefore leads to low yields. Suppose investors believe the macroeconomic uncertainty is high over the first few years, but then relatively low in the following years. Then, other things equal, you should expect to see low short-term yields and higher longer-term yields, i.e., an increasing yield curve.

The above considerations relate to the desire to move consumption opportunities over time and are therefore determining real interest rates and real yields, i.e., yields on inflation-indexed bonds that deliver a given purchasing power or, equivalently, a given level of consumption. Most traded bonds are nominal bonds delivering a given payment in some currency. The prices of nominal bonds are affected by expectations and uncertainty about future inflation rates in addition to the above-mentioned real determinants.

If you expect a high inflation over the life of a dollar-denominated bond, the fixed final dollar payment is really not that valuable, since you expect that it can only buy you few consumption goods. The high expected inflation lowers the price of the nominal bond and thus increases its yield. Other things equal, if the general market expects high inflation in the first few years and then lower inflation in subsequent years, you should expect to see a downward-sloping yield curve. A high degree of uncertainty about the inflation rate over the life of the nominal bond makes the purchasing power it delivers more risky, i.e., it makes the bond more risky in real terms. Again this lowers the demand for the bond, decreases the price, and increases the yield. However, the expectation and uncertainty about future inflation rates can also affect the expectation and uncertainty about the future growth rate of the economy and thus also influence yields through that channel.

Inflation risk affects all nominal securities and is therefore a systematic risk factor. Empirical research indicates that investors require a premium for buying assets exposed to inflation risk. Because of the inflation risk premium, nominal bonds will have lower prices and therefore higher yields. Finding the determinants of the magnitude and the fluctuations over time in the inflation risk premium is still an active research area.

5.6.2 Older theories

The **expectation hypothesis** relates the current interest rates and yields to expected future interest rates or returns. The basic idea dates back to [Fisher \(1896\)](#) and was further developed and concretized by [Hicks \(1939\)](#) and [Lutz \(1940\)](#). The original motivation of the hypothesis is that when lenders (bond investors) and borrowers (bond issuers) decide between long-term or short-term bonds, they will compare the price or yield of a long-term bond to the expected price or return on a roll-over strategy in short-term bonds. Hence, long-term rates and expected future short-term rates will be linked. Of course, a cornerstone of modern finance theory is that, when comparing different strategies, investors will also take the risks into account. So even before going into the specifics of the hypothesis you should really be quite skeptical, at least when it comes to very strict interpretations of the expectation hypothesis.

The vague idea that current yields and interest rates are linked to expected future rates and returns can be concretized in a number of ways. One specific version says that the forward rates implicit in current yield curve reflect expectations of future short-term yields. For example, the forward rate $f_{n,n+1}$ set today should equal the expectation of the one-year yield y_1 that will prevail n periods into the future. [Cox, Ingersoll, and Ross \(1981a\)](#) find that some versions of the expectation hypothesis are equivalent, some versions are inconsistent. They also show that, based on economic theory, none of the versions of the

expectation hypothesis should hold. And many, many empirical studies conclude that it does not hold in real-life bond markets.

Current yields may certainly be related to expectations about future interest rates, but if you want to apply such a relation you would have to think about how to form expectations about the future interest rates. Maybe such expectations are the result of certain expectations of future economic activity and inflation and then we end up in the more modern theories introduced in the preceding subsection.

Moreover, we know that prices of financial assets are not just determined by expectations. Risks are also important. The expectation hypothesis was developed long before modern finance theory with its focus on risk and investor preferences, so it is not surprising that it is too simplistic. It is really more surprising that some economists still apply it.

Another traditional explanation of the shape of the yield curve is given by the **liquidity preference hypothesis** introduced by [Hicks \(1939\)](#). He realized that the expectation hypothesis basically ignores investors' aversion towards risk and argued that expected returns on long-term bonds should exceed the expected returns on short-term bonds to compensate for the higher price fluctuations of long-term bonds. As will be discussed in Section 5.8, long-term bonds are generally more risky than short-term bonds. Hence, it makes sense that investors typically require a larger yield on long-term bonds than on short-term bonds, which explains why the yield curve tends to be increasing. Note that the word "liquidity" in the name of the hypothesis is not used in the modern sense of the word. Short-term bonds are not necessarily more liquid than long-term bonds. A better name would be "the maturity preference hypothesis".

In contrast the **market segmentation hypothesis** introduced by [Culbertson \(1957\)](#) claims that investors will typically prefer to invest in bonds with time-to-maturity in a certain interval, a maturity segment, perhaps in an attempt to match liabilities with similar maturities. For example, a pension fund with liabilities due in 20-30 years can reduce risk by investing in bonds of similar maturity. On the other hand, central banks typically operate in the short end of the market. Hence, separated market segments can exist without any relation between the bond prices and the interest rates in different maturity segments. If this is really the case, we cannot expect to see continuous or smooth yield curves and discount functions across the different segments.

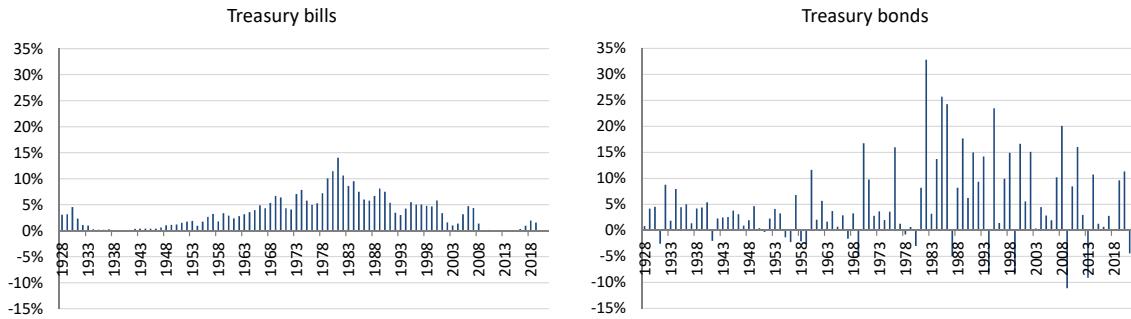
A more realistic version of this hypothesis is the **preferred habitats hypothesis** put forward by [Modigliani and Sutch \(1966\)](#). An investor may prefer bonds with a certain maturity, but should be willing to move away from that maturity if she is sufficiently compensated in terms of a higher yield.² The different segments are therefore not completely independent of each other, and yields and discount factors should depend on maturity in a smooth way.

It is really not possible to quantify the market segmentation or the preferred habitats hypothesis without setting up an economy with agents having different favorite maturities. The resulting equilibrium yield curve will depend heavily on the degree of risk aversion of the various agents as illustrated by an analysis of [Cox, Ingersoll, and Ross \(1981a\)](#).

5.7 Stylized facts about bond returns and interest rates

Average returns and volatilities increase with maturity, whereas Sharpe ratios decrease with maturity. Figure 5.8 shows the nominal rates of return on 3-month U.S. Treasury bills (left panel) and 10-year U.S. Treasury bonds (right panel) in each of the years from 1928 to 2021. Consistent with the characteristics of yield curves presented in Section 5.3,

²In a sense the liquidity preference hypothesis simply says that all investors prefer short bonds.

**Figure 5.8: Time series of Treasury returns.**

The graphs show annual returns on 3-month Treasury bills and 10-year Treasury bonds in the U.S. over the period 1928-2021. The data are taken from the homepage of Professor Aswath Damodaran at the Stern School of Business at New York University, see <http://pages.stern.nyu.edu/~adamodar>.

	Inflation	1M	3M	1Y	2Y	5Y	7Y	10Y	20Y	30Y
Avg return	3.67%	3.84%	4.25%	4.72%	4.95%	5.50%	5.82%	5.59%	6.18%	6.16%
Standard dev	1.59%	0.92%	1.05%	1.79%	2.75%	4.99%	6.14%	7.36%	9.86%	11.35%
Sharpe ratio		0.398	0.492	0.406	0.334	0.323	0.238	0.238	0.204	

Table 5.2: Return statistics for U.S. Treasury bonds.

The statistics are based on monthly observations over the period from January 1946 to December 2021 downloaded from CRSP U.S. Treasury and Inflation Indexes on July 4, 2022. The returns are nominal. The statistics shown are annualized from monthly statistics. For the average return and standard deviation for both bond returns and the inflation rate, the annualization follows Eqs. (3.83) and (3.84). The annualized Sharpe ratio for a given maturity is calculated as the difference of the annualized average return for that maturity minus the annualized average return on 1-month bills, divided by the annualized standard deviation for the given maturity.

we see that the T-bill returns are generally lower and less volatile than the T-bond returns. The arithmetic average returns are 3.33% and 5.11%, and the geometric average returns are 3.28% and 4.84%, respectively. The standard deviation is 3.04% for T-bills and 7.68% for T-bonds.

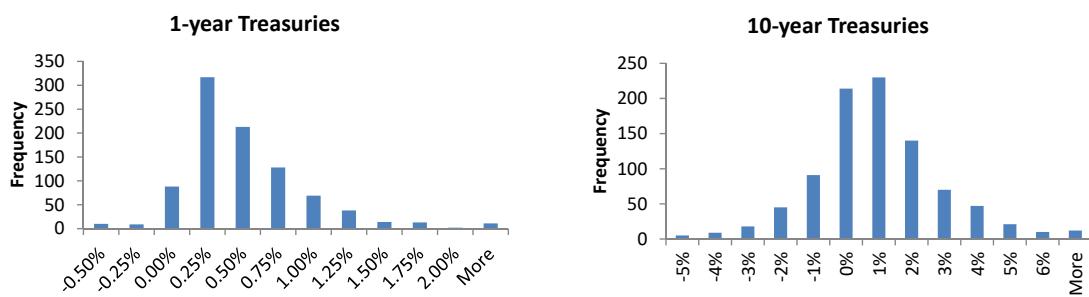
Based on monthly returns on nominal U.S. Treasury bonds from 1946 to 2021, Table 5.2 shows the annualized average return, standard deviation, and Sharpe ratios for bonds with maturities ranging from one month to 30 years. In line with the statistics described above, we see that average returns and standard deviations are increasing in maturity. However, the Sharpe ratio is decreasing in maturities above three months. For example, the Sharpe ratio on one-year bonds is more than twice the Sharpe ratio on 10-year bonds. By leveraging up the one-year bond to the same standard deviation as the 10-year bond, you could obtain a much larger expected return than on the 10-year bond. As discussed in Chapter 6, the S&P 500 stock index has an average annual excess return around 8% and an annual standard deviation of around 18%, which gives a Sharpe ratio of 0.444. The short-term Treasury bonds thus have a Sharpe ratio of the same magnitude as the stock market index. In Section 6.5 we dig deeper into the historical performance of stocks and bonds, and there we present similar numbers for other countries.

van Binsbergen and Kooijen (2017) present statistics for corporate bonds of two different

	Intermediate maturity (≈ 5 years)				Long maturity (≈ 10 years)			
	AAA	AA	A	BAA	AAA	AA	A	BAA
Average excess return	2.38%	2.53%	2.76%	3.44%	3.12%	3.80%	3.75%	4.60%
Standard deviation	5.02%	4.99%	5.28%	5.48%	10.45%	9.74%	9.67%	9.82%
Sharpe ratio	0.47	0.51	0.52	0.63	0.30	0.39	0.39	0.47

Table 5.3: Return statistics for U.S. corporate bonds.

The statistics are annualized and based on data from Barclays corporate bond indexes over the period from January 1973 to August 2014. Intermediate maturity corresponds to a duration of about 5 years, and long maturity to a duration of about 10 years. Source: Table 5 in van Binsbergen and Kojen (2017).

**Figure 5.9: Return distributions of Treasury bonds.**

The graphs are histograms of the monthly nominal returns from January 1946 to December 2021. The left graph is for one-year Treasuries, the right graph for ten-year Treasuries. The data were downloaded from CRSP U.S. Treasury and Inflation Indexes on July 4, 2022.

maturities, intermediate (duration about 5 years) and long term (duration about 10 years), and four different credit qualities as described by their credit ratings (see discussion in Section 5.10.2). Table 5.3 reproduces their results. We see that in each rating category, the average excess return and the standard deviation are increasing in maturity, whereas the Sharpe ratio is decreasing in maturity—as for Treasury bonds. For a fixed maturity, the average excess return and the Sharpe ratio tend to decrease with credit quality, i.e. they are larger for relatively low-quality (BAA) bonds than for high-quality (AAA) bonds, whereas there is no clear pattern in the standard deviations. Corporate bonds seem to have larger average returns, larger standard deviations, and larger Sharpe ratios than Treasury bonds of a similar maturity, although the statistics in these tables do not allow a direct comparison.

Bond returns are not normally distributed. Figure 5.9 shows histograms of the monthly nominal returns on one- and ten-year Treasury securities over the period from 1946 to 2021. The mean and standard deviation of the two series can be seen in Table 5.2. For the one-year bonds, the skewness is 2.6, the kurtosis 18.6, the minimum observation -1.72% (February 1980), and the maximum 5.61% (April 1980). For the ten-year bonds, the skewness is 0.5, the kurtosis 2.09, the minimum observation -6.68% (July 2003), and the maximum 10.00% (October 1982). Both distributions—and especially that of the one-year bond—thus deviate from the normal distribution, but the normal distribution does not seem to be a completely terrible approximation.

Government bonds with close maturities have high correlations. Two bonds that are

	1M	3M	6M	1Y	2Y	3Y	5Y	7Y	10Y	20Y	30Y
1M	1.00	0.97	0.93	0.88	0.72	0.60	0.46	0.37	0.36	0.30	0.30
3M	0.97	1.00	0.98	0.93	0.78	0.67	0.53	0.45	0.43	0.37	0.36
6M	0.93	0.98	1.00	0.97	0.83	0.73	0.58	0.49	0.46	0.39	0.38
1Y	0.88	0.93	0.97	1.00	0.91	0.82	0.67	0.58	0.54	0.46	0.44
2Y	0.72	0.78	0.83	0.91	1.00	0.97	0.86	0.77	0.72	0.63	0.59
3Y	0.60	0.67	0.73	0.82	0.97	1.00	0.95	0.87	0.82	0.73	0.68
5Y	0.46	0.53	0.58	0.67	0.86	0.95	1.00	0.98	0.94	0.87	0.83
7Y	0.37	0.45	0.49	0.58	0.77	0.87	0.98	1.00	0.99	0.94	0.90
10Y	0.36	0.43	0.46	0.54	0.72	0.82	0.94	0.99	1.00	0.97	0.95
20Y	0.30	0.37	0.39	0.46	0.63	0.73	0.87	0.94	0.97	1.00	0.99
30Y	0.30	0.36	0.38	0.44	0.59	0.68	0.83	0.90	0.95	0.99	1.00

Table 5.4: Bond correlations.

The table shows correlations between monthly changes in yields of U.S. Treasury bonds of different maturities in the period January 2012 to December 2021. Source: <http://www.federalreserve.gov/releases/h15/data.htm>, data retrieved on July 4, 2022.

issued by the same issuer, follow the same amortization principle, and have close maturities are very similar assets. Hence, we would expect the rates of return of the two bonds to be highly correlated. As shown in the next section, the rate of return on a bond can be approximated by (a multiple of) the change in its yield, so instead of looking at return correlations we consider correlations between changes in yields. Table 5.4 shows correlations between monthly yield changes in U.S. Treasuries of various maturities over a period from 2012 to 2021. The table confirms that government bonds of similar maturities exhibit high correlations. For example, the correlation between the yield changes of the two-year and the three-year Treasury bonds was 0.97 in the period considered. As the maturity distance is increased, the correlation drops. For example, the correlation between the one-year bond and the ten-year bond was 0.54, which is lower but still relatively high. As explained in Chapter 4, we can generally reduce the risk of our return by spreading the investment over several assets, but the size of risk reduction and thus the diversification gain depend heavily on the correlation between the assets. Apparently, little is gained by simultaneously investing in government bonds with very similar maturities.

Short-term interest rates exhibit considerable persistence. The autocorrelation in monthly nominal returns on one-month Treasury bills was 0.97 over the period 1946-2021, compared to 0.43 for one-year Treasury bonds, and 0.08 for 10-year Treasury bonds. If we form real monthly returns by subtracting the realized inflation rate from the nominal returns, we get a monthly autocorrelation of 0.48 for one-month bills, 0.40 for one-year bonds, and 0.12 for 10-year bonds. While the inflation rate is responsible for a large part of the autocorrelation in short-maturity bond returns, even the real returns show a sizeable autocorrelation. Another feature of interest rates is that they tend to mean revert, in particular short-term interest rates. Also, the volatility of short-term interest tends to increase with the level of interest rates, see for example Chan, Karolyi, Longstaff, and Sanders (1992).

The excess returns on U.S. Treasury bonds relative to very short-term interest rates are predictable by changes in the yield spreads over time. Campbell and Shiller (1991) and Campbell, Lo, and MacKinlay (1997, Ch. 10) find that a high yield spread between a long-term and a short-term interest rate forecasts an increase in short-term interest rates in the long run and a decrease in the yields on long-term bonds in the near future. Other studies indicate that a combination of forward rates can predict bond returns, see for example Fama and Bliss (1987), Stambaugh (1988), and Cochrane and Piazzesi (2005).

5.8 Interest rate risk

The riskiness of a bond investment is highly dependent on the length of the period over which the investor intends to hold the bond. First, think of default-free, inflation-indexed zero-coupon bonds. If you purchase such a bond and hold it to maturity, the investment is riskfree: at the time of purchase, you know exactly which purchasing power the bond's final payment will offer you. A five-year zero-coupon bond is the riskfree asset for a five-year investment horizon. A one-year zero-coupon bond is the riskfree asset for a one-year horizon. However, if you purchase a five-year zero-coupon bond and intend to sell it after one year, the bond is not riskfree. Your return is then depending on the selling price of the bond or, equivalently, what the yield of the bond is when you sell it. When you purchase the bond, you do not know what the bond's yield is going to be at the selling date. Hence, in this case the return of the bond investment is subject to interest rate risk. The prices of default-free, inflation-indexed bonds fluctuate with real interest rates. The prices of default-free nominal bonds fluctuate with nominal interest rates, i.e. both real interest rates and inflation rates. The prices of defaultable bonds are also affected by reassessments of the default risk and the recovery rates in case of default (see Section 5.10.2).

The aim of this section is to determine simple measures that indicate how sensitive the price of a bond is to changes in the market interest rates. It is clear from the general relation (5.4) between the discount rate and the bond price that bond prices increase when interest rates fall and decrease when interest rates rise. But prices of different bonds react differently to interest rate changes. We want to measure *how* sensitive the price is to interest rate movements.

5.8.1 Duration

The most frequently applied measure of interest rate risk is the duration of the bond, which was introduced already by Macaulay (1938). As in earlier sections suppose we are currently at time 0 and consider a bond paying M_1 at time 1, M_2 at time 2, etc., up to M_n at time n . The **duration** of the bond is defined as

$$D_0 = \sum_{i=1}^n i w_i, \quad (5.28)$$

where

$$w_i = \frac{M_i(1+y)^{-i}}{B_0} \quad (5.29)$$

is the relative weight of the i 'th payment in the bond price $B_0 = \sum_{i=1}^n M_i(1+y)^{-i}$. In particular,

$$\sum_{i=1}^n w_i = \sum_{i=1}^n \frac{M_i(1+y)^{-i}}{B_0} = \frac{\sum_{i=1}^n M_i(1+y)^{-i}}{B_0} = 1.$$

Due to its definition, the duration is interpreted as an effective or weighted time-to-maturity. The time until each payment date is weighted by the payment's relative importance in present value terms. If two bonds of the same maturity have different payment schedules, they have different durations reflecting the distribution of payments over time. Practitioners often use the so-called **modified duration** defined as

$$D_0^* = \frac{D_0}{1+y}, \quad (5.30)$$

The following theorem establishes the duration as a measure of interest rate risk.

Theorem 5.2

The duration satisfies

$$D_0 = -\frac{1+y}{B_0} \frac{\partial B_0}{\partial y}. \quad (5.31)$$

A first-order approximation of the relative bond price change in response to an immediate yield change of Δy is

$$\frac{\Delta B_0}{B_0} \approx -\frac{\Delta y}{1+y} D_0 = -D_0^* \Delta y. \quad (5.32)$$

Proof

The derivative of the bond price with respect to its own yield is

$$\begin{aligned} \frac{\partial B_0}{\partial y} &= -\sum_{i=1}^n i M_i (1+y)^{-i-1} \\ &= -(1+y)^{-1} \sum_{i=1}^n i M_i (1+y)^{-i} = -(1+y)^{-1} \sum_{i=1}^n i w_i B_0 \\ &= -\frac{B_0}{1+y} \sum_{i=1}^n i w_i = -\frac{B_0}{1+y} D_0 \end{aligned} \quad (5.33)$$

and now Eq. (5.31) follows.

The derivative $\frac{\partial B_0}{\partial y}$ gives the price sensitivity for a marginal (i.e., extremely small) change in the yield. For a given yield change Δy , the resulting price change ΔB_0 is approximated as

$$\frac{\Delta B_0}{\Delta y} \approx \frac{\partial B_0}{\partial y} = -\frac{B_0}{1+y} D_0,$$

which implies

$$\Delta B_0 \approx \frac{\partial B_0}{\partial y} \Delta y = -\frac{B_0}{1+y} D_0 \Delta y,$$

and thus leads to Eq. (5.32).

Suppose we are currently at date $t \in (0,1)$ between two payment dates. Then the duration is defined as

$$D_t = \sum_{i=1}^n (i-t) w_i, \quad w_i = \frac{M_i (1+y)^{-(i-t)}}{\sum_{i=1}^n M_i (1+y)^{-(i-t)}}. \quad (5.34)$$

where $i-t$ is the time until payment i and w_i is still the relative present value weight of the i 'th payment.

For any single bond or any other payment stream, the duration can be calculated directly using (5.28) by deriving the present value weights of each payment.

Year i	1	2	3	4	5	6	7	8	sum
Cash flow M_i	90	90	90	90	90	90	90	1090	
Present value	81.818	74.380	67.618	61.471	55.883	50.803	46.184	508.493	946.651
Weight w_i	0.0864	0.0786	0.0714	0.0649	0.0590	0.0537	0.0488	0.05371	1
$i \times w_i$	0.0864	0.1571	0.2143	0.2597	0.2952	0.3220	0.3415	4.2972	5.9735

Table 5.5: Computation of the duration.

The bond is an 8-year bullet bond with annual payments. It has a face value of 1000, a coupon rate of 9%, and a yield of 10%.

Example 5.7

Table 5.5 shows how to compute the duration of a bond. The bond is an 8-year bullet bond with annual payments. It has a face value of 1000, a coupon rate of 9%, and a yield of 10%. The price is the sum of the present values of the individual payments, which is 946.651. The duration D_0 is 5.9735 in this case as shown in the bottom row of the table.

For the standard bond types, we have the expressions for durations shown in the next theorem. Note that Excel's built-in functions DURATION and MDURATION deliver the duration and modified duration for bullet bonds.

Theorem 5.3

Let y denote the current yield of the bond.

- (a) For a zero-coupon bond maturing at time n , the duration at time $t < n$ is the time to maturity, i.e.

$$D_t = n - t. \quad (5.35)$$

- (b) Let time 0 denote a time point with a full period until the next payment date and n full periods until maturity. Let q denote the coupon rate. Then the time 0 durations of different coupon bonds are

$$\text{Bullet bond: } D_0 = \frac{1+y}{y} - \frac{1+y-n(y-q)}{q[(1+y)^n - 1] + y}, \quad (5.36)$$

$$\text{Perpetuity: } D_0 = \frac{1+y}{y}, \quad (\text{provided } y > 0) \quad (5.37)$$

$$\text{Annuity bond: } D_0 = \frac{1+y}{y} - \frac{n}{(1+y)^n - 1}, \quad (5.38)$$

$$\text{Serial bond: } D_0 = \frac{1+y}{y} \left(1 - \frac{qA(y,n) + n(y-q)(1+y)^{-n-1}}{qn + (y-q)A(y,n)} \right). \quad (5.39)$$

- (c) If there is only a fraction $1-t \in (0, 1)$ of a period until the next payment date, then the duration at time t is given by

$$D_t = D_0 - t, \quad (5.40)$$

where the above formulas for D_0 can be used for standard coupon bond types.

Proof

(a) If a bond only has a single payment left, that payment naturally has a weight of one and this weight is multiplied by the time until the payment is received, which is the time to maturity of the bond.

(b) The bond price as a function of the yield follows by replacing r by y in the formulas of Theorem 5.1. Now calculate the derivative of B_0 with respect to y and substitute this into (5.31) to get the duration. For the perpetuity the computation is simple: the price-yield relation is $B_0 = F_0 q/y$, cf. Eq. (5.7), so the relevant derivative is

$$\frac{\partial B_0}{\partial y} = -\frac{qF_0}{y^2} = -\frac{1}{y} \frac{qF_0}{y} = -\frac{B_0}{y}.$$

Hence, the duration of the perpetuity is

$$D_0 = -\frac{1+y}{B_0} \frac{\partial B_0}{\partial y} = \frac{1+y}{B_0} \frac{B_0}{y} = \frac{1+y}{y}.$$

We skip the tedious details of the calculations for the other bond types.

(c) Note that the weights in (5.34) satisfy

$$w_i = \frac{(1+y)^t M_i (1+y)^{-i}}{(1+y)^t \sum_{i=1}^n M_i (1+y)^{-i}} = \frac{M_i (1+y)^{-i}}{\sum_{i=1}^n M_i (1+y)^{-i}}$$

so they are exactly the same as they were at the preceding payment date 0, presuming of course that the yield was the same. Hence, it follows that

$$D_t = \sum_{i=1}^n (i-t) w_i = \sum_{i=1}^n i w_i - t \sum_{i=1}^n w_i = D_0 - t.$$

Equation (5.34) shows that the duration declines steadily with time as long as the yield remains unchanged and we do not pass a payment date of the bond.

Figure 5.10 shows how the duration depends on the yield for the three bonds considered in Example 5.1. For any level of the yield, the bullet bond has a higher duration than the annuity bond, which again has a higher duration than the serial bond. If we think of the duration as a weighted time-to-maturity, this ranking is natural given the different payment schedules of the three bonds, cf. Figure 5.1. If we think of the duration as a measure of interest rate risk, the bullet bond is the most risky since a given change in the yield will affect the present value of payments far into the future more than the present value of payments in the near future. The serial bond has the shortest weighted time-to-maturity of the three bonds and is therefore the least sensitive to yield changes, hence it has the lowest duration.

It is clear from Figure 5.10 that the duration of a bond decreases with its yield. A higher yield reduces the present value of distant payments more than the payments in the near future, so the relative weights are shifted towards the earlier payments leading to a reduction in the weighted time-to-maturity.

Not surprisingly, the duration of a bond generally increases with the time-to-maturity of the bond. This is true for all bonds having a yield smaller than or equal to the coupon rate and also for bonds having a yield somewhat larger than the coupon rate. For coupon

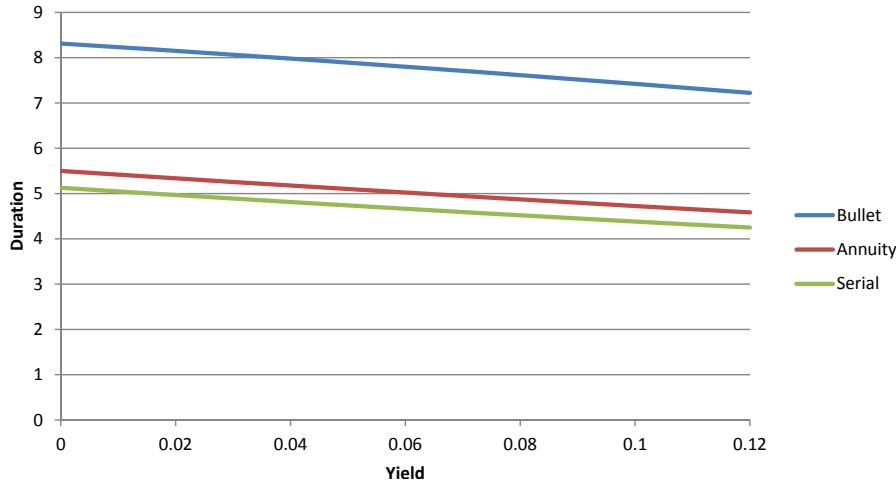


Figure 5.10: Yields and durations.

The graphs show the duration as a function of the yield for a bullet bond, an annuity bond, and a serial bond. All bonds have $F = 100$, $q = 0.06$, and $n = 10$.

bonds trading at a yield much higher than the coupon rate, the duration may in some cases decrease slightly with maturity.

How does the duration depend on the coupon rate? First note that for an annuity bond, the duration is independent of the coupon rate. If you increase the coupon rate of an annuity bond, all payments increase by exactly the same amount, so the relative weights remain the same. In contrast, for a bullet bond or a serial bond, the duration is decreasing in the coupon rate. For a bullet bond a modest increase in the coupon rate will only cause a small relative change in the terminal payment as this is dominated by the repayment of the face value, whereas the relative change in all the earlier payments is much higher. Hence, the relative weights of the earlier payments increase and the relative weight of the terminal payment will decrease, leading to a lower weighted time-to-maturity. For a serial bond an increase in the coupon rate has a bigger relative impact on the first payment than the second payment and so forth, again leading to higher relative weights on the early payments and thus a lower duration.

5.8.2 Convexity

Eq. (5.32) gives a duration-based approximation of the bond price change as a function of the yield change. According to the approximation, the price change is proportional to the yield change. This corresponds to assuming a linear relation between the bond price and the yield. However, we know that the true relation is non-linear as illustrated, for example, in Figure 5.2. The linear approximation may be sufficiently accurate for small yield changes, but can be imprecise for larger yield changes. In fact, we know the price-yield relation is convex so the linear approximation underestimates the price increase when the yield falls and overestimates the price drop when the yield goes up. We can improve upon the approximation by including a quadratic term, and this term is closely related to the so-called convexity of the bond.

Formally, we define the **convexity** of the bond at time 0 as

$$C_0 = \sum_{i=1}^n i(i+1)w_i, \quad (5.41)$$

where w_i is still given by (5.29). Slightly different definitions of the convexity are sometimes used, e.g., in other textbooks. We further define the **modified convexity** as

$$C_0^* = \frac{C_0}{(1+y)^2}. \quad (5.42)$$

The next theorem links the convexity and the second-order derivative $\partial^2 B_0 / \partial y^2$.

Theorem 5.4

The convexity satisfies

$$C_0 = \frac{(1+y)^2}{B_0} \frac{\partial^2 B_0}{\partial y^2} \quad (5.43)$$

and

$$C_0 = \sum_{i=1}^n i^2 w_i + D_0. \quad (5.44)$$

The convexity is linked to the yield-sensitivity of the duration through

$$\frac{\partial D_0}{\partial y} = \frac{D_0(1+D_0) - C_0}{1+y} \Leftrightarrow C_0 = D_0(1+D_0) - (1+y) \frac{\partial D_0}{\partial y}. \quad (5.45)$$

A second-order approximation of the relative bond price change in response to an immediate yield change of Δy is

$$\frac{\Delta B_0}{B_0} \approx -\frac{D_0}{1+y} \Delta y + \frac{1}{2} \frac{C_0}{(1+y)^2} (\Delta y)^2 = -D_0^* \Delta y + \frac{1}{2} C_0^* (\Delta y)^2. \quad (5.46)$$

Proof

The second-order derivative of the bond price follows by differentiation of the first-order derivative in (5.33):

$$\frac{\partial^2 B_0}{\partial y^2} = \sum_{i=1}^n i(i+1) M_i (1+y)^{-i-2}.$$

Now we can rewrite the convexity as

$$C_0 = \frac{\sum_{i=1}^n i(i+1) M_i (1+y)^{-i}}{B_0} = \frac{(1+y)^2}{B_0} \sum_{i=1}^n i(i+1) M_i (1+y)^{-i-2} = \frac{(1+y)^2}{B_0} \frac{\partial^2 B_0}{\partial y^2},$$

which confirms (5.43). Eq. (5.44) follows easily:

$$C_0 = \sum_{i=1}^n i(i+1) w_i = \sum_{i=1}^n i^2 w_i + \sum_{i=1}^n i w_i = \sum_{i=1}^n i^2 w_i + D_0.$$

Exercise 5.15 asks for a proof of the first equality in (5.45). The second equality follows by isolating C_0 in the first equality.

A second-order approximation of B_0 around the current yield y means

$$\Delta B_0 \approx \frac{\partial B_0}{\partial y} \Delta y + \frac{1}{2} \frac{\partial^2 B_0}{\partial y^2} (\Delta y)^2.$$

By substitution of the relations (5.31) and (5.43) between the partial derivatives and the duration and convexity, we get (5.46).

While the duration is a weighted average time-to-payment, we can loosely think of the convexity as a measure of how much the payments are spread out over time. Why? The sum in Eq. (5.44) is the weighted average of the squared time-until-payment. The more spread out the payments are, the higher this sum and thus the convexity. For example, consider a bond with a single payment in 5 years. Since that payment has a weight of 1, the duration is 5 and $\sum_{i=1}^n i^2 w_i = 5^2 \times 1 = 25$. Hence the convexity is $C_0 = 25 + 5 = 30$. Now consider a bond with a payment in 1 year and a payment in 9 years, each having a present value weight of 0.5. Then the duration is $0.5 \times 1 + 0.5 \times 9 = 5$ as for the first bond. Since now $\sum_{i=1}^n i^2 w_i = 1^2 \times 0.5 + 9^2 \times 0.5 = 41$, this bond has a convexity of $C_0 = 41 + 5 = 46$, higher than that of the first bond.

The second-order approximation (5.46) is certainly a better approximation than the first-order approximation (5.32). In fact, an even better approximation is

$$\frac{\Delta B_0}{B_0} \approx \exp \left\{ -D_0^* \Delta y + \frac{1}{2} [C_0^* - (D_0^*)^2] (\Delta y)^2 \right\} - 1, \quad (5.47)$$

as was demonstrated by Barber (1995). This approximation stems from a second-order Taylor expansion of the log bond price, $\ln B_0(y)$.

With modern computers and financial calculators, it is very easy to compute the exact price change caused by a change in the yield of the bond. Hence, the above-mentioned approximations are not so important any longer. What is more important is the role of duration and convexity in the measurement of interest rate risk. As explained in the preceding subsection, the duration reflects the price sensitivity to the yield of the bond and is thus a natural risk measure. Since convexity is defined in terms of the second-order derivative of the price, you can think of the convexity as a second-order risk measure.

The above discussion assumes that the bond in consideration has a full period until the next payment date, i.e., that the previous payment has just been made. We can define the convexity between two payment dates similarly: at time $t \in (0,1)$, the convexity is

$$C_t = \sum_{i=1}^n (i-t)(i+1-t)w_i,$$

where $i-t$ is the time distance to the next payment date.

5.8.3 Duration and convexity for portfolios of bonds

Investors typically hold a portfolio of different bonds. How are the risk measures—duration and convexity—of the portfolio related to the risk measures of the individual bonds in the portfolio? In the following we assume we are at time 0 and drop the usual 0-subscripts to simplify the notation. Instead, we use a p -subscript to indicate a portfolio

and subscripts $j = 1, 2, \dots, J$ to indicate the individual bonds in the portfolio. If we let B_p denote the market price of the portfolio and M_{p1}, \dots, M_{pn} its cash flow, the portfolio yield y_p is implicitly defined by

$$B_p = \sum_{i=1}^n M_{pi}(1 + y_p)^{-i},$$

just as for an individual bond. Similarly, the duration and convexity of a portfolio are defined as

$$D_p = \sum_{i=1}^n i \frac{M_{pi}(1 + y_p)^{-i}}{B_p}, \quad C_p = \sum_{i=1}^n i(i+1) \frac{M_{pi}(1 + y_p)^{-i}}{B_p}. \quad (5.48)$$

We are searching for a hopefully simple relation between the portfolio values y_p , D_p , C_p on the one hand and the bond values y_j , D_j , C_j on the other hand.

To see the complexity of this problem, think of a two-bond portfolio with N_1 units of bond 1 and N_2 units of bond 2. The two bonds are assumed to have overlapping payment dates. Let M_{1i} and M_{2i} denote the payments of the two bonds at payment date i . The prices and yields of the two bonds at time 0 are related through the equations

$$B_1 = \sum_{i=1}^n M_{1i}(1 + y_1)^{-i}, \quad B_2 = \sum_{i=1}^n M_{2i}(1 + y_2)^{-i}.$$

The payments of the portfolio are given by $M_{pi} = N_1 M_{1i} + N_2 M_{2i}$ and the price of the portfolio is $B_p = N_1 B_1 + N_2 B_2$. Consequently, the portfolio yield and the bond yields are connected via the relation

$$\underbrace{\sum_{i=1}^n (N_1 M_{1i} + N_2 M_{2i})(1 + y_p)^{-i}}_{=B_p} = N_1 \underbrace{\sum_{i=1}^n M_{1i}(1 + y_1)^{-i}}_{=B_1} + N_2 \underbrace{\sum_{i=1}^n M_{2i}(1 + y_2)^{-i}}_{=B_2}. \quad (5.49)$$

In general, we cannot solve this for y_p , and the exact portfolio yield y_p is thus not just a simple weighted average of the bond yields. The next theorem presents a weighted bond yield which is typically a good approximation of y_p . However, in the special case where $y_1 = y_2$, Eq. (5.49) is satisfied when $y_p = y_1 = y_2$, so when all the bonds in the portfolio have identical yields, then the portfolio yield is equal to that common bond yield.

Theorem 5.5

A bond portfolio's yield y_p , duration D_p , and convexity C_p are approximately given by

$$y_p \approx \sum_{j=1}^J k_j y_j, \quad D_p \approx \sum_{j=1}^J \pi_j D_j, \quad C_p \approx \sum_{j=1}^J \pi_j C_j, \quad (5.50)$$

where π_j is the portfolio weight of bond j and $k_j = \pi_j D_j / \sum_{m=1}^J \pi_m D_m$. If all bonds in the portfolio have identical yields, then the approximations are exact.

Proof

For a portfolio of two bonds with identical yields, we have

$$\begin{aligned}\frac{\partial B_p}{\partial y} &= N_1 \frac{\partial B_1}{\partial y} + N_2 \frac{\partial B_2}{\partial y} = N_1 \left(-\frac{B_1}{1+y} D_1 \right) + N_2 \left(-\frac{B_2}{1+y} D_2 \right) \\ &= -\frac{1}{1+y} (N_1 B_1 D_1 + N_2 B_2 D_2),\end{aligned}$$

and, consequently, the portfolio duration is

$$\begin{aligned}D_p &= -\frac{1+y}{B_p} \frac{\partial B_p}{\partial y} = \frac{1+y}{B_p} \frac{1}{1+y} (N_1 B_1 D_1 + N_2 B_2 D_2) \\ &= \frac{N_1 B_1}{B_p} D_1 + \frac{N_2 B_2}{B_p} D_2 = \pi_1 D_1 + \pi_2 D_2,\end{aligned}$$

so simply a weighted average of the durations of the bonds in the portfolio. Similarly for the convexity. In other cases, the relations hold only as approximations.

Note that $\sum_{j=1}^J k_j = 1$. If all bonds have the same yield y , then

$$\sum_{j=1}^J k_j y_j = \sum_{j=1}^J k_j y = y \sum_{j=1}^J k_j = y.$$

As argued before the theorem, we know that the portfolio yield is also y in this case, so also the first approximation in (5.50) holds exactly in this special situation.

The following example examines the accuracy of the approximations of the yield, duration, and convexity of a bond portfolio.

Example 5.8

Consider a simple bond portfolio consisting only of one unit of a 2-year 3% bond and one unit of a 5-year 4% bond. Both bonds are bullet bonds with one annual payment and exactly one year until the next payment, and their face value is \$1000. We will calculate the exact and the approximate values of the portfolio's yield, duration, and convexity for three different shapes of the zero-coupon yield curve for the relevant maturity range from 1 to 5 years:

1. Flat yield curve at 3%.
2. Increasing yield curve: the 1-year zero-coupon yield is 1%, the 2-year zero-coupon yield is 2%, etc., i.e., the zero-coupon yield curve is linearly increasing in maturity with a slope of 1.
3. Decreasing yield curve: the 1-year zero-coupon yield is 5%, the 2-year zero-coupon yield is 4%, etc., i.e., the zero-coupon yield curve is linearly decreasing in maturity with a slope of -1.

The non-flat yield curves are very steep so we are testing the precision of the approximation in rather extreme cases. For each of the three yield curves, we calculate the price,

	Flat curve			Increasing curve			Decreasing curve		
	Yield, %	Dura	Conv	Yield, %	Dura	Conv	Yield, %	Dura	Conv
2-year bond	3.0000	1.9709	5.8835	1.9852	1.9712	5.8846	4.0148	1.9706	5.8824
5-year bond	3.0000	4.6393	27.1251	4.8342	4.6219	26.9929	1.1414	4.6564	27.2553
Portfolio, exact	3.0000	3.3350	16.7421	3.9638	3.3137	16.5750	1.9234	3.3590	16.9310
Portfolio, approx	3.0000	3.3350	16.7421	3.9484	3.2591	16.1408	1.9093	3.4132	17.3621

Table 5.6: Immunization outcomes.

The table shows bond and portfolio yield and risk measures with three different yield curves as explained in Example 5.8.

yield (using Excel's Solver), duration, and convexity of each of the two bonds and of the portfolio, as well as the approximate values for the portfolio. Table 5.6 summarizes the results. When the yield curve is flat at 3%, both bonds and the portfolio have a yield of 3%, and the approximations of the portfolio yield, duration, and convexity are exact as explained above. For both the increasing and the decreasing yield curve, the approximation underestimates the yield of the portfolio. For the increasing yield curve, the approximation underestimates the portfolio's duration and convexity. Conversely, for the decreasing yield curve, the approximation overestimates the duration and convexity. Even with these extreme curves, the approximate values deviate by less than 3% from the exact values so the approximations appear to be sufficiently accurate for most purposes.

5.8.4 Duration and convexity as risk measures

The Macaulay duration of a bond measures how sensitive the bond's price is to changes in its own yield. Generally, different bonds have different yields. And the changes in the yields of different bonds over a given time period are generally also different. A certain shift of the zero-coupon yield curve from one day to the next may have very different effects on different bonds. Hence, in general, it seems inappropriate to compare the interest rate sensitivity of different bonds by comparing their Macaulay durations.

Recall the duration-based first-order approximation in Eq. (5.32) of the relative bond price change in response to a yield change. In terms of the modified duration D_0^* , the relative price change is $\Delta B_0/B_0 \approx -D_0^* \Delta y$. Therefore, if we want to compare the relative price change of two different bonds, we can do so by comparing their modified durations under the condition that the yield change is the same for both bonds. More generally, if the yield curve changes only in the form of parallel shifts, the modified duration is an appropriate risk measure. In terms of the non-modified duration D_0 , the relative price change is $\Delta B_0/B_0 \approx -\frac{\Delta y}{1+y} D_0$. If the yield curve only changes so that $\Delta y/(1+y)$ is the same for all bonds, then the non-modified duration is an appropriate risk measure. This requires proportional yield curve shifts in the sense that $\Delta y = k(1+y)$ for some value of k which is common to all bonds.

But yield curves shifts are not always parallel or proportional. Typically short-maturity yields move by more than long-maturity yields. Some yield curve shifts move the entire curve in the same direction, but generally not by the same distance for all maturities. Other yield curve shifts move short-maturity yields in one direction and long-maturity yields in the other direction—a so-called twist of the yield curve—so that the prices of short-maturity and long-maturity bonds can move in opposite directions.

Macaulay (1938) also defined an alternative duration measure based on the zero-coupon yield curve rather than the bond's own yield. After decades of neglect, this duration measure was revived by Fisher and Weil (1971) who demonstrated the relevance of the measure for constructing immunization strategies. We will refer to this duration measure as the **Fisher-Weil duration**. Assume that we are currently at time 0 and let y_i denote the current zero-coupon yield for a maturity in i periods. The Fisher-Weil duration of a bond with cash flow M_1, \dots, M_n is defined as

$$D_0^{\text{FW}} = \sum_{i=1}^n i w_i^{\text{FW}}, \quad w_i^{\text{FW}} = \frac{M_i(1+y_i)^{-i}}{B_0}. \quad (5.51)$$

Also these weights sum to one since the price of a coupon bond is given by

$$B_0 = \sum_{i=1}^n M_i(1+y_i)^{-i}, \quad (5.52)$$

cf. Eqs. (5.15) and (5.19). The weight w_i^{FW} has the interpretation of the relative weight of the present value of the i 'th payment, but now the payment M_i is discounted using the zero-coupon yield appropriate for date i instead of the bond's own yield y as in the definition of the Macaulay duration. If the yield curve is flat, the bond's own yield and all the zero-coupon yields are identical, in which case there is no difference between the Macaulay and the Fisher-Weil durations. For non-flat yield curves, the two durations are different, but typically quite close.

The bond price in (5.52) depends on all the zero-coupon yields y_1, y_2, \dots, y_n . If these yields change simultaneously, a first-order approximation of the total change in the bond price is given by

$$\Delta B_0 \approx \sum_{i=1}^n \frac{\partial B_0}{\partial y_i} \Delta y_i. \quad (5.53)$$

If the changes in the zero-coupon yields are proportional in the sense that³

$$\Delta y_i = (1+y_i)\delta, \quad i = 1, 2, \dots, n, \quad (5.54)$$

for some constant δ , we can write the right-hand side of (5.53) in terms of the Fisher-Weil duration as stated in the following theorem.

Theorem 5.6

Suppose that zero-coupon yields change proportionally as stated in (5.54). Then

$$\sum_{i=1}^n \frac{\partial B_0}{\partial y_i} \Delta y_i = -\delta B_0 D_0^{\text{FW}} \quad (5.55)$$

so that a first-order approximation of the relative bond price change in response to an immediate yield curve shift is

$$\frac{\Delta B_0}{B_0} \approx -\delta D_0^{\text{FW}}. \quad (5.56)$$

³Of course, the change Δy_i is not proportional to the yield y_i but rather proportional to $1+y_i$.

Proof

The sensitivity to a specific zero-coupon yield is captured by the partial derivative

$$\frac{\partial B_0}{\partial y_i} = -iM_i(1+y_i)^{-i-1}.$$

If yield curve shifts satisfy (5.54), we thus have

the zero-coupon yield y_i changes by Δy_i and none of the other zero-coupon yields changes, the bond price will change by approximately $\frac{\partial B_0}{\partial y_i}\Delta y_i$. Summing up over the changes in all the zero-coupon yields, the price change of the coupon bond is

$$\sum_{i=1}^n \frac{\partial B_0}{\partial y_i} \Delta y_i = - \sum_{i=1}^n iM_i(1+y_i)^{-i-1}(1+y_i)\delta = -\delta \sum_{i=1}^n iM_i(1+y_i)^{-i} = -\delta B_0 D_0^{\text{FW}}, \quad (5.57)$$

where the last equality is due to (5.51). From (5.53), we now get

$$\Delta B_0 \approx -\delta B_0 D_0^{\text{FW}},$$

from which (5.56) follows.

We see that, with proportional yield curve shifts, the relative bond price change is proportional to the bond's Fisher-Weil duration.

We can also introduce a **Fisher-Weil convexity**

$$C_0^{\text{FW}} = \sum_{i=1}^n i(i+1)w_i^{\text{FW}}, \quad (5.58)$$

which differs from the Macaulay convexity only with respect to the definition of the weights. Then a second-order expansion of the price will lead to

$$\frac{\Delta B_0}{B_0} \approx -\delta D_0^{\text{FW}} + \frac{1}{2}\delta^2 C_0^{\text{FW}}, \quad (5.59)$$

again assuming the proportional shift of the zero-coupon yield curve formalized by (5.54).

Relative to the Macaulay measures, the Fisher-Weil duration and convexity have several advantages. First, it makes good sense to compute the relative importance of the individual payments using the appropriate zero-coupon yield. Secondly, it can be shown that the Fisher-Weil duration and convexity of a portfolio are exactly given by a weighted average of the durations and convexities of the bonds in the portfolio, whereas this only holds as an approximation for the Macaulay counterparts.

However, the assumption of proportional shifts is not fully realistic since several other types of yield curve shifts are observed in real life. Moreover, for typical yield curves the Macaulay weights and the Fisher-Weil weights are very close, and hence the Macaulay duration and the Fisher-Weil duration are close, so it does not make much of a difference whether you use one or the other.

Some practitioners like to represent the yield curve by a few key yields or key rates, such as the 3-month yield, the 2-year yield, the 5-year yield, and the 10-year yield. For each bond and each key rate, we can define a duration in terms of the bond price sensitivity with respect to a change in the value of that particular key rate, assuming the other key

rates remain unchanged. By combining these **key rate durations**, the price sensitivity to any combination of key rate changes can be approximated. See Ho (1992) for more information.

Empirical studies have shown that a yield curve is generally well-described by its **level, slope, and curvature**. Yield curve changes can be seen as a result of changes in these three factors. The level can be represented by a short-maturity yield, say, the 1-month yield. The slope is the difference between a long-maturity yield (like the 10-year yield) and the short-maturity yield. The curvature is typically computed as two times a medium-term yield minus a short-term yield minus a long-term yield so that a positive curvature corresponds to a concave yield curve. We can measure the price sensitivity of each bond to each of the three factors and thus define **factor durations**. The factor sensitivities can either be (i) estimated from an empirical analysis of historical movements in factors and bond prices or (ii) derived theoretically in models that describe possible factor movements over time and how bond prices are related to the factors.

For risk management the focus should then be on the factor durations of the net position. For example, immunization (see below) would involve picking a portfolio that can match all the factor durations of the liabilities. Matching only the Macaulay duration is basically taking care only of the level factor. The position might still be exposed to changes in the slope or the curvature of the yield curve.

5.9 Immunization

Some individuals or corporate investors are investing in the bond market either to ensure that some future liabilities can be met or just to obtain some desired future cash flow. For example, a pension fund often has a relatively precise estimate of the size and timing of the future pension payments to its customers. For such an investor it is important that the value of the investment portfolio remains close to the value of the liabilities. Some financial institutions are even required by law to keep the value of the investment portfolio at any point in time above the value of the liabilities by some percentage margin.

A cash flow or portfolio is said to be immunized (against interest rate risk) if the value of the cash flow or portfolio is not negatively affected by any possible change in the term structure of interest rates. An investor who has to pay a given payment stream can obtain a perfectly immunized total position by investing in a replicating portfolio. For example, if an investor has to pay 10 million dollars in 5 years, he can invest in default-free 5-year zero-coupon bonds with a total face value of 10 million dollars. The present value of his total position is completely immune to interest rate movements. An investor who has a desired cash flow consisting of several future payments can obtain perfect immunization by investing in a portfolio of zero-coupon bonds that exactly replicates the cash flow.

In many cases, however, all the necessary zero-coupon bonds are neither traded on the bond market nor possible to construct by a static portfolio of traded coupon bonds. Or perfect cash flow matching may require investments in a large number of zero-coupon bonds. In these cases, the desired cash flow can be matched by constructing a dynamically rebalanced portfolio of a relatively small number of bonds.

Intuitively, we have to set up a portfolio with the same present value as the liability (desired cash flow) and the same sensitivity to interest rate changes so that the portfolio continues to have the same value as the liability after any possible change in interest rates. If we take duration as the measure of interest rate sensitivity, we want to *match present value and duration*. Given that duration is a first-order measure of the interest rate sensitivity, we refer to this as **first-order immunization**. To satisfy two conditions (matching present value and duration) we need a portfolio of (at least) two bonds.

Suppose that the liabilities have a present value of \bar{B} and a duration of \bar{D} . Take two bonds with durations D_1 and D_2 , respectively. We need $D_1 \neq D_2$. Let π_1 denote the portfolio weight of the first bond and $\pi_2 = 1 - \pi_1$ the portfolio weight of the second bond. Using the approximation formula (5.50) for the portfolio duration, we aim at satisfying the equation

$$\bar{D} = \pi_1 D_1 + \pi_2 D_2 = \pi_1 D_1 + (1 - \pi_1) D_2 = \pi_1 (D_1 - D_2) + D_2, \quad (5.60)$$

which we can solve for π_1 . Given the portfolio weights, we can determine the number of units of each bond that we have to buy to match the present value of the liabilities. We summarize the results in the following theorem.

Theorem 5.7

Given liabilities with present value \bar{B} and duration \bar{D} , a first-order immunization is obtained by a portfolio of any two bonds with durations $D_1 \neq D_2$. The portfolio weights of the two bonds are

$$\pi_1 = \frac{\bar{D} - D_2}{D_1 - D_2}, \quad \pi_2 = \frac{D_1 - \bar{D}}{D_1 - D_2}. \quad (5.61)$$

The number of units of each bond in the portfolio is

$$N_1 = \frac{\pi_1 \bar{B}}{B_1} = \frac{\bar{B}}{B_1} \frac{\bar{D} - D_2}{D_1 - D_2}, \quad N_2 = \frac{\pi_2 \bar{B}}{B_2} = \frac{\bar{B}}{B_2} \frac{D_1 - \bar{D}}{D_1 - D_2}, \quad (5.62)$$

where B_1 and B_2 are the prices of the two bonds.

Proof

Solving (5.60), we get the expression for π_1 in (5.61). The expression for π_2 then follows from

$$\pi_2 = 1 - \pi_1 = 1 - \frac{\bar{D} - D_2}{D_1 - D_2} = \frac{D_1 - \bar{D}}{D_1 - D_2}.$$

The value of the bond portfolio must equal \bar{B} , the value of the liabilities. Hence, we need to invest the amounts $\pi_1 \bar{B}$ in bond 1 and $\pi_2 \bar{B}$ in bond 2. The number of units of each bond is the amount invested in the bond divided by its price, which leads to (5.62).

The next example illustrates the procedure in a case with a single liability.

Example 5.9

A company has made a promise to pay \$1,000,000 in four years from now. The CFO of the company wants to immunize the interest rate risk on this liability by investing in an asset or portfolio with the same present value and duration as the liability. Suppose that the yield curve is currently flat at 3%. Obviously, a perfect immunization would be obtained by investing in four-year zero-coupon bonds with a total face value of \$1,000,000, but suppose such bonds are not available.

i	Bond 1				Bond 2			
	M_i	PV	w_i	$w_i \times i$	M_i	PV	w_i	$w_i \times i$
1	30	29.1262	0.0291	0.0291	60	58.2524	0.0481	0.0481
2	1030	970.8738	0.9709	1.9417	60	56.5558	0.0467	0.0934
3					60	54.9085	0.0454	0.1361
4					60	53.3092	0.0440	0.1761
5					60	51.7565	0.0428	0.2138
6					60	50.2491	0.0415	0.2490
7					60	48.7855	0.0403	0.2821
8					1060	836.7738	0.6912	5.5297
Sum		1000	1	1.9709		1210.5908	1	6.7284

Table 5.7: Price and duration calculations.

The table shows how to calculate the price and duration of the bonds in Example 5.9.

Instead, the CFO tries to set up a portfolio of a 3% bullet bond maturing in two years and a 6% bullet bond maturing in eight years, both having a single annual payment and a face value of \$1,000. The portfolio is constructed to match the liability's present value

$$\bar{B} = \$1,000,000 \times (1.03)^{-4} \approx \$888,487.05$$

and duration which is clearly $\bar{D} = 4$ years. The next step is to find the price and duration of each of the bonds. This is done in Table 5.7. The conclusion is that

$$B_1 = \$1,000.00, \quad D_1 \approx 1.9709, \quad B_2 = \$1,210.59, \quad D_2 \approx 6.7284.$$

Substituting into (5.61), we see that the portfolio weights must be

$$\pi_1 = \frac{4 - 6.7284}{1.9709 - 6.7284} \approx 0.5735, \quad \pi_2 = 1 - \pi_1 \approx 0.4265.$$

The total amount invested in the two-year bond has to be $\pi_1 \bar{B} = \$509,536.53$ which with a unit price of \$1,000 means a purchase of (approximately) 509.54 units of that bond. Likewise, an amount of $0.4265 \times \$888,487.05 = \$378,950.52$ must be invested in the eight-year bond which with a unit price of \$1,210.59 means a purchase of (approximately) 313.03 units of that bond.

If the yield curve is flat when the immunization portfolio is constructed and can only shift in the form of a parallel shift (i.e. to a new flat yield curve), then the first-order immunization works well. Note that when the initial yield curve is flat, there is no difference between a parallel shift and a proportional shift of the curve. In this case, the portfolio duration formula (5.60) is exact, cf. Theorem 5.5, and the Macaulay duration is an appropriate risk measure. If the curve shifts in a proportional manner, but the initial curve is not flat, the first-order immunization strategy typically works well, but a small deviation may occur because (5.60) is not exact. If the initial curve is not flat or may shift in non-proportional ways (or both), the first-order immunization strategy is less precise.

Example 5.10

Consider the setting of Example 5.9. Let us look at the performance of the immunization strategy over the first six months. Suppose that after six months the yield curve is still flat at 3%. The present value of the liability would then be $\$1,000,000 \times (1.03)^{-3.5} = \$901,715.87$. The price of the two bonds would be \$1014.89 and \$1228.62, respectively, implying that the value of the immunization portfolio would be

$$509.54 \times \$1014.89 + 313.03 \times \$1228.62 = \$901,715.87$$

and thus identical to the value of the liability as desired. If you calculate the duration-matching portfolio again, you will find exactly the same portfolio. Hence, in this situation you do not have to rebalance the portfolio to stay immunized.

Suppose instead that the yield curve after six months is flat at 4%. Then the values would be \$871,732.65 for the liability, \$1000.57 and \$1157.13 for each of the bonds, and thus \$872,040.99 for the immunization portfolio. As the portfolio value exceeds the liability value, the immunization has proved successful, in fact with a profit of \$308.34. As yields have changed, you need to rebalance the portfolio to stay immunized against further yield curve changes. Based on the new liability duration of 3.5 years and bond durations of 1.4706 and 6.1807, the appropriate portfolio weights have changed to 0.5691 and 0.4309. If you invest in 495.86 units of bond 1 and 324.59 units of bond 2, you will match both the duration and new present value of the liability. Hence, you have to sell 13.68 units of bond 1 and purchase additional 11.56 units of bond 2, which will give you net proceeds of \$308.34, exactly identical to the profit realized on the immunization over the first six months.

Finally, suppose that the yield curve after the first six months is no longer flat. The 3.5-year yield is still 3%, but the yield on bond 1 has dropped to 2% and the yield on bond 2 has increased to 4%. In this case, the value is \$901,715.87 for the liability (as above), \$1029.56 for bond 1, and \$1157.13 for bond 2. Hence, the value of the original immunization portfolio is only \$886,812.39, which is \$14,903.48 less than the value of the liability so the immunization was not fully successful. If, on the other hand, the yields were 4% for bond 1 and 2% for bond 2, the portfolio value would have been \$918,607.23 and thus above the value of the liability.

The above example shows that the portfolio needs to be rebalanced after a change in the yield curve. This is also the case at the payment dates of the bonds involved (or the liability stream), whether or not the yield curve has changed since the portfolio was originally constructed. For example, this is clearly the case when the shortest of the bonds mature and has to be replaced by a different (short-term) bond in the immunizing portfolio.

As the yield curve changes almost continuously in time, an immunizing bond portfolio should in theory be continuously rebalanced to make sure that (5.60) always holds. Of course, continuous rebalancing of a portfolio is not practically implementable nor desirable considering real-world transaction costs. If the portfolio is only rebalanced periodically, a perfect immunization cannot be guaranteed. The durations may be matched each time the portfolio is rebalanced, but between these dates the durations may diverge due to interest rate movements and the passage of time. With different durations the portfolio and the desired cash flow will not have the same sensitivity towards another interest rate change so the portfolio value may diverge from the value of the liabilities.

As shown in (5.31), the convexity is closely related to the sensitivity of the duration towards interest rate changes. If both the durations and the convexities of the portfolio and the liabilities are matched each time the portfolio is rebalanced, then the durations are more likely to stay close even after several interest rate changes. Therefore, *matching the convexities* should improve the effectiveness of the immunization strategy. We refer to the strategy of matching both present values, durations, and convexities as a **second-order immunization**. This requires a portfolio of (at least) three bonds. We have the following theorem:

Theorem 5.8

Given liabilities with present value \bar{B} , duration \bar{D} , and convexity \bar{C} , a second-order immunization is obtained by a portfolio of any three bonds with durations D_j and convexities C_j provided that $(D_3 - D_2)C_1 + (D_1 - D_3)C_2 + (D_2 - D_1)C_3 \neq 0$. The portfolio weights of the three bonds are

$$\pi_1 = \frac{(D_3 - D_2)\bar{C} + (\bar{D} - D_3)C_2 + (D_2 - \bar{D})C_3}{(D_3 - D_2)C_1 + (D_1 - D_3)C_2 + (D_2 - D_1)C_3}, \quad (5.63)$$

$$\pi_2 = \frac{(D_3 - \bar{D})C_1 + (D_1 - D_3)\bar{C} + (\bar{D} - D_1)C_3}{(D_3 - D_2)C_1 + (D_1 - D_3)C_2 + (D_2 - D_1)C_3}, \quad (5.64)$$

$$\pi_3 = \frac{(\bar{D} - D_2)C_1 + (D_1 - \bar{D})C_2 + (D_2 - D_1)\bar{C}}{(D_3 - D_2)C_1 + (D_1 - D_3)C_2 + (D_2 - D_1)C_3}. \quad (5.65)$$

The number of units of each bond $j = 1, 2, 3$ in the portfolio is $N_j = \pi_j \bar{B}/B_j$, where B_j is the price of bond j .

Proof

We are assuming that the portfolio duration and portfolio convexity are well approximated by weighted averages as stated in (5.50). Since $\pi_3 = 1 - \pi_1 - \pi_2$, durations and convexities will be matched if π_1 and π_2 are chosen such that

$$\begin{aligned}\bar{D} &= \pi_1 D_1 + \pi_2 D_2 + [1 - \pi_1 - \pi_2] D_3, \\ \bar{C} &= \pi_1 C_1 + \pi_2 C_2 + [1 - \pi_1 - \pi_2] C_3.\end{aligned}$$

We leave it to the reader to verify that these equations are satisfied when π_1 and π_2 are given by (5.63) and (5.64). Then substitute these values of π_1 and π_2 into $\pi_3 = 1 - \pi_1 - \pi_2$ to get (5.65).

If only durations are matched, the convexity of the portfolio and hence the effectiveness of the immunization strategy will be highly dependent on which two bonds the portfolio consists of. If the convexity of the investment portfolio is larger than the convexity of the liability, a big change (positive or negative) in the short rate will induce an increase in the net value of the total position. The downside is that if the short rate stays almost constant, the net value of the position will decrease. The converse conclusions hold in case the convexity of the portfolio is less than the convexity of the cash flow.

5.10 Bonds with risky payments

5.10.1 Floating rate bonds

The coupon rate of a floating rate bond is reset periodically over the life of the bond. Let us consider the most common floating rate bond, which is a bullet bond. The coupon rate effective for the payment at the end of one period is set at the beginning of the period at a rate equal to the current market interest rate for that period, which is the yield of a bond maturing at the end of the period.

Suppose the bond matures at time n , i.e., in n periods from now. At time $n - 1$, the coupon is reset to the one-period yield at that time, which we denote by $y_1(n - 1)$. If F denotes the face value, the bond payment at time n is thus $(1 + y_1(n - 1))F$. The present value at time $n - 1$ of that payment is the product of the payment and the appropriate discount factor over the following period, which is $(1 + y_1(n - 1))^{-1}$:

$$B_{n-1} = [(1 + y_1(n - 1))F] (1 + y_1(n - 1))^{-1} = F.$$

Hence, the bond is priced at par immediately after the last coupon reset date.

At time $n - 2$, the coupon is set for the following period equal to the one-period yield $y_1(n - 2)$ prevailing at that date. This implies a coupon payment of $y_1(n - 2)F$ at time $n - 1$. In addition, when purchasing the bond at time $n - 2$, the investor obtains a claim to the repayment of the face value and a yet undetermined coupon payment at time n . But already at time $n - 2$ we know that last coupon rate is set so that the terminal payment has a present value at time $n - 1$ equal to F . So at time $n - 1$ we receive $y_1(n - 2)F$ and have a claim worth F , giving a total value of $(1 + y_1(n - 2))F$. Discounting back one period using the appropriate discount rate $y_1(n - 2)$, we find that the bond value at time $n - 2$ is

$$B_{n-2} = [(1 + y_1(n - 2))F] (1 + y_1(n - 2))^{-1} = F.$$

Continuing this line of argumentation, we can conclude that the price of such a floating rate bond is equal to the face value immediately after any reset of the coupon. In between payment dates, the bond price may deviate from the face value if the yield for the period until the next payment date deviates from the current coupon rate, but typically the bond price stays close to par.

5.10.2 Corporate bonds and default risk

In the bond pricing formulas presented so far we have assumed that the bond issuer always delivers the promised payments on a timely basis. An issuer not delivering the promised payments is said to default on the bond. The risk that this may happen is referred to as **default risk** or **credit risk**. A default does not imply that the holder of that contract walks away empty-handed. The holder will typically get a **recovery payment** either in cash or in the form of a new claim on the issuer.

The assumption of no default on a bond is in most cases very reasonable if the bond is issued by the government or treasury department of a country having a limited public debt relative to the tax incomes or the GDP of the country. If a government has issued bonds and subsequently faces financial troubles, it may raise taxes, cut public spending, or—if the bonds are denominated in the domestic currency—print enough money so that it can honour its nominal debts. Of course in the latter case the purchasing power of the money received by the bondholders may be lower than expected. It should be noted however that there are plenty of historical examples of countries defaulting on part or all

of their debt.⁴

For many bonds issued by private corporations it is necessary to take default risk into account. The prime example is corporate bonds, where a firm has borrowed money by issuing bonds promising a prespecified future payment stream. For various reasons the firm may end up in a situation where it cannot or will not continue paying the promised amounts to bondholders and therefore the firm defaults on its debt. Of course, when potential investors value a corporate bond they anticipate the possibility of a default of the issuing firm before the maturity date of the bond.

It is intuitively clear that, other things equal, a higher risk of default leads to a lower price. If we compute the yield of the bond from equating the market price of the bond and the sum of the discounted promised payments, a higher default risk implies a higher yield. The so-called **credit spread** is the difference between the yield on such a defaultable bond and the yield on a default-free bond having the same (or at least very similar) promised payment stream.

As a simple example, consider a bullet bond maturing in one period. The promised payment is $(1 + q)F$, where F is the face value and q is the coupon rate. Suppose that the probability that the issuer defaults in that period is $p \in [0,1]$. If default occurs, the bondholder only receives a recovery payment equal to a fraction $R \in [0,1]$ of the promised payment. With a probability of $1 - p$, the bondholder receives the full promised payment. If we discount the expected payment using the one-period yield y_1 on a default-free bond, the price of the defaultable bond becomes

$$\begin{aligned}\tilde{B}_0 &= [pR(1 + q)F + (1 - p)(1 + q)F](1 + y_1)^{-1} \\ &= (1 + q)F[1 - p(1 - R)](1 + y_1)^{-1}.\end{aligned}\quad (5.66)$$

The term $p(1 - R)$ is the expected loss due to the default as a fraction of the promised payment. On the other hand the yield \tilde{y}_1 on the defaultable bond is defined so that

$$\tilde{B}_0 = (1 + q)F(1 + \tilde{y}_1)^{-1}.$$

Aligning the right-hand sides of the two preceding expressions, we conclude that

$$(1 + \tilde{y}_1)^{-1} = [1 - p(1 - R)](1 + y_1)^{-1},$$

which implies that

$$p(1 - R) = 1 - \frac{1 + y_1}{1 + \tilde{y}_1} = \frac{\tilde{y}_1 - y_1}{1 + \tilde{y}_1} \approx \tilde{y}_1 - y_1 = s_1, \quad (5.67)$$

where s_1 is the one-period credit spread. The credit spread is thus (approximately) the product of the default probability and the fractional loss in case of default. Similar relations can be shown to hold for bonds of longer maturities.

Good models of corporate bond valuation are quite complex, however, since they have to model both the possible variations in future default-free interest rates as well as default

⁴Tomz and Wright (2007) report that 106 countries have defaulted a total of 250 times in the period 1820–2004. Recent examples of such defaults include the Russian government's default on the domestically issued GKO bonds in August 1998 and Argentina's default on USD 142 billion of domestic public debt in 2001 and roughly USD 1 billion debt to the World Bank in 2002. Government debt defaults are often due to a period of bad performance of the domestic economy, but Tomz and Wright (2007) point out that there are also many examples of countries defaulting in “good times,” for example, following a major change in the political regime in the country.

Moody's	S&P and Fitch
Aaa	AAA
Aa	AA
A	A
Baa	BBB
Ba	BB
B	B
Caa	CCC
Ca	CC
C	C

Table 5.8: Credit ratings.

The table shows the rating categories of the main credit rating agencies. The best rating is the upper (Aaa, AAA) reflecting a very small risk of default. Lower ratings reflect increasingly higher default risks.

probabilities and recovery rates. Moreover, many corporate bonds are callable so that the issuing corporation has the right (option) to buy back the bonds at some pre-specified price, which is usually the sum of the outstanding debt and some premium often depending on the remaining time to maturity. Such an option is particularly valuable when current market interest rates are below the coupon rate of the bonds, because then the issuer can buy back the high-coupon debt and issue new debt with a lower coupon. The benefits have to be compared to the call premium and any costs involved in the transactions.

The call feature complicates corporate bond valuation even further and also implies that standard measures of interest rate risk can be misleading. In particular, when market interest rates are low the price of a callable bond may react differently to a further decline in interest rates than a similar non-callable bond.

Rating agencies such as Standard & Poor's, Moody's Investor Service, and Fitch evaluate the credit quality of many bonds and assign a corresponding **credit rating** to each such bond. Table 5.8 lists the primary rating categories of the main agencies. Some of the rating categories are sometimes subdivided into finer groups either by adding a plus or a minus (S&P and Fitch) or by adding a 1, 2, or 3 (Moody's). Securities with a rating in the upper four categories are called investment grade securities, while those with a lower rating are often referred to as non-investment grade, speculative grade, or *junk* securities. Ratings for short-term bonds are often given according to slightly different categorizations.

Ratings are assigned to bonds of all sorts of issuers, including national and local governments, private corporations, and non-profit organizations. Note that bonds of the same issuer may have different terms (maturity, priority, etc.) that make them more or less credit risky and they can therefore have different ratings. Ratings are also routinely assigned to so-called collateralized debt obligations (CDOs) as well as to mortgage-backed bonds and collateralized mortgage obligations (CMOs).

Table 5.9 indicates that the credit ratings on average provide a reasonable ranking of default risk. The historical default frequencies are decreasing in the rating, i.e., higher for lower rated companies and close to zero for top-rated companies. However, the credit rating agencies have been criticized for being too slow in downgrading companies. An example, probably the most severe, is Enron that remained investment grade until 4 days before its 2001 bankruptcy, although its financial problems were apparently well-known to the agencies long before. The lag in rating changes can at least partially be explained by the fact that credit rating agencies take a “through-the-cycle” perspective so that credit ratings are set to reflect long-term default risk rather than the short-term default

	Maturity (years)								
	1	2	3	4	5	7	10	15	20
Aaa	0.000	0.013	0.013	0.037	0.107	0.250	0.508	0.955	1.139
Aa	0.017	0.054	0.087	0.157	0.234	0.388	0.551	1.074	2.194
A	0.025	0.118	0.272	0.432	0.612	1.025	1.752	3.111	5.102
Baa	0.164	0.472	0.877	1.356	1.824	2.770	4.397	8.009	11.303
Ba	1.113	2.971	5.194	7.523	9.639	13.263	18.276	27.220	34.845
B	4.333	9.752	15.106	19.864	24.175	32.164	41.088	52.190	56.101
Caa-C	16.015	25.981	34.154	40.515	45.800	52.702	63.275	68.873	70.922
Inv. Grade	0.068	0.215	0.416	0.651	0.894	1.399	2.237	3.966	5.952
Spec. Grade	4.113	8.372	12.467	16.093	19.245	24.520	30.637	39.343	45.498
All rated	1.401	2.844	4.193	5.360	6.344	7.938	9.802	12.608	15.125

Table 5.9: Default rates and ratings.

The table shows the average cumulative global default rates (in percent) based on Moody's ratings for 1970-2008. Source: Emery, Ou, Tennant, Matos, and Cantor (2009).

probability which might be more relevant to some investors. There is some empirical evidence that the market-determined credit spread increases *before* the rating of the bond is downgraded when the credit quality of a bond-issuing company deteriorates.

Rating agencies are sometimes accused of having too close relations to the companies they are rating, which might generate conflicts of interest questioning their objectivity. Rating agencies are paid by the issuing companies and can be tempted to provide ratings that please issuers rather than serve the investors and regulatory bodies relying on the ratings. And if an agency tells an issuer that it will be poorly rated, the issuer might take its business to a less pessimistic (or less honest) rating agency. In the recent financial crisis (starting around 2007) the major credit rating agencies have been exposed to harsh critique, in particular for their high original ratings of structured products such as collateralized debt obligations and collateralized mortgage obligations of which many have later been substantially downgraded or even defaulted. Rating structured products became a very profitable business for the credit rating agencies who often worked together with the issuer to structure the product such that the agency was willing to give a certain high rating. Independence between the credit rating agency and the issuer was illusory. Reforms of the legislative basis of the credit rating industry are currently being discussed and implemented both in the U.S. and Europe. The reforms involve, among other things, mandatory disclosure of rating practices and conflicts of interest as well as tightened supervision by governmental authorities.

5.10.3 Mortgage-backed bonds

A mortgage is a loan offered by a financial institution to the owner of a given real estate property that serves as collateral for the loan. In many countries, a mortgage is the standard way to finance a large part of the purchase of residential property. Traditionally mortgages are long-term (often 30 years when initiated) annuity loans where the initial debt is gradually paid back in such a way that the sum of the repayment and the interest payment are the same for all scheduled payment dates. Numerous alternative mortgages have been developed in recent decades, however.

In some countries mortgages are typically financed by the issuance of bonds. A large

number of similar mortgages are pooled either by the original lending institution or by some other financial institution. The pooling institution issues bonds with payments that are closely linked to the payments on the underlying mortgages. Afterwards, the bonds are traded publicly. Large markets for mortgage-backed bonds exist in the United States and Denmark, for example.

Mortgage-backed bonds differ from government bonds in several respects. Most importantly, the cash flow to the bond owners depends on the payments that borrowers make on the underlying mortgages. While a mortgage specifies an amortization schedule, most mortgages allow the borrower to pay back the debt earlier than scheduled. A big challenge in the valuation of mortgage-backed bonds is to predict or model the fraction of the total outstanding debt in the corresponding mortgage pool that will be prepaid in a given period. Obviously, the current refinancing rate is central, but factors such as the slope of the yield curve and the development in real estate prices are also relevant.

Of course, some homeowners may default and not make the promised payments on their mortgage. For the holders of the mortgage-backed bonds it is important to know how that will affect the payments on their bonds. Often the financial institution issuing the bonds or some other institution will guarantee the payments to the bondholders and basically chip in any missing mortgage payments.

Because of the prepayment options, quite complicated models are necessary to compute prices of mortgage-backed bonds. Some insights are possible without specific models, however. When current long-term rates are high relative to the coupon rate of a pool of existing mortgages, it is unlikely that many of these mortgages will be prepaid right now. In this case, the pricing of the corresponding mortgage-backed bonds is similar to the pricing of a government bond with the same promised payment schedule. In contrast, when current long-term rates are significantly lower than the coupon rate of the mortgage pool, most of the mortgages are expected to be prepaid soon so that the bondholders can expect to get roughly the face value (with any interest since the previous payment date) right away. The price of the mortgage-backed bond is then close to the face value.

It is unlikely that the price of the mortgage-backed bond can exceed the face value by a significant margin, because that would require current interest rates to be much lower than the coupon rate and then massive prepayments are expected. A further reduction in current interest rates can only lead to a very small increase in the bond price. Hence, the usual convex relation between interest rates and bond prices breaks down for low interest rates in the case of mortgage-backed bonds and other callable bonds. Such bonds can have a negative convexity when the current interest rate is below the coupon rate. It also follows that the duration of a mortgage-backed bond differs from that of a similar government bond for relatively low market interest rates. Duration and convexity measures of mortgage-backed bonds have to be computed from pricing models that take prepayment behavior into account. For such bonds it should be clear that the relevant risk measure is not the price sensitivity towards the bond's own yield, but rather the price sensitivity with respect to the level of market interest rates (and maybe the slope and the curvature of the yield curve).

5.11 Bond portfolio management

Some investors set up their bond portfolio to match certain future liabilities using some sort of immunization approach as explained in Section 5.9. More generally, by investing in one or more default-free bonds investors can ensure a minimum cash flow stream in the future. But how should bonds enter the portfolio of investors who want to maximize their risk-return tradeoff over a given investment horizon? How should households saving for

retirement invest in bonds?

Government bonds as an asset class is often seen as providing a good diversification with stocks, because bond prices frequently go up in severe stock market downturns in which government bonds are considered a safe haven (the *flight to quality* mechanism). However, the stock-bond correlation varies considerably over time as we illustrate in Chapter 6, see, for example, Figure 6.13 on page 257.

Within the class of government bonds there is no point in spreading your investment over a large number of bonds. As explained earlier, government bonds offer small, if any, diversification gains as they are all sensitive to the same factors. Among government bonds it should normally suffice to include a short-term, a medium-term, and a long-term bond in the portfolio. In that way the investor can adjust his exposure to the level, slope, and curvature factors and pick up any risk premium related to those factors. If you are looking for a portfolio with a specific sensitivity to these factors, observe that most stock prices are also sensitive to the same factors (and more). The interest rate sensitivity of stock prices is yet not fully explored, neither theoretically nor empirically.

Along the same lines, you may consider investing in a few representative mortgage-backed bonds of different maturities and having different exposures to prepayment risk, as well as a few representative corporate bonds of different maturities and different levels of default risk as indicated for example by their credit ratings. For households the optimal investments in general, and the optimal bond position in particular, should be seen in connection with the magnitude and risk characteristics of the human capital of the household as well as any real estate investments and related mortgages. The research of these issues is still developing, and we discuss some results in Chapter 8.

Some market analysts and portfolio managers believe they can identify bonds that are mispriced relative to other bonds. Once in a while, the arbitrage relations between prices may temporarily break down and the alert trader may seize a profit. But such opportunities are rare.

Other analysts and portfolio managers believe they can predict movements in the yield curve. If you believe that yields are going to decline, you should shift to bonds with a higher duration since their prices will increase a lot when yields go down. Conversely, if you believe yields are going to increase, you may want to eliminate your bond portfolio or at least reduce its duration to limit your loss when yields increase. However, since markets are generally very efficient, you will only profit if you have superior information and superior forecasting skills relative to the market. Also note that interest rate forecasters have a notoriously poor track record.

5.12 Concluding remarks

Fabozzi (2010) has more information about basic bond market concepts and relations as well as the structure of the U.S. bond market. Batten, Fetherston, and Szilagyi (2004) provide an overview of various European markets for fixed income. To price bonds, bond derivatives, and other interest rate derivatives, analysts and traders use fairly advanced mathematical models of how interest rates vary over time. Such models are also useful for deriving measures of interest rate risk for different securities. Numerous models have been suggested in the academic literature since the early 1970's. There are many books surveying such models including Brigo and Mercurio (2006), Veronesi (2010), and—my personal favorite—Munk (2011).

5.13 Exercises

Exercise 5.1. Consider the following default-free bullet bonds:

Price	Face value	Coupon rate	Maturity
71.5805	100	0%	10 years
100	100	3%	10 years
121.7326	100	6%	10 years

Is there any arbitrage opportunity in the market? If yes, construct the arbitrage strategy and compute the profit.

Exercise 5.2. In the Kingdom of Far Far Away two bonds are currently traded. Both have an initial face value of \$100, a single annual payment, and mature in exactly two years from now. The first bond is a 7% bullet bond traded at a price of \$98.10. The second bond is a 10% serial bond traded at a price of \$101.70.

- (a) Show how you from these bonds can construct a zero-coupon bond maturing in one year and a zero-coupon bond maturing in two years.
- (b) What are the one- and two-year zero-coupon yields, that is y_1 and y_2 ?
- (c) What are the forward rates over the first and the second year, that is $f_{0,1}$ and $f_{1,2}$?

The Far Far Away bond exchange is now opening up for trade in a third bond, namely a 2% bullet bond maturing in two years, also having a single annual payment and a face value of \$100.

- (d) What do you expect the price of the new bond to be? Why?
- (e) Suppose the new bond will trade at a lower price than the one you expect. What do you do?

Exercise 5.3. Consider a market where a number of zero-coupon bonds with a face value of 1000 are traded. The zero-coupon yields for maturities from 1, 2, 3, 4, and 5 years are:

$$y_1 = 0.5\%, \quad y_2 = 1\%, \quad y_3 = 2\%, \quad y_4 = 3\%, \quad y_5 = 4\%.$$

- (a) What are the prices of the zero-coupon bond maturing in 1, 2, 3, 4, and 5 years?

Bullet bonds with various maturities are also traded. They all have a face value of 1000 and a coupon rate of 5%.

- (b) What are the no-arbitrage prices of such bullet bonds maturing in 1, 2, 3, 4, and 5 years?
- (c) What are the corresponding yields on these bullet bonds?

Furthermore, annuity bonds with various maturities are also traded. They also have an initial face value of 1000 and a coupon rate of 5%.

- (d) What are the no-arbitrage prices of such annuity bonds maturing in 1, 2, 3, 4, and 5 years?
- (e) What are the corresponding yields on these annuity bonds?
- (f) For each bond type (zero-coupon, bullet, annuity) illustrate how the yield depends on the maturity date of the bond. Compare the yields of the different bond types.

Exercise 5.4. Suppose that you can trade in two bullet bonds, both having a face value of 100, exactly two years until maturity, and one payment date per year. The first bond has a coupon rate of 5% and trades at a price of 100.0458. The second bond has a coupon rate of 8% and trades at a price of 105.6515.

- (a) Construct a portfolio of the two bonds that replicates a one-year zero-coupon bond with a face value of 100. What is the one-year zero-coupon yield y_1 ?
- (b) Construct a portfolio of the two bonds that replicates a two-year zero-coupon bond with a face value of 100. What is the two-year zero-coupon yield y_2 ?

- (c) What is the forward rate $f_{0,1}$ over the first year? What is the forward rate $f_{1,2}$ over the second year?

Exercise 5.5. Consider a 5% bullet bond maturing in exactly 10 years from now. It has a single annual payment and a face value of \$1000. The yield of the bond is currently 3%. You plan to buy the bond today and resell it after a year, immediately after receiving the coupon payment.

- (a) What is your holding-period return going to be if the yield of the bond one year from now is still 3%?
- (b) What is your holding-period return going to be if the yield of the bond one year from now is 1%?
- (c) What is your holding-period return going to be if the yield of the bond one year from now is 5%?

Exercise 5.6. Imagine a market where two bonds are traded. Both have a face value of \$1000 and a single annual payment. The first bond is a 5% bullet bond maturing in one year and is traded at \$1034.48. The second bond is a 4% bullet bond maturing in two years and is traded at \$1019.71. Assume that fractions of bonds can be traded.

- (a) Show how you from these bonds can construct a zero-coupon bond with face value \$1000 maturing in one year and a zero-coupon bond with face value \$1000 maturing in two years.
- (b) What are the one- and two-year zero-coupon yields, that is y_1 and y_2 ?
- (c) What are the forward rates over the first and the second year, that is $f_{0,1}$ and $f_{1,2}$?

Suppose now that a third bond is introduced in the market. It is a 6% annuity bond maturing in two years, also having a single annual payment and a face value of \$1000.

- (d) What is the unique no-arbitrage price and the corresponding yield of the annuity bond? Set up an arbitrage strategy involving the two bullet bonds and the annuity bond if the annuity bond trades at a higher price.

Exercise 5.7. You can invest in the following government bonds:

Price	Par value	Coupon rate	Time to maturity
82.2703	100	0%	4 years
100	100	5%	5 years
96.4541	100	4%	4 years

Coupons are paid annually.

- (a) If you believe that interest rates will fall later today, in which bond should you invest now?
- (b) If the fall in interest rates is 0.2%, what is the *approximate* return on the 5% coupon bond?

Exercise 5.8. You are thinking about investing \$2,000,000 in a portfolio of the following two bonds:

1. 2% bullet bond maturing in 3 years from now, one annual payment date, face value of \$1,000, current price \$1,020
2. 5% bullet bond maturing in 10 years from now, one annual payment date, face value of \$1,000, current price \$1,050.

Each bond would have a weight of 50% in your portfolio.

- (a) For each bond compute the yield, duration, and convexity.
- (b) For the portfolio compute the yield, duration, and convexity using the approximations in Section 5.8.3.
- (c) Determine the payment schedule of the portfolio. Based on this schedule, determine the exact yield, duration, and convexity of the portfolio. Compare with the approximate values computed in the preceding question.

Exercise 5.9. A company has made a promise to pay \$4,000,000 in five years from now. The CFO of the company wants to immunize the interest rate risk on this liability by investing in a portfolio consisting of a 4% bullet bond maturing in four years and a 5% bullet bond maturing in eight years, both having a single annual payment and a face value of \$1000. Assume that the yield curve is currently flat at 3%.

- (a) What is the present value and the duration of the liability?
- (b) What is the price and the duration of each of the two bonds?
- (c) Exactly which portfolio of the two bonds will immunize the interest rate risk on the liability?
- (d) Can you use the same portfolio until the liability is due? Why or why not?

Exercise 5.10. A company has made a promise to pay \$500,000 in three years from now. The CFO of the company wants to immunize the interest rate risk on this liability by investing in an asset or portfolio with the same present value and duration as the liability. Suppose that the yield curve is currently flat at 4%.

- (a) What is the present value and the duration of the liability?
- (b) It would be sufficient for the company to invest in a single bond (if this is traded). Which bond?

Unfortunately, the bond considered in (b) is not traded. Instead, the CFO will try to set up a portfolio of a 6% bullet bond maturing in two years and a 5% bullet bond maturing in four years, both having a single annual payment and a face value of \$100.

- (c) Find the present value and the duration of each of the bonds.
- (d) Exactly which portfolio of the two bonds will immunize the interest rate risk on the liability?
- (e) Can you use the same portfolio until the liability is due?

Exercise 5.11. Consider a 7% bullet bond maturing in 10 years. It has a single annual payment and a face value of \$1000. The yield of the bond is currently 4%.

- (a) Compute the price, duration, and convexity of the bond.
- (b) Find the exact price change if the yield immediately changes by -2%, -1%, 1%, and 2%, respectively.
- (c) Find the approximate price change for the same yield changes based on an approximation using the duration only.
- (d) Find the approximate price change for the same yield changes based on an approximation using the duration and the convexity.

Exercise 5.12. The company V-O-U has made a promise to pay \$5,000,000 in six years from now. Harry Hedger, the CFO of the company, wants to immunize the interest rate risk on this liability by investing in a portfolio consisting of a 3% bullet bond maturing in three years and a 5% bullet bond maturing in ten years, both having a single annual payment and a face value of \$1000. Assume that the yield curve is currently flat at 2% per year.

- (a) What is the present value and the duration of the liability? What is the price and the duration of each of the two bonds?
- (b) Exactly which portfolio of the two bonds immunizes the interest rate risk on the liability?

Suppose Harry has invested in the immunization portfolio found in (b) and holds on to that over the next year. Now consider the situation after that year has passed so that there are five years remaining until the promised payment is due.

- (c) Suppose that the yield curve after the first year is flat at 3% per year. What is the value at that point in time of the liability and of the bond portfolio (including any coupon payments received)? Has the immunization strategy been successful so far? Should Harry rebalance the portfolio at this point?
- (d) Suppose that the yield curve after the first year is increasing so that the yield on the 3% bullet bond is 1%, the five-year zero-coupon yield is 3%, and the yield on the 5% bullet bond is 5%. What is then the value at that point in time of the liability and of the bond portfolio (including any coupon payments received)? Has the immunization strategy been successful so far? Why or why not?

Exercise 5.13. Suppose that the current zero-coupon yield curve includes the following values:

Years to maturity, n	1	2	3	4	5
Zero-coupon yield, y_n	0.5%	1.0%	1.5%	2.0%	2.5%

- (a) Consider a bullet bond with a face value of \$1000, a coupon rate of 2%, two years to maturity, and one annual payment date. State the payment schedule of the bond. Compute its price. Determine the yield-to-maturity of the bond. Compute the duration of the bond.
- (b) Consider a bullet bond with a face value of \$1000, a coupon rate of 4%, five years to maturity, and one annual payment date. State the payment schedule of the bond. Compute its price. Determine the yield-to-maturity of the bond. Compute the duration of the bond.

Now suppose that suddenly the zero-coupon yield curve changes to the following:

Years to maturity, n	1	2	3	4	5
Zero-coupon yield, y_n	4.5%	4.0%	3.5%	3.0%	2.5%

- (c) By how much does the price of each of the two bonds considered above change? What is the new yield-to-maturity of each of the two bonds? Discuss your findings in relation to the durations of the two bonds computed above.

Exercise 5.14. Suppose that the current zero-coupon yield curve includes the following values:

Years to maturity, n	1	2	3	4	5	6
Zero-coupon yield, y_n	0.67%	0.99%	1.26%	1.48%	1.65%	1.81%

- (a) Consider a bullet bond with a face value of \$1000, a coupon rate of 1%, two years to maturity, and one annual payment date. State the payment schedule of the bond. Compute its price. Determine the yield-to-maturity of the bond. Compute the duration of the bond.
- (b) Consider a bullet bond with a face value of \$1000, a coupon rate of 3%, six years to maturity, and one annual payment date. State the payment schedule of the bond. Compute its price. Determine the yield-to-maturity of the bond. Compute the duration of the bond.
- (c) Suppose you have issued a guarantee to pay \$1,000,000 in five years from now. What is the present value of the guarantee? Which portfolio of the two-year bullet bond and the six-year bullet bond should immunize the guarantee?

Now suppose that the zero-coupon yield curve immediately changes to a flat yield curve so that $y_n = 2\%$ for $n = 1, 2, \dots, 6$.

- (d) By how much does the price of each of the two bonds considered above change? Discuss your findings in relation to the durations of the two bonds computed above.
- (e) What is now the value of the guarantee? What is the value of the immunization portfolio constructed in question (c)? Has the immunization worked successfully? Discuss your findings.

Exercise 5.15. Give a mathematical proof of Eq. (5.31).

Exercise 5.16. Suppose that the current zero-coupon yield curve includes the following values:

Years to maturity, n	1	2	3	4	5
Zero-coupon yield, y_n	1%	2%	3%	2%	1%

All the bonds considered in the following are assumed to be bullet bonds with a face value of \$1,000 and a single annual payment, and they have exactly one year to the next payment date.

- (a) State the cash flow of a 5-year bond with a coupon rate of 3%. What is the price, the yield, the duration, and the convexity of the bond?

Suppose that you have issued a guarantee to pay \$1,000,000 in three years from now.

- (b) What is the present value, the duration, and the convexity of the guaranteed payment?

You are worried that the present value of the guarantee might increase, so you intend to immunize the risk by investing in a bond portfolio consisting of the 5-year bond studied above as well as one of the following bonds:

Maturity and coupon	Price	Yield	Duration	Convexity
2-year 2% bond	\$1000.19	1.9900%	1.9804	5.9216
4-year 4% bond	\$1075.46	2.0175%	3.7841	18.5632

In the following, calculate the duration of a portfolio as a weighted average of the durations of the bonds in the portfolio. Similarly for the convexity.

- (c) Determine the portfolio of the 2-year bond and the 5-year bond that matches the present value and the duration of the guarantee.

Now assume that immediately after you invested in the portfolio, the zero-coupon yields change to the following values

Years to maturity, n	1	2	3	4	5
Zero-coupon yield, y_n	3%	2%	1%	2%	3%

- (d) What is now the present value of the guarantee? What is the value of the bond portfolio? Was the immunization successful? Explain why or why not.
(e) Determine the portfolio of all three bonds mentioned above that would have matched the present value, the duration, and the convexity of the guarantee before the change in the yield curve.
(f) After the change in the yield curve, what would be the value of the bond portfolio determined in the previous question? Would that portfolio have led to a successful immunization? Explain why or why not.

Exercise 5.17. Your company has made a promise to pay \$1,000,000 on April 1, 2026.

On April 1, 2022 you decide to immunize the interest rate risk on this liability using a portfolio of two bullet bonds. Bond 1 is a two-year bond with a coupon rate of 4%. Bond 2 is a six-year bond with a coupon rate of 2%. Both bonds have a face value of \$1,000, and they have annual payments on April 1 starting in 2023. The zero-coupon yield curve on April 1, 2022 is given by $y_n = 1\% + 0.2\% \times n$, so that the one-year zero-coupon yield is 1.2%, the two-year zero-coupon yields is 1.4%, etc. Make the following calculations assuming that the current date is April 1, 2022.

- (a) What is the present value and the duration of the liability?
(b) What is the price and the duration of bond 1? What is the price and the duration of bond 2?
(c) Which portfolio of the two bonds immunizes the interest rate risk on the liability?

Suppose now that the date is October 1, 2022 so that six months have passed since you set up the immunization portfolio. Now you want to evaluate whether your immunization portfolio has been successful so far. This depends on what the zero-coupon yield curve is on October 1, 2022.

- (d) Suppose the zero-coupon yield curve on October 1, 2022 is still given by $y_n = 1\% + 0.2\% \times n$. For example, this means that the zero-coupon yield for payments at April 1, 2023 (i.e. half a year later) is $1\% + 0.2\% \times 0.5 = 1.1\%$. What is now the present value of the liability? What is the value of the immunization portfolio? Has the immunization strategy been successful so far?

- (e) Instead, suppose the zero-coupon yield curve on October 1, 2022 is given by $y_n = 1\% + 0.6\% \times n$. For example, this means that the zero-coupon yield for payments at April 1, 2023 is $1\% + 0.6\% \times 0.5 = 1.3\%$.

What is now the present value of the liability? What is the value of the immunization portfolio? Has the immunization strategy been successful so far? Briefly compare with your conclusion in Question (d) and discuss your findings.

Exercise 5.18. In the bond market, four bullet bonds are traded, all with a face value of \$1000, annual payments, and exactly one year until the next payment. You have the following information about the bonds:

Bond number	Maturity (years)	Coupon rate	Price
1	1	2%	\$1004.93
2	2	3%	\$1021.50
3	3	4%	\$1052.20
4	4	0%	\$909.50

- (a) Determine the yield and the duration of each of the four bonds.
- (b) Determine a portfolio of the 1-year, 2-year, and 3-year bullet bonds that replicates a 3-year zero-coupon bond with a face value of \$1000. What is the price of the portfolio? What is the 3-year zero-coupon yield?
- (c) A new bond is introduced in the bond market. Also this bond has annual payments. It pays \$250 after one year and after two years, and then it pays \$300 after three years and after four years. What is the no-arbitrage price of this new bond? Explain. If the new bond is traded at a price which is \$5 lower than the no-arbitrage price, which strategy would generate an arbitrage profit?
- (d) Later today, the Central Bank is making an announcement that might cause a significant and immediate change in bond yields. You believe that the change in the yield of the 4-year bullet bond is normally distributed with a mean of 0.1 percentage points and a standard deviation of 0.2 percentage points. What do you expect the price of the 4-year bullet bond to be at the end of the day? Determine a 95% confidence interval for the end-of-day price of the 4-year bullet bond.

CHAPTER 6

Stocks

A stock is issued by a company and gives its holder rights to a share of specific future payouts made by the company as well as rights to vote on matters of corporate policy. Most stocks are subsequently listed and traded at one or more stock exchanges. Stocks and stock markets were introduced in Section 1.2. Stock markets are huge and offer vast opportunities for both short- and long-term investments.

The first part of this chapter explains the basic theoretical approaches to valuing stocks. Section 6.1 shows how to price a stock by calculating the present value of the future dividends paid to the stock owner, whereas Section 6.2 links stock prices to the earnings or the free cash flows of the issuing company. Section 6.3 introduces a decomposition of stock returns, whereas Section 6.4 generalizes the duration introduced for bonds in the previous chapter to the case of stocks.

The second part of the chapter presents key properties of historical stock returns. Section 6.5 considers the overall stock market and discusses the historical average, variability, and distribution of returns, as well as the question whether or not future stock market returns are predictable. Section 6.6 provides empirical evidence of systematic differences in the average return and the return variability across stocks, whereas Section 6.7 zooms in on individual stocks. Section 6.8 presents empirical correlations both between different stocks and between different asset classes. Finally, Section 6.9 provides an illustration of how much risk a stock investor can diversify away by investing in many different stocks rather than a few stocks.

6.1 The dividend discount model

By purchasing a share of stock in a given company, an investor obtains the right to receive her proportionate share of the future dividend payments made by the company. Hence, it seems reasonable to value the stock by computing the present value of all future dividends. This is the idea of the *dividend discount model* which can be seen as an application of the more general *discounted cash flow model* to the case where the cash flows being discounted are dividends. We can either think of this at the company level with dividends representing the total cash flows floating from the company to all its shareholders or at the level of an individual share of stock with the dividends representing the cash flow per share. Or we could even take a market view and relate the value of the stock market to the total dividends paid out by all listed companies.

What are dividends? The traditional way for companies to pay dividends is in form of an *ordinary cash dividend* where shareholders receive a certain cash amount per share they hold. Such dividends are typically paid at a quarterly or an annual frequency. But money can flow from companies to shareholders in other ways.

Instead of paying cash dividends, the company can spend the same amount on *stock repurchases* by buying back some of its own shares either directly on the stock exchange or through a fixed-price tender offer to existing shareholders. Other things equal, the reduced supply of shares causes the price per share to increase to the benefit of the shareholders. (Sometimes the stock price drops if investors interpret a stock buyback as a negative signal about the company's future earnings.) While shareholders generally have to pay immediate taxes on cash dividends, the taxation of capital gains can often be deferred until the gains are realized through a stock sale, and for some investors capital gains are taxed at a lower rate than cash dividends. Company managers may have a personal preference for buy-backs over cash dividends if they hold call options on the stocks of the company as part of their compensation package or if their compensation is increasing in the earnings per share of the company. Stock repurchases were uncommon in the U.S. until the 1980's but have since 1997 dominated cash dividends in terms of the total cash flow received by shareholders (Zeng and Luk 2020). From 1980 to 2018, the proportion of cash dividend-paying U.S. companies decreased from 78% to 43%, whereas the proportion of companies implementing stock repurchases increased from 28% to 53%. The flip side of a stock buyback is a *stock issuance* where a company issues new shares of stock in return for cash which can be seen as a negative dividend payment by the company.

A *liquidation dividend* can be paid by a company to its shareholders in case of a partial or full liquidation of the company which may occur if the company is insolvent or terminates (some of) its business activities for other reasons. A form of liquidation dividend is seen in mergers and acquisitions (M&As). If firm A acquires firm B by buying the shares of firm B at \$10 per share, this corresponds to the shareholders of firm B receiving a \$10 liquidation dividend. In periods with higher M&A activity, a large part of the total cash flows from companies to shareholders stems from such M&A liquidation dividends.

We first look at what we can say about the stock price without imposing any structure on how dividends change from period to period. Afterwards, we discuss various special cases with such a structure imposed, for example by assuming that dividends forever are expected to grow from period to period at a constant rate.

6.1.1 A general dividend discount model

Assume the company pays dividends regularly, e.g. once every quarter or once every year with the same time distance between consecutive payments. We refer to this time distance as "a period." Let t denote the current point in time, and assume the company has just paid a dividend of D_t per share. Hence, there is exactly one period to the next dividend D_{t+1} , two periods to the subsequent dividend D_{t+2} , etc. Assume the dividends continue forever. The stock price P_t at time t is the value of all dividends arriving after time t , i.e. it is an ex-dividend price, excluding the dividend paid out at time t .

Future dividends are generally uncertain. The present value of a future dividend payment is generally calculated by discounting the expected dividend back to the valuation date using some appropriately risk-adjusted discount rate. Of course, your expectation of the dividend you will receive at a future date is depending on the information you have. In particular, the expectation may vary over time as you obtain new information about the company or the general macroeconomic prospects. Hence, the relevant expectation is a conditional expectation. We indicate the conditioning by a t -subscript on the expec-

tations operator, e.g., $E_t[D_{t+1}]$ denotes the expectation of the dividend in period $t + 1$ computed conditional on the information available at time t . If we assume that the same risk-adjusted discount rate r applies to all future dividend payments, the stock price at time t is given by

$$P_t = \frac{E_t[D_{t+1}]}{1+r} + \frac{E_t[D_{t+2}]}{(1+r)^2} + \frac{E_t[D_{t+3}]}{(1+r)^3} + \dots = \sum_{j=1}^{\infty} \frac{E_t[D_{t+j}]}{(1+r)^j}. \quad (6.1)$$

Of course, it is an enormous task to estimate expected dividend in all future periods. A typical approach is to estimate expected dividends maybe 5-10 years into the future and then make some simplifying assumption about how dividends evolve from that point on. We discuss this approach in more detail in the next subsection.

Determining the appropriate discount rate is also tricky. Generally, investors dislike uncertainty. They would rather have a \$10 dividend next year for sure than a claim to a dividend with an expected value of \$10 but with some uncertainty. In other words, the present value of an expected dividend of \$10 is lower than the present value of a certain \$10 dividend. The flip side of this statement is that the risk-adjusted discount rate exceeds the riskfree rate. Intuitively, you might expect the risk-adjusted discount rate to be increasing in the degree of uncertainty as, for example, measured by the standard deviation of the dividend. But, as we discuss extensively in several later chapters, this is not so simple. We need to figure out exactly how investors quantify the relevant risk of an uncertain future dividend and how big a compensation investors require to take on that risk. As most investors hold a portfolio of various assets, the appropriate risk measure and thus the appropriate risk-adjusted discount rate for each individual asset are depending on how this asset is contributing to the overall riskiness of the investors' portfolios. This involves the covariance between the different assets in the portfolio. We return to the question of determining the appropriate risk-adjusted discount rates in Chapters 10 and 11.

The pricing formula in (6.1) and the special cases derived below assume a full period to the next dividend payment date. What is the stock price at other dates? If a fraction $\tau \in [0,1)$ of a period has passed since the previous payment date, then a fraction $1-\tau$ of a period is left until the next payment. Hence, the first payment should only be discounted over that period, i.e. we divide the expected dividend by $(1+r)^{1-\tau} = (1+r)(1+r)^{-\tau}$. Similarly, the second dividend comes in $2-\tau$ periods so its present value equals the expected dividend divided by $(1+r)^{2-\tau} = (1+r)^2(1+r)^{-\tau}$. Similarly, for all future dividends. So, compared to (6.1), each right-hand side term is divided by an extra $(1+r)^{-\tau}$, i.e. multiplied by $(1+r)^{\tau}$. Hence, the stock price equals $(1+r)^{\tau}$ times what the stock price would be at the previous payment date. This is similar to Eq. (5.10) for bond prices.

While Eq. (6.1) states the stock price at time t , there is a similar expression for the stock price at future dates. Let $T > t$ denote a future dividend payment date, which means that $T - t$ is a positive integer. The stock price at time T is, of course, the present value at T of the dividends arriving after T . We can think of the stock price today as the present value of the dividends up to and including time T plus the present value of the stock price at T since the latter embodies the value of all dividends after time T . This leads to the recursive pricing formula stated in the next theorem.

Theorem 6.1

In the general dividend discount model, the stock price satisfies the recursive relation

$$P_t = \frac{E_t[D_{t+1}]}{1+r} + \frac{E_t[D_{t+2}]}{(1+r)^2} + \cdots + \frac{E_t[D_T]}{(1+r)^{T-t}} + \frac{E_t[P_T]}{(1+r)^{T-t}} \quad (6.2)$$

for any $T > t$. In particular,

$$P_t = \frac{E_t[D_{t+1} + P_{t+1}]}{1+r}, \quad (6.3)$$

so that the risk-adjusted discount rate equals the expected rate of return per period, i.e.

$$r = E_t \left[\frac{D_{t+1} + P_{t+1} - P_t}{P_t} \right]. \quad (6.4)$$

Proof

From Eq. (6.1), the stock price at time T is

$$P_T = \frac{E_T[D_{T+1}]}{1+r} + \frac{E_T[D_{T+2}]}{(1+r)^2} + \frac{E_T[D_{T+3}]}{(1+r)^3} + \dots$$

Seen from time t , you do not know exactly which information you will have at time T . Therefore, at time t , a conditional expectation like $E_T[D_{T+1}]$ is really a random variable. However, it can be shown that, for any $T > t$ and any random variable X , $E_t[E_T[X]] = E_t[X]$. In words: what you expect today to expect tomorrow about some value realized in the future is equal to what you expect today about that value. Furthermore, we know the expectation of a sum equals the sum of the expectations, so we get

$$E_t[P_T] = \frac{E_t[D_{T+1}]}{1+r} + \frac{E_t[D_{T+2}]}{(1+r)^2} + \frac{E_t[D_{T+3}]}{(1+r)^3} + \dots,$$

and, consequently,

$$\frac{E_t[P_T]}{(1+r)^{T-t}} = \frac{E_t[D_{T+1}]}{(1+r)^{T+1-t}} + \frac{E_t[D_{T+2}]}{(1+r)^{T+2-t}} + \frac{E_t[D_{T+3}]}{(1+r)^{T+3-t}} + \dots$$

If we write out some more terms in Eq. (6.1), the time t stock price satisfies

$$\begin{aligned} P_t &= \frac{E_t[D_{t+1}]}{1+r} + \frac{E_t[D_{t+2}]}{(1+r)^2} + \cdots + \frac{E_t[D_T]}{(1+r)^{T-t}} + \frac{E_t[D_{T+1}]}{(1+r)^{T+1-t}} + \frac{E_t[D_{T+2}]}{(1+r)^{T+2-t}} + \cdots \\ &= \frac{E_t[D_{t+1}]}{1+r} + \frac{E_t[D_{t+2}]}{(1+r)^2} + \cdots + \frac{E_t[D_T]}{(1+r)^{T-t}} + \frac{E_t[P_T]}{(1+r)^{T-t}}, \end{aligned}$$

which confirms (6.2). Eq. (6.3) follows from (6.2) by putting $T = t + 1$. Finally, Eq. (6.4) follows by rearranging (6.3).

Eq. (6.4) shows that the risk-adjusted discount rate is identical to the expected rate of return per period on the stock. Later chapters study models for the equilibrium expected

stock return and these model are therefore also telling us how to set the risk-adjusted discount rate appropriately.

6.1.2 Constant growth models

The so-called *Gordon's growth model*, named after Gordon (1962), assumes that dividends are expected to grow forever at the same periodic growth rate g . This means that, for any time t , having just observed the dividend D_t over the most recent period, the expected dividend over the next period is

$$\mathbb{E}_t[D_{t+1}] = (1 + g)D_t, \quad (6.5)$$

and, more generally, expected dividends in later periods are given by

$$\mathbb{E}_t[D_{t+s}] = (1 + g)^s D_t, \quad s = 1, 2, \dots$$

By substituting this into the general dividend discount formula (6.1), we get a so-called infinite geometric series that leads to a simple pricing formula:

Theorem 6.2

In Gordon's growth model with an expected dividend growth rate g and a discount rate $r > g$, the stock price is given by

$$P_t = \frac{(1 + g)D_t}{r - g} = \frac{\mathbb{E}_t[D_{t+1}]}{r - g}. \quad (6.6)$$

Proof

If we substitute the above expressions for expected dividends into (6.1), we get

$$\begin{aligned} P_t &= \frac{(1 + g)D_t}{1 + r} + \frac{(1 + g)^2 D_t}{(1 + r)^2} + \frac{(1 + g)^3 D_t}{(1 + r)^3} + \dots \\ &= D_t \left(\frac{1 + g}{1 + r} + \left(\frac{1 + g}{1 + r} \right)^2 + \left(\frac{1 + g}{1 + r} \right)^3 + \dots \right) = D_t \sum_{j=1}^{\infty} \left(\frac{1 + g}{1 + r} \right)^j. \end{aligned}$$

This is an infinite geometric series, and a general result is that (e.g. Sydsæter, Hammond, Strøm, and Carvajal 2021, Sec. 10.4)

$$\sum_{j=1}^{\infty} \alpha^j = \alpha + \alpha^2 + \alpha^3 + \dots = \frac{\alpha}{1 - \alpha}, \quad \text{if } |\alpha| < 1.$$

We apply this result with $\alpha = \frac{1+g}{1+r}$ in which case we get

$$\frac{\alpha}{1 - \alpha} = \frac{\frac{1+g}{1+r}}{1 - \frac{1+g}{1+r}} = \frac{1 + g}{1 + r - (1 + g)} = \frac{1 + g}{r - g}.$$

Therefore,

$$P_t = D_t \sum_{j=1}^{\infty} \left(\frac{1+g}{1+r} \right)^j = D_t \frac{1+g}{r-g},$$

which shows the first equality in (6.6). The second equality follows by using (6.5).

Note that we need $r > g$ for (6.6) to hold. If $r < g$, expected dividends would grow faster than the discount factor so the present value of each dividend payment would increase with the time distance until the payment. With infinitely many dividend payments, the total present value is infinite. If $r = g$, each dividend payment has the same present value, and with infinitely many dividends, we get an infinite total present value. We need the discount rate to exceed the dividend growth rate to make sure the future dividends have decreasing present values as we look further into the future. A company cannot expect to forever maintain a growth rate in dividends above the general growth rate of the economy since the company would then make up an ever increasing share of the economy. Hence, the dividend growth rate should generally not be set higher than the long-run expected growth rate in GDP, which is probably in the range of 1-3% per year.

A stock price is often seen in relation to the dividends paid by the company in the most recent year. Clearly, Gordon's growth model leads to a constant price-dividend ratio of

$$\frac{P_t}{D_t} = \frac{1+g}{r-g}. \quad (6.7)$$

If dividends are paid every quarter, and r and g refer to the discount rate per quarter and the expected growth rate per quarter, then Eq. (6.7) gives the ratio of the stock price to the quarterly dividends. This ratio then has to be divided by four to give the price to dividend-per-year ratio typically used by analysts and in the financial media.

Other things equal, a low price-dividend ratio indicates a more attractive stock, as the investor seemingly pays less for the same dividends. But, obviously, the price-dividend ratio may be low because of a small expected dividend growth rate or because of a large risk-adjusted discount rate due to highly risky future dividends. Other things equal, *risky and low-growth companies have low price-dividend ratios*. The sensitivity of the price-dividend ratio to g and r is captured by the partial derivatives

$$\frac{\partial(P_t/D_t)}{\partial g} = \frac{1+r}{(r-g)^2} > 0, \quad \frac{\partial(P_t/D_t)}{\partial r} = -\frac{1+g}{(r-g)^2} < 0.$$

The price-dividend ratio increases with the dividend growth rate and decreases with the risk-adjusted discount rate. The derivative with respect to r is decreasing in g , i.e. more negative when g is large. Hence, following a drop in the risk-adjusted discount rate, the price-dividend ratio increases more for high-growth firms than for low-growth firms.

In the model, the stock price next period is $P_{t+1} = D_{t+1}(1+g)/(r-g) = D_{t+1}P_t/D_t$ so that $P_{t+1}/P_t = D_{t+1}/D_t$. Prices grow at the same rate as dividends. The expected percentage capital gain per period is thus equal to g . The *dividend yield* over the next period is D_{t+1}/P_t and its expectation is

$$E_t \left[\frac{D_{t+1}}{P_t} \right] = \frac{E_t[D_{t+1}]}{P_t} = \frac{(1+g)D_t}{P_t} = (1+g) \frac{r-g}{1+g} = r-g,$$

using (6.5) and (6.7). The expected rate of return over the next period, r , can thus be divided into an expected dividend yield of $r - g$ and an expected capital gain of g .

The price computed using Gordon's growth model is very sensitive to the discount rate r and the dividend growth rate g . This is illustrated by the following example, which is taken from Brealey, Myers, and Allen (2009, Sec. 12.3).

Example 6.1

In the year 2000, the companies in the S&P 500 index paid dividends totaling 154.6 million USD. With a discount rate of 9.2% and an expected dividend growth rate of 8%, the value of the index at the beginning of the year should have been

$$P = \frac{154.6}{0.092 - 0.08} \approx 12,883 \text{ mn USD},$$

which was close to the actual value at that point in time.

Suppose the dividend growth rate is expected to drop just a little, to 7.4%, and stay at that level forever. Then the value would fall to

$$P = \frac{154.6}{0.092 - 0.074} \approx 8,589 \text{ mn USD},$$

which is close to the actual value in October 2002 after the so-called burst of the dot-com bubble. The small drop in the dividend growth rate causes a value reduction of one third.

The constant price-dividend ratio is at odds with the empirical findings presented later in this chapter. But the assumption of a constant dividend growth rate forever is also quite unrealistic. Companies are often built up around an innovative idea of a product or service. In the first years such companies typically experience relative high growth rates, but the growth rate of the demand then slows down and maybe competitors enter the market and force down profitability.

The dividend discount model can be generalized to a *multi-stage model* with different growth rates and possibly different risk-adjusted discount rates applying to different periods. After the last change in the dividend growth rate or the discount rate, we are in the stable, infinite-horizon situation of Gordon's basic model. The stock price at that point in time is sometimes referred to as a *terminal value*. If the growth and discount rates are assumed constant from time T , the terminal value is

$$P_T = D_T \frac{1 + g}{r - g} \tag{6.8}$$

provided that $g < r$. At any time $t < T$, we need to make assumptions about the dividend growth and the risk-adjusted discount rate between t and T . One tractable assumption is that dividends are expected to grow at a different rate G until T , but the risk-adjusted discount rate is r throughout. Then we can first calculate the present value of the dividends until T as a finite geometric series. The expected dividend at time T is then $D_t(1+G)^{T-t}$, which leads to an expected stock price at time T of $E_t[P_T] = D_t(1+G)^{T-t}(1+g)/(r-g)$ that we discount back to time t . The next theorem states the result.

Theorem 6.3

Suppose the expected growth rate of the dividends of a stock is G until time T and g thereafter, whereas the risk-adjusted discount rate is $r > g$ for all periods. The stock price at time $t < T$ is then

$$P_t = \begin{cases} D_t \frac{1+G}{r-G} \left[1 - \left(\frac{1+G}{1+r} \right)^{T-t} \right] + D_t \frac{1+g}{r-g} \left(\frac{1+G}{1+r} \right)^{T-t}, & \text{if } G \neq r, \\ (T-t)D_t + D_t \frac{1+g}{r-g}, & \text{if } G = r. \end{cases} \quad (6.9)$$

Proof

At time t , you expect the dividend at time T to be $D_t(1+G)^{T-t}$ and the dividends in the intermediate periods to be of the form $D_t(1+G)^k$ for $k = 1, 2, \dots, T-t$. Therefore, an application of (6.2) implies a price at time t of

$$\begin{aligned} P_t &= \frac{D_t(1+G)}{1+r} + \frac{D_t(1+G)^2}{(1+r)^2} + \dots + \frac{D_t(1+G)^{T-t}}{(1+r)^{T-t}} + \frac{D_t(1+G)^{T-t} \frac{1+g}{r-g}}{(1+r)^{T-t}} \\ &= D_t \left[\frac{1+G}{1+r} + \left(\frac{1+G}{1+r} \right)^2 + \dots + \left(\frac{1+G}{1+r} \right)^{T-t} \right] + D_t \frac{1+g}{r-g} \left(\frac{1+G}{1+r} \right)^{T-t} \\ &= D_t \frac{1+G}{r-G} \left[1 - \left(\frac{1+G}{1+r} \right)^{T-t} \right] + D_t \frac{1+g}{r-g} \left(\frac{1+G}{1+r} \right)^{T-t}, \end{aligned}$$

which shows the first case in (6.9). The last equality applies the mathematical result that

$$\sum_{k=1}^n \alpha^k = \alpha + \alpha^2 + \dots + \alpha^n = \frac{\alpha}{1-\alpha} (1 - \alpha^n), \quad (6.10)$$

provided that $\alpha \neq 1$. Of course, if $\alpha = 1$, the sum is simply n . (We applied this result already in Section 4.4, see Eq. (4.48).) In our case, this means that if $G = r$, the total present value of the dividends up to time T is equal to $(T-t)D_t$, which leads to the second case in (6.9).

In line with the discussion following Theorem 6.2, we note that a long-run growth rate g above an assumed 1-3% annual long-run growth in real GDP is hard to justify. The next example illustrates the two-stage model.

Example 6.2

The company ABC pays annual dividends and this year's dividends of \$5 per share have just been paid out. You expect dividends to grow by 10% per year over the next 5 years, after which you believe the growth rate slows down to 2% per year. The risk-adjusted

discount rate is 12% per year. According to (6.9), the current stock price should then be

$$P_t = \$5 \times \frac{1.10}{0.12 - 0.10} \times \left[1 - \left(\frac{1.10}{1.12} \right)^5 \right] + \$5 \times \frac{1.02}{0.12 - 0.02} \times \left(\frac{1.10}{1.12} \right)^5 \\ \approx \$23.69 + \$46.61 \approx \$70.30,$$

where the \$23.69 is the present value of the first five dividend payments and the \$46.61 is the present value of the rest. If the company could have maintained the 10% dividend growth forever, its stock price would have been $\$5 \times 1.10/0.02 = \275 . On the other hand, if the growth rate would have been 2% starting immediately, the stock price had been $\$5 \times 1.02/0.10 = \51 .

6.1.3 Discount rates versus expected returns

The above analysis has shown that when the appropriate periodic discount rate r is constant, then the expected rate of return on the stock over a period is identical to that discount rate. Now suppose that investors believe that the appropriate discount rate changes, say, to a lower level, but expected dividends stay the same. Eventually, the expected rate of return on the stock is going to be equal to that lower level and thus lower than before. However, in the transition period, the drop in the discount rate induces an increase in the stock price and thus an expected stock return higher than before.

To see this mathematically, assume that the infinite-horizon Gordon's growth model holds. Suppose that at time t , the discount rate is r and therefore the expected rate of return is also r . Now suppose that at time $t+1$, the discount rate has unexpectedly changed to $\hat{r} = r + \delta$, where the change δ can be positive or negative. We assume that the dividend growth rate remains g . From Eq. (6.6), we therefore have that

$$P_t = \frac{(1+g)D_t}{r-g}, \quad P_{t+1} = \frac{(1+g)D_{t+1}}{\hat{r}-g},$$

provided that the inequalities $r > g$ and $\hat{r} > g$ hold. The realized rate of return over the period is

$$r_{t+1} = \frac{D_{t+1} + P_{t+1}}{P_t} - 1 = \frac{D_{t+1} + \frac{(1+g)D_{t+1}}{\hat{r}-g}}{\frac{(1+g)D_t}{r-g}} - 1 = \frac{D_{t+1}}{D_t} \frac{(1+\hat{r})(r-g)}{(1+g)(\hat{r}-g)} - 1.$$

If $D_{t+1} = D_t(1+g)$, as expected, the realized return is going to be

$$r_{t+1} = \frac{(1+\hat{r})(r-g)}{\hat{r}-g} - 1 = \frac{(1+\hat{r})(r-g) - (\hat{r}-g)}{\hat{r}-g} \\ = \frac{\hat{r}(r-g) + r - \hat{r}}{\hat{r}-g} = \frac{(r+\delta)(r-g) - \delta}{r+\delta-g}.$$

If $\delta = 0$, so the discount rate is not changing, we see that $r_{t+1} = r$ as expected. If $\delta > 0$, we get $r_{t+1} < r$, so that the realized return temporarily declines when the appropriate discount rate—and thus future expected returns—increase. Conversely, if $\delta < 0$, we find $r_{t+1} > r$, that is, the realized return temporarily increases when the appropriate discount rate and the future expected returns decrease. Note the similarity to the analysis in Section 5.3.2 of bond returns when yields change.

In practice, the transition from one discount rate level to another may last several years. In this case, we can apply the above formula repeatedly to find the returns over the transition period.

Example 6.3

Suppose the current discount rate is $r = 8\%$ per year, that the expected discount growth rate is $g = 2\%$ per year, and that the dividend indeed grows by that rate over the next year. If the discount rate over the next year changes to $\hat{r} = 10\%$, then the realized return on the stock over this year is -17.5% , after which you expect the annual returns to be 10% . Alternatively, if the discount rate over the next year changes to $\hat{r} = 6\%$, the realized return on the stock over this year is 59.0% , after which you expect the annual returns to be 6% .

Now consider the case where the change in the discount rate happens gradually over a five-year period, with an annual increase or decrease of 0.4 percentage points. We assume that the realized dividend in each year is equal to the expected dividend given the assumed growth rate. If the discount rate increases from 8% to 10% over the five years, the annual rate of return in the five years would be 1.6%, 2.4%, 3.1%, 3.8%, and 4.5% after which the expected annual returns stabilize at 10%. If the discount rate decreases from 8% to 6% over the five years, the annual rate of return in the five years would be 15.3%, 15.4%, 15.7%, 16.1%, and 16.5% after which the expected annual returns stabilize at 6%.

A change in the discount rate has a larger effect on stocks with a larger growth rate. Suppose that $g = 4\%$ instead of the 2% assumed above. If the 8% discount rate increases to 10% over the next year, the realized return is then -26.7% compared to -17.5% for the stock with a 2% expected growth rate. If the discount rate decreases to 6% over the next year, the realized return is 112.0% instead of 59.0%.

The discount rate appropriate for a given stock reflects investors' required expected return. We can think of the required expected return as the sum of a riskfree rate and a risk premium. As discussed in Chapter 5, riskfree rates are generally associated with yields of zero-coupon government bonds and depend on the future payment date. If, for simplicity, we assume a flat yield curve, then the riskfree rate component of the required expected stock return should be the same for all stocks. Intuitively, the risk premium should depend on how risky future dividends are, and Chapters 10 and 11 aim at determining which risks investors demand a premium for taking and how big the associated risk premium is.

Changes in discount rates are due to changes in riskfree rates or risk premiums. When riskfree rates decline, as they did from 1981 to 2014 (with only short intermittent periods of minor increases), but risk premiums remain constant—or at least not increase more than riskfree rates fall—then discount rates decline. This signals lower expected stock returns in the future, but with relatively high returns during the period of declining discount rates, in particular for stocks of high-growth companies.

Investors' required risk premiums can also depend on other aspects of their preferences than risk aversion. An example is the increasing focus among investors on ESG (Environmental, Social, and Governance) issues. If two stocks deliver the same expected dividends in the future, some investors would prefer to invest in the stock of the company having the best ESG performance, e.g. by polluting less, emitting less carbon, or treating their employees better. The price is therefore higher for the good-ESG stock than the bad-ESG stock or, equivalently, the discount rate is lower for the good-ESG stock than the bad-ESG

stock. By implication, expected returns should be lower for good-ESG stocks than similar bad-ESG stocks. Again, returns are different through a transition phase of growing ESG awareness among investors. During a period of continued increasing ESG awareness, good-ESG stocks provide higher returns than comparable low-ESG stocks. But this does not change the fact that, in the longer run, expected returns on good-ESG stocks should be lower than on similar bad-ESG stocks. For more discussion and some empirical facts, see van der Beck (2021) and Pastor, Stambaugh, and Taylor (2022).

Investors might also revise the expected growth rate of the dividends from a stock. An increase in the growth rate leads to an increase in the stock price and thus an unexpectedly high return. Conversely, a decline in the growth rate leads to a drop in the stock price and a low, maybe negative, return, cf. Example 6.1. However, if the discount rate r remains the same, expected returns after the revision of the growth rate revert to r .

6.1.4 The observed price-dividend relation

Figure 6.1 shows that the price-dividend ratio and its reciprocal, the dividend yield, for the S&P 500 index have varied substantially between 1871 and 2021. A constant price-dividend ratio, as implied by the simple Gordon model, is not supported by the data. It might be tempting to assume that the price-dividend ratio is fluctuating around some fixed long-run level. If this is true, then the high current price-dividend ratio suggests that stocks are relatively expensive, and we might expect stock prices to decrease in the near future so that the price-dividend ratio reverts back to the historical average.

But just because prices are high relative to current dividends, prices are not necessarily *too* high. Extrapolating from the price-dividend ratio $(1 + g)/(r - g)$ in the Gordon model, it seems fair to assume that the price-dividend ratio is varying over time due to fluctuations in expected growth rates and discount rates. A high current price-dividend ratio could be due to low current discount rates or high expected growth rates instead of stocks being over-valued. The low interest rates seen through most of the post-2000 period certainly contributes to the high recent valuation ratios observed in Figure 6.1.

As mentioned in beginning of Section 6.1, cash dividends have to a large extent been replaced by stock repurchases since the 1980's. If the dividend measure used in the denominator of the price-dividend ratio excludes repurchases, the price-dividend ratio is going to increase and its reciprocal, the dividend yield, is going to decrease, other things equal, and this contributes to the much-higher-than-average price-dividend ratio seen in recent decades, as illustrated by Figure 6.1. If the shift in the payout policy from cash dividends to stock buybacks is permanent, we should not expect the ratio of stock prices to cash dividends to revert back. It seems appropriate to replace cash dividends by a measure of total payouts or net total payouts to shareholders and focus on the price-payout ratio or its reciprocal, the payout yield. In fact, time series of the price-payout ratio and the payout yield do appear to be more in line with the idea of variables moving around a stable long-term level, see, e.g., Eaton and Paye (2017).

6.2 Prices and other fundamentals than dividends

6.2.1 Earnings

Companies pay dividends out of their earnings. Let e_t denote the earnings per share in period t . Suppose the company has the policy to reinvest a fraction b of its earnings—the so-called plowback ratio—and pay out the remaining fraction $1 - b$ in dividends to shareholders. Then the dividend in period t is simply $D_t = (1 - b)e_t$ and similarly in other periods. Suppose that the company obtains a return on investment or equity (sometimes

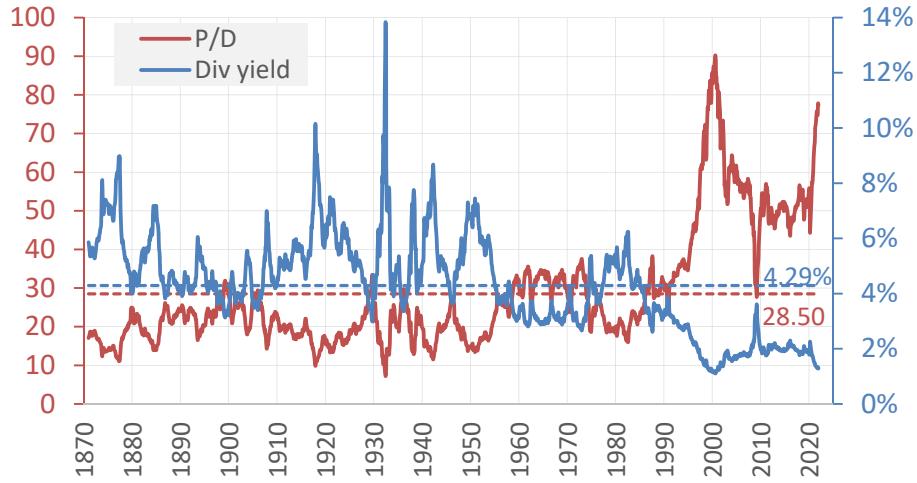


Figure 6.1: Stock prices and dividends.

Data is from the U.S. stock market over the period 1871-2021. The red graph (left vertical axis) shows for each month the price-dividend ratio, that is the ratio of the value of the S&P 500 index to an annualized measure of the dividends paid out during the month. The average value of 28.50 is indicated by the dotted red line. The blue graph (right vertical axis) shows the dividend yield, which is just the reciprocal of the price-dividend ratio. The average value of 4.29% is indicated by the dotted blue line. The data were downloaded on March 4, 2022 from the homepage of Professor Robert Shiller at Yale University, see <http://www.econ.yale.edu/~shiller/>.

abbreviated ROE) of r_e . The assets and thus the earnings are then growing at a rate of br_e , and since dividends are proportional to earnings they also grow at that rate. Hence, the valuation formulas developed in the preceding section applies with a dividend growth rate of $g = br_e$. From (6.6), we obtain

$$P_t = \frac{(1 - b)(1 + br_e)e_t}{r - br_e} = \frac{(1 - b)\mathbb{E}_t[e_{t+1}]}{r - br_e}. \quad (6.11)$$

Therefore, the price-earnings ratio is

$$\frac{P_t}{e_t} = \frac{(1 - b)(1 + br_e)}{r - br_e}, \quad (6.12)$$

whereas the so-called forward price-earnings ratio is

$$\frac{P_t}{\mathbb{E}_t[e_{t+1}]} = \frac{1 - b}{r - br_e}. \quad (6.13)$$

The term ‘forward’ is because next year’s expected earnings are used. Earnings are expected to grow at the same rate $g = br_e$ as dividends, so the current price-earnings ratio is constant, just as the price-dividend ratio, under the stated assumptions. The earnings per year are typically used in the calculation of the price-earnings ratio. Sometimes the *earnings yield* is used, which is simply the reciprocal of the price-earnings ratio.

The price-earnings ratio increases in r_e as you would expect. The larger the return on investments inside the company, the more valuable the stocks are. Mathematically the result follows since in the right-most ratio in Eq. (6.12), the numerator is increasing and the denominator is decreasing in r_e , and therefore the ratio increases with r_e . Clearly, the price-earnings ratio is decreasing in the risk-adjusted discount rate r . With a higher risk-adjusted discount rate, the earnings are worth less to the stock holders.

The dependence of the price-earnings ratio on the plowback ratio b is less clear. With a plowback ratio of zero, the price-earnings ratio is simply $1/r$. In this case, all earnings are paid out immediately as dividends. Since nothing is reinvested in the assets of the company, the growth rate is zero so the dividend stream is an infinite annuity stream having a present value equal to the current dividend (identical to current earnings) divided by the discount rate r . If you calculate the derivative of the price-earnings ratio with respect to b and evaluate the derivative for $b = 0$, you get $(r_e - r + r_e r)/(r - br_e)^2$, which is positive when $r_e \geq r$ (and even for r_e slightly below r) so in this case a positive b leads to a larger price-earnings ratio than when $b = 0$. The intuition is that when the company's return on equity exceeds the investor's risk-adjusted discount rate, then the investor prefers that the company reinvests at least some part of its earnings. But there is also a limit to how big the plowback ratio should be. If the company reinvests all earnings in the company, the stock holder would never receive any dividends. In fact, if $r_e \geq r$ and you let b increase towards r/r_e , then the price-earnings ratio goes to infinity since the denominator $r - br_e$ goes to zero. In this case the dividend growth rate br_e is approaching the discount rate, which implies that the present value of the dividends goes to infinity. Note that this analysis relies on the questionable assumptions that the company's return on equity can be upheld forever and is the same no matter how much of the earnings you reinvest.

Example 6.4

Suppose that company XYZ applies a 40% plowback ratio and can obtain a 12% return on investments. If the risk-adjusted discount rate is 10%, the forward price-earnings ratio of XYZ stocks should be

$$\frac{1 - 0.4}{0.1 - 0.4 \times 0.12} \approx 11.5.$$

If the expected earnings next year are \$2 per share, the stock price should be around \$23.

If the company could increase the return on investment to 15%, without changing the plowback ratio or the risk-adjusted discount rate, the forward price-earnings ratio would increase to

$$\frac{1 - 0.4}{0.1 - 0.4 \times 0.15} = 15,$$

which constitutes a 30% increase.

The value of a company can be divided into the value of its current assets and the value of its growth opportunities. If the company would just stick to its current assets, it would pay out all earnings as dividends and not grow at all. The (expected) dividend stream would then be a perpetuity with a present value of e_t/r , i.e., the price-earnings ratio (as well as the price-dividend ratio) would be $1/r$. Any value on top of this no-growth value must be due to growth opportunities. If we let O_t denote the present value at time t of

the growth opportunities per share, we have by construction

$$P_t = \frac{e_t}{r} + O_t$$

and

$$\frac{P_t}{e_t} = \frac{1}{r} + \frac{O_t}{e_t} = \frac{1}{r} \left(1 + \frac{O_t}{e_t/r} \right). \quad (6.14)$$

The last expression is interpreted as follows: the no-growth price-earnings ratio $1/r$ is scaled up by one plus the ratio of the value of growth opportunities to the value of the assets in place. This emphasizes that, to a large extent, the price-earnings ratio reflects the potential growth of the company.

In order to apply (6.14), you need the value of the growth opportunities. Some growth opportunities are very similar to the firm's existing operations and, assuming a constant growth rate, these growth opportunities can be valued as shown earlier in this chapter. Other and more complex growth opportunities are difficult to value. In some cases growth opportunities involve one or more strategic options, and then you have to invoke option valuation methods as explained in Chapter 15.

Price-earnings ratios differ substantially across industries. According to January 2022 data published by Professor Aswath Damodaran at New York University's Stern School of Business, the lowest U.S. forward price-earnings ratios are found in the industries labelled "life insurance" (ratio of 8.2), "steel" (8.5), "rubber and tires" (8.6), and "basic chemical" (9.1). Very high ratios are found in "telecom, services" (ratio of 3798), "software, internet" (839), and "real estate development" (712). The price-earnings ratio for the overall U.S. stock market is reported to be 61.¹

The earnings number entering the price-earnings ratio is generally taken from the company's reported earnings in its financial statements. The reported earnings are affected by various accounting rules involving historical costs. Therefore, the reported earnings may give a distorted picture of the company's current earnings. Furthermore, companies have some degrees of freedom when determining the accounting rules, and sometimes managers may have incentives to artificially inflate or deflate the reported earnings. Finally, the earnings and thus the price-earnings ratio are sensitive to the company's degree of leverage.

Earnings in any particular year can be affected by company-specific shocks in that year or by business cycle variations. And, as just mentioned, earnings can to some extent be manipulated by management through a creative use of accounting techniques. For these reasons, the stock price of a company is sometimes related to an average of annual earnings over a number of years. A frequently applied measure is the so-called CAPE—the cyclically adjusted price-earnings ratio—defined as the current stock price level divided by an average of inflation-adjusted earnings over the past ten years, cf. Campbell and Shiller (1988b) and Siegel (2016). The CAPE is also known as the P/E 10 ratio. Just as the basic price-dividend or price-earnings ratio, the CAPE can be calculated both for individual companies, for industries, or for the entire stock market.

Figure 6.2 shows that the CAPE of the U.S. stock market has fluctuated significantly over the period 1881-2021. Again a constant price-earnings ratio as in the Gordon model is not supported by the data. The most recent CAPE value of 38.66 is more than double the long-run average of 17.24. Does this mean that stocks are currently over-valued? Not necessarily. The figure suggests that the price-earnings ratio tends to be negatively related to long-term interest rates. This makes perfect sense since, other things equal, a stream

¹See http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/pedata.html.

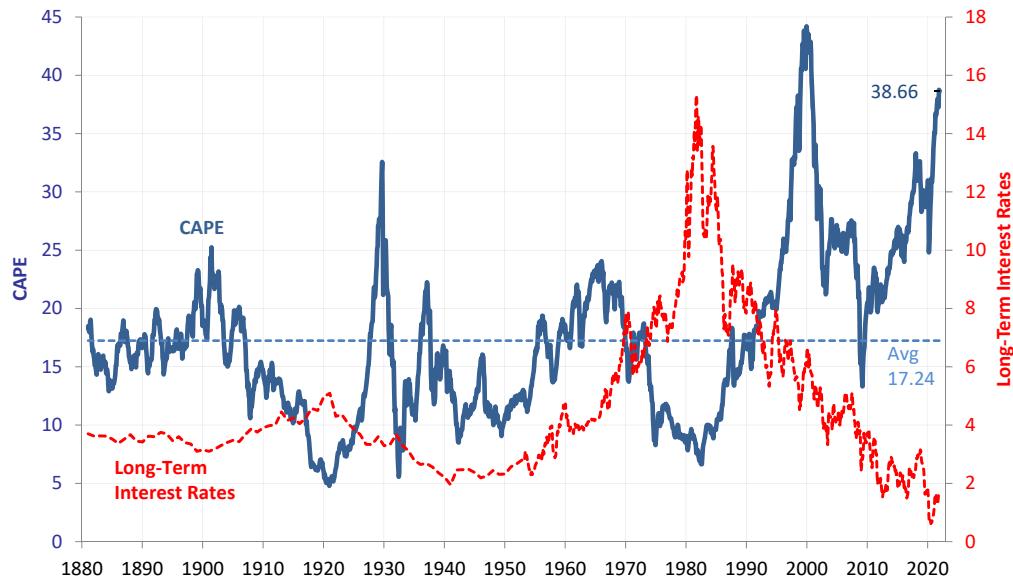


Figure 6.2: The history of the CAPE.

The solid blue curve shows how the CAPE of the S&P 500 index has developed from January 1881 to December 2021. The dashed blue line shows the average of the CAPM which is 17.24. Then numerical values of both blue curves are on the left axis. The red curve shows the 10-year Treasury bond yield and is to be read off the right axis. Monthly observations are used for both the CAPE and the 10-year yield. The graph is an edited version of the one published together with the underlying data on the homepage of Professor Robert Shiller at Yale University, see <http://www.econ.yale.edu/~shiller/data.htm>. The data were downloaded on March 4, 2022.

of earnings has a higher present value when interest rates are low. Hence, the high recent values of the CAPE can, at least to some extent, be justified by low interest rates.

6.2.2 Free cash flows

Some companies currently pay no dividends and may have negative earnings, and then the above valuation approach cannot be used. An often used alternative is to base the valuation on the free cash flows, which does not require positive dividends or earnings.

The free cash flow of a company in a given period is the after-tax cash flow generated by the firm's operations which is available for distribution among shareholders and creditors. More precisely, the free cash flow is the earnings before interest and taxes (EBIT) in that period multiplied by one minus the corporate tax rate, plus depreciation, minus the net increase in working capital, and minus capital expenditures. The entire value of the firm can, in principle, be found as the present value of all future free cash flows, where the discount rate should reflect the risk of the cash flows and is often represented by the company's WACC, i.e., weighted average cost of capital. Formally, the computation would be of the form

$$V_t = \sum_{j=1}^{\infty} \frac{E_t [FCF_{t+j}]}{(1 + r_{\text{firm}})^j} = \sum_{j=1}^{T-t} \frac{E_t [FCF_{t+j}]}{(1 + r_{\text{firm}})^j} + \frac{E_t [V_T]}{(1 + r_{\text{firm}})^{T-t}}, \quad (6.15)$$

similar to the recursive price-dividend relation (6.2).

Often a two-stage approach is taken, just as discussed in the dividend discount model above. For the first $T - t$ years, expected free cash flows are estimated, maybe based on a thorough, fundamental analysis of the company. Then the firm value at time T ahead is estimated assuming that expected free cash flows will grow at a constant rate from that point on, i.e.,

$$V_T = \frac{(1 + g)FCF_T}{r_{\text{firm}} - g}.$$

The value of the stocks can next be found as the difference between the entire firm value and the market value of the debt, where the latter is often close to the face value (the outstanding debt).

Alternatively, the equity can be valued from discounting the free cash flows made to shareholders by a discount rate reflecting the risk of these payments. Here, the free cash flow to equity is the total free cash flow less after-tax interest payments plus the value of any net increase in the debt.

6.2.3 Other fundamentals

In the models described above, the stock price is a certain multiple of current dividends, earnings, or free cash flows. Financial analysts also apply other multipliers, i.e. they relate stock prices to other variables that are important for the particular company, the industry in which it operates, or the entire economy. One example is the market-to-book ratio which is simply the market price per share of the stock relative to the book value per share as recorded in the balance sheet of the company. Often, the reciprocal book-to-market ratio is used. Another example is the price-to-sales ratio. For start-up companies or companies in specific industries, it may make sense to consider the company value relative to other variables that are likely to predict future profits to the company. For more on valuation and the link to financial statements, the reader is referred to books like [Petersen and Plenborg \(2012\)](#) and [Penman \(2013\)](#). If the goal is to assess the value of a country's entire stock market, it may make sense to look at the ratio of the total stock market value to the GDP of the country, the aggregate consumption level in the country, or some other macroeconomic variables.

All the valuation formulas above aim at computing the fair current value of the stocks given your expectations about the company's future performance. Of course, analysts may have different expectations (or apply different risk-adjusted discount rates) so that they arrive at different prices. These different price estimates are the basis for trading in the stock.

Sometimes stock prices (or the prices of other assets) seem hard to justify from reasonable expectations about the firm's future earnings and dividends. Following surprisingly large earnings in the recent past, some investors may extrapolate that into sustained large earnings in the future as well, but most often extraordinarily large earnings *are* just extraordinary and cannot be sustained for long. As illustrated by Example 6.1, just a small increase in the expected growth rate may translate into a significant increase in the current valuation estimate.

Other investors may think that a stock is currently overpriced relative to fundamentals (earnings or dividends with reasonable growth rates), but if they expect the mispricing to increase further, it may still be profitable for them to buy the stock and sell it again later on. Buying a stock gives you dividends, but also the right to sell the stock again, potentially at a price you think is higher than justified by the "fundamentals" of the firm.

These mechanisms may lead to price bubbles which seem to appear at rare occasions in stock markets and other asset markets. It is much easier to detect a price bubble after

it has burst, though. If prices suddenly drop a lot, they must have been too high before, right? But, again, small changes in expectations or discount rates may lead to large changes in present values. Even if you believe in a price bubble, it may be difficult to profit from it, at least in the short run. Maybe you cannot short-sell the bubbly assets and, even if you could, you run the risk that the bubble grows bigger before it bursts, and in the meantime you lose money on your short position while the long investors continue to profit.

In Section 6.5.5 we present some empirical evidence on the variations in the price-dividend and price-earnings ratios over time.

6.3 A decomposition of stock returns

The rate of return over a period consists of a dividend yield and a relative capital gain:

$$r_{t+1} = \frac{D_{t+1}}{P_t} + \frac{P_{t+1} - P_t}{P_t} = \frac{D_{t+1}}{P_t} + \frac{P_{t+1}}{P_t} - 1.$$

For any variable F , we can rewrite the price ratio P_{t+1}/P_t as

$$\frac{P_{t+1}}{P_t} = \frac{P_{t+1}/F_{t+1}}{P_t/F_t} \frac{F_{t+1}}{F_t} = (1 + G_{t+1}^{PF}) (1 + G_{t+1}^F),$$

where G_{t+1}^{PF} and G_{t+1}^F denote the percentage growth rates in the P/F -ratio and in F itself, respectively. This is useful when F represents some “fundamental” such as the dividend, the earnings, or any other variable that prices seem related to, cf. the discussion in the previous sections. The ratio P/F is then the price-fundamental ratio or the “valuation ratio.” Combining the two equations above, we get

$$\begin{aligned} r_{t+1} &= \frac{D_{t+1}}{P_t} + (1 + G_{t+1}^{PF}) (1 + G_{t+1}^F) - 1 \\ &= \frac{D_{t+1}}{P_t} + G_{t+1}^{PF} + G_{t+1}^F + G_{t+1}^{PF} \times G_{t+1}^F. \end{aligned} \quad (6.16)$$

We see that the rate of return is decomposed into the dividend yield, the growth in the valuation ratio, the growth in the fundamental itself, plus the product of these growth rates. The product tends to be small, so that the three first terms dominate. The expected rate of return next period is thus the sum of the expected dividend yield, the expected growth in the valuation ratio, the expected growth in the fundamental, plus the small product term.

Let us take Gordon’s growth model as an example. In this model, prices are related to dividends so the fundamental is the dividend itself. From (6.6), the valuation ratio in the model is $P_t/D_t = (1 + g)/(r - g)$ which is a constant and thus $G_{t+1}^{PF} = 0$. Moreover, the model assumes that the expected dividend growth is g so $G_{t+1}^F = g$. The expectation of the dividend yield D_{t+1}/P_t is the constant $r - g$. Hence, the expected return is the sum of the expected dividend yield $r - g$ and the expected dividend growth rate g , which equals r and this is indeed the expected rate of return in Gordon’s growth model.

Alternatively, we can use earnings as the fundamental. In the Gordon-style model of Section 6.2.1, the expectation of the dividend yield is $r - br_e$, the growth rate of the valuation ratio, i.e. the price-earnings ratio, is zero, and the expected growth rate of the fundamental is br_e . Again the terms add up to r , the expected rate of return.

As we shall see later in this chapter, the valuation ratios typically considered are not constant over time. For example, Figure 6.1 shows that the price-dividend ratio of the

S&P 500 stocks has varied substantially over time. Other things equal, we will observe relatively high returns in periods where the valuation ratio increases and low returns when the valuation ratio decreases. Valuation ratios cannot increase indefinitely and most valuation ratios seem to vary around some long-run average. This implies that if high returns over an extended time period are due to a hike in the valuation ratio, then we should expect lower future returns when the valuation ratio reverts towards its long-run level.

6.4 Equity duration

In Section 5.8, the duration of a bond was defined as a present-value-weighted average of the time to the payments of the bond. Also among stocks, we can think of short-duration stocks for which a large share of the total value is due to payments to stockholders in the near future whereas long-duration stocks are mainly valuable because of payments that are expected many years from now. Firms with higher growth rates have longer durations, all else equal, as their expected cash flows far into the future are relatively large compared to cash flows in the near future.

More precisely, let the dividend in any given period represent the cash flow per share from the company to the shareholders in that period. The price per share P_t at time t then equals the sum over all future periods of the discounted expected future dividends per share as stated in Eq. (6.1). To distinguish duration from dividends, we now use DUR instead of D to denote the duration. We can then define the duration of the stock at time t as

$$\text{DUR}_t = \sum_{i=1}^{\infty} i \times w_{i,t}, \quad w_{i,t} = \frac{(1+r)^{-i} \mathbb{E}_t[D_{t+i}]}{P_t}, \quad (6.17)$$

where r is the appropriate discount rate per period. Theorem 5.2 showed that, besides being a weighted average time-to-maturity, the bond duration is also measuring the price sensitivity with respect to the bond's yield. We have a similar result for stock durations:

$$\text{DUR}_t = -\frac{1+r}{P_t} \frac{\partial P_t}{\partial r}. \quad (6.18)$$

Note, however, that this relation assumes that expected future dividends are not changing with the discount rate. Such an assumption is questionable. For example, both expected future earnings or dividends and discount rates are likely to vary over the business cycle.

As discussed for the dividend discount model, analysts often assume a constant growth rate either for all future periods or after some initial periods with a different dividend growth pattern. Such assumptions allows us to derive more explicit expressions for the stock duration as summarized in the next theorem.

Theorem 6.4

Assume a constant discount rate r per period. We have the following formulas for the stock duration DUR_t :

- (a) If the expected dividend growth rate is a constant $g < r$ in all future periods, then

$$\text{DUR}_t = \frac{1+r}{r-g}. \quad (6.19)$$

(b) If the expected dividend growth rate is G until time $T > t$ and $g < r$ thereafter, then

$$\text{DUR}_t = \begin{cases} \frac{1+r}{r-G} + \frac{T-t+\frac{1+g}{r-g}}{1-\frac{r-g}{G-g}\left(\frac{1+r}{1+G}\right)^{T-t-1}}, & \text{if } G \neq r, \\ \frac{\frac{1}{2}(T-t)(T-t+1)+\left(\frac{1+r}{r-g}+T-t\right)\frac{1+g}{r-g}}{T-t+\frac{1+g}{r-g}}, & \text{if } G = r. \end{cases} \quad (6.20)$$

(c) If the expected dividend growth rate is $g < r$ after time T , then

$$\text{DUR}_t = \frac{\sum_{i=1}^{T-t} i(1+r)^{-i} \mathbb{E}_t[D_{t+i}]}{P_t} + \frac{P_t - \sum_{i=1}^T (1+r)^{-i} \mathbb{E}_t[D_{t+i}]}{P_t} \left(T - t + \frac{1+r}{r-g} \right). \quad (6.21)$$

Proof

(a) This is Gordon's growth model so the stock price is given by Eq. (6.6). The derivative $\partial P_t / \partial r$ and the duration DUR_t are therefore

$$\frac{\partial P_t}{\partial r} = -\frac{(1+g)D_t}{(r-g)^2} = -\frac{P_t}{r-g}, \quad \text{DUR}_t = -\frac{1+r}{P_t} \times \left(-\frac{P_t}{r-g} \right) = \frac{1+r}{r-g}.$$

(b) Under the stated assumptions, the stock price is given by Eq. (6.9). If $G \neq r$, the derivative can be written as

$$\frac{\partial P_t}{\partial r} = -D_t \left\{ \frac{1+G}{(r-G)^2} + \left(\frac{1+G}{1+r} \right)^{T-t} \left[\frac{1+g}{(r-g)^2} - \frac{1+G}{(r-G)^2} + \frac{T-t}{1+r} \left(\frac{1+g}{r-g} - \frac{1+G}{r-G} \right) \right] \right\}.$$

By substituting this into (6.18) and rewriting the resulting expression, we eventually arrive at the duration shown in the first case of (6.20).

If $G = r$, the price is $P_t = D_t \left(T - t + \frac{1+g}{r-g} \right)$ according to Eq. (6.9). But here we cannot just differentiate with respect to r since that would mean that also G changes along with r , which is not what we want. However, we can calculate the duration directly using (6.17). Let $\hat{P}_{t,T+}$ denote the present value at time t of all payments after time T . We can split the duration defined in (6.17) into a sum over the first T dividends and a sum of the remaining dividends, and if we multiply and divide the second sum by $\hat{P}_{t,T+}$, we get

$$\text{DUR}_t = \sum_{i=1}^{T-t} i \times w_{i,t} + \frac{\hat{P}_{t,T+}}{P_t} \sum_{i=T-t+1}^{\infty} i \times \frac{(1+r)^{-i} \mathbb{E}_t[D_{t+i}]}{\hat{P}_{t,T+}}, \quad (6.22)$$

and the second sum in this expression is exactly the duration at time t of the claim to the dividends arriving after time T . Given the constant growth rate g after time T , we know from (a) that this claim has a duration of $(1+r)/(r-g)$ at time T , and therefore $T - t + (1+r)/(r-g)$ at time t , cf. Eq. (5.40). With $G = r$, each of the first $T - t$ payments has a present value equal to D_t and thus a weight of $w_{i,t} = 1/(T - t + \frac{1+g}{r-g})$. The payments after time T have a present value at time t of $\hat{P}_{t,T+} = D_t(1+g)/(r-g)$.

Substituting into (6.22), we find that the duration of the entire cash flow stream is

$$\text{DUR}_t = \frac{1}{T-t+\frac{1+g}{r-g}} \left(\sum_{i=1}^{T-t} i \right) + \frac{\frac{1+g}{r-g}}{T-t+\frac{1+g}{r-g}} \left(T-t+\frac{1+r}{r-g} \right),$$

and since $\sum_{i=1}^{T-t} i = (T-t)(T-t+1)/2$, this confirms the second part of (6.20).

(c) This follows from (6.22) by using, as in the proof of (b), the result of part (a) for the payments arriving after time T . Also, we replace $\hat{P}_{t,T+}$ by $P_t - \sum_{i=1}^T (1+r)^{-i} E_t[D_{t+i}]$, which holds simply because the value of the payments after time T equals the value of all payments less the payments arriving before and at time T .

The constant-growth duration in Eq. (6.19) generalizes the duration for the zero-growth perpetuity in Eq. (5.37). For example, with a discount rate of $r = 10\%$ per year, the duration is $1.1/0.1 = 11$ years without growth, but 22 years with a growth rate of $g = 5\%$ per year. Clearly, the duration is increasing in g . The next example considers the two-stage case covered by Part (b) of the above theorem. Among the conclusions are that duration is increasing in growth rates and decreasing in the discount rate and thus in the risk of the company. For example, companies in manufacturing, coal and gas, and financial services tend to have low duration, whereas computer technology and bio-science companies tend to have high duration (Weber 2018; Dechow, Erhard, Sloan, and Soliman 2021).

Example 6.5

For various combinations of growth rates and discount rates, Table 6.1 shows the price-dividend ratio, the duration, as well as present value weights for each of the first 10 years, the sum of those weights, and the present value weight for the payments arriving more than 10 years into the future. For example, suppose that the discount rate is 8% and the expected dividends grow by 5% per year the first 10 years and by 1% thereafter. Then the price-dividend ratio is 19.48 and the duration is 16.54 years. The present value of the dividend in the first year constitutes 5.0% of the total price. The present value of the first 10 years of dividends is 44.1% of the total price, so the present value of all later dividends is 55.9% of the total price.

The table confirms that the price-dividend ratio is increasing in both the short-term growth rate G and the long-term growth rate g . Also the duration is increasing in both G and g since an increase in anyone of them means that dividends in the distant future are larger and constitute a larger present value share of the total price. The duration is more sensitive to the long-run growth rate g than the short-run growth rate G .

A higher discount rate leads to lower price-dividend ratios and also to lower durations as the present value is reduced more for dividends far into the future than for soon-to-come dividends. The latter results comply with Figure 5.10 that shows how bond durations are decreasing in the yield.

The expected dividends in the duration expression (6.21) represent cash flows to shareholders that can be forecast based on accounting reports and various assumptions. This

	$r = 8\%$				$r = 12\%$			
	$G = 5\%$		$G = 10\%$		$G = 5\%$		$G = 10\%$	
	$g = 1\%$	$g = 3\%$	$g = 1\%$	$g = 3\%$	$g = 1\%$	$g = 3\%$	$g = 1\%$	$g = 3\%$
P/D	19.48	24.14	28.41	35.83	11.95	13.14	16.74	18.63
Duration	16.54	22.23	17.72	23.58	11.10	12.96	12.15	14.12
Year 1	5.0%	4.0%	3.6%	2.8%	7.8%	7.1%	5.9%	5.3%
Year 2	4.9%	3.9%	3.7%	2.9%	7.4%	6.7%	5.8%	5.2%
Year 3	4.7%	3.8%	3.7%	2.9%	6.9%	6.3%	5.7%	5.1%
Year 4	4.6%	3.7%	3.8%	3.0%	6.5%	5.9%	5.6%	5.0%
Year 5	4.5%	3.6%	3.9%	3.1%	6.1%	5.5%	5.5%	4.9%
Year 6	4.3%	3.5%	3.9%	3.1%	5.7%	5.2%	5.4%	4.8%
Year 7	4.2%	3.4%	4.0%	3.2%	5.3%	4.8%	5.3%	4.7%
Year 8	4.1%	3.3%	4.1%	3.2%	5.0%	4.5%	5.2%	4.6%
Year 9	4.0%	3.2%	4.2%	3.3%	4.7%	4.3%	5.1%	4.6%
Year 10	3.9%	3.1%	4.2%	3.4%	4.4%	4.0%	5.0%	4.5%
Year 1-10	44.1%	35.6%	39.0%	30.9%	59.7%	54.3%	54.2%	48.7%
Rest	55.9%	64.4%	61.0%	69.1%	40.3%	45.7%	45.8%	51.3%

Table 6.1: Present value weights and equity duration.

The table assumes a two-stage model where the expected dividend growth rate is G for the first 10 years and g thereafter, whereas the discount rate is r throughout. For various combinations of r , G , and g , the top rows show the current price-dividend ratio P_t/D_t and the equity duration DUR_t . The next 10 rows show the present value weight of the dividend in each of the years 1-10. Row two from below gives the total present value weight of the first 10 years of dividends, whereas the final row gives the present value weight of all dividends arriving more than 10 years out.

approach is used by Dechow, Sloan, and Soliman (2004) and Weber (2018), among others. In some cases, publicly available analyst forecast can be used in the firm-level duration calculation, see Schröder and Esterer (2016). For some of the larger European and US stocks, futures on dividends in different business years are traded at the Eurex exchange. For example, in March 2022, you can buy various dividend futures on the company Adidas. One contract is on the dividends paid out in 2023 and naturally matures at the end of 2023. Another contract is on the dividends in 2024 and matures at the end of 2024, etc. Typically the offered contracts cover dividends of up to 3-7 years into the future. The futures prices for the different calendar-year contracts contain information about the present value today of the dividends coming in each of these years, which can be used in the calculation and studies of the equity duration, see Gormsen and Lazarus (2023).

The equity duration is a useful tool for understanding differences in the timing of cash flows across firms or industries. Empirical results indicate that stocks with low duration tend to deliver higher returns than stocks with high duration. This phenomenon is sometimes referred to as a downward-sloping *equity term structure*.

The equity duration can also be considered at an index or market level. The specific duration of a stock index can be used as a benchmark that individual stock durations can be compared with, e.g. to get an impression of whether the cash flow of a given stock is more front loaded or back loaded than the cash flow of an average stock. Dividend futures are traded on several leading stock indices with maturities of up to 10 years, i.e. covering dividends of up to 10 years out. The futures prices and returns across maturities provide information about how the stock market values payments at different horizons, see van Binsbergen, Hueskes, Koijen, and Vrugt (2013), van Binsbergen and Koijen (2017), and Gormsen (2021). The prices of index options of various maturities can also be useful

for this purpose, see van Binsbergen, Brandt, and Kojen (2012). Also these studies indicate that the term structure of equity returns is typically downward sloping but that it tends to be upward sloping in bad times.

6.5 Stylized facts about stock market returns

This section intends to provide an overview of the historical performance of the stock market. We focus mostly on the S&P 500 index which is one of the most prominent stock indices in the World, but we also report some international evidence. Section 6.6 focuses on differences in historical performance across stocks by looking at various portfolios formed by sorting stocks according to some criteria, e.g. the market capitalization of the company. Section 6.7 takes the analysis to the level of individual stocks.

6.5.1 Historical returns on the S&P 500 index

The S&P 500 index measures the performance of the stocks of 500 large companies listed on U.S. stock exchanges and covers about 80% of the value of all listed U.S. stocks. Several variations of the index exist, but the one typically referred to in the media is a value-weighted price return index, not accounting for dividend payments. The historical changes in the index are therefore underestimating the total return investors have made in the stock market. A total return version of the index that includes dividends is published since 1988.

The CRSP database includes reliable data on the returns on both a value-weighted and an equally-weighted portfolio of the stocks in the S&P 500 index dating back to 1926. Of course, the portfolio weights are regularly rebalanced. These returns include dividend payments. Table 6.2 lists summary statistics on the monthly and annual returns on these portfolios over the period 1927-2019, the post World War II period 1946-2019, and the most recent 30-year period 1990-2019. Figure 6.3 shows histograms of annual and monthly returns on both portfolios. Figure 6.4 shows the return on the S&P 500 U.S. stock index in each year over the period 1927-2019.

First, focus on nominal returns. In the full 1927-2019 sample, the arithmetic [geometric] average annual return is 11.9% [10.0%] for the value-weighted and 14.3% [11.6%] for the equally-weighted portfolio. The average return is very similar for the shorter samples as for the full sample. The average annual return is close to, but not identical to, $(1 + \bar{r}_{\text{mon}})^{12} - 1$, where \bar{r}_{mon} is the average monthly return, cf. Equation (3.83). To illustrate the role of dividends, the average monthly change in the S&P 500 index in 1927-2019 was 0.64%. The gap to the 0.94% average return on the value-weighted portfolio shown in the table is due to dividend payments.

Some of the other moments differ across time periods. In particular, the standard deviation is larger in the full sample than in the two more recent samples, and for monthly returns the kurtosis is a lot larger when the early part of the sample is included. These differences arise from a large number of extreme returns in the early part of the sample, such as monthly returns of 68.0% (August 1932), 55.2% (July 1932), -31.0% (September 1931), and -29.9% (May 1932) for the equally-weighted portfolio. In fact, in all years from 1927 to 1945 the return was either negative or above 20%, both for the value-weighted (see Figure 6.4) and the equally-weighted portfolio. The histograms also show more frequent extreme returns when the 1927-1945 is included.

If we use the 1946-2019 sample, the standard deviation of annual returns (also known as the volatility) is 17% for the value-weighted and 19% for the equally-weighted portfolio, and the annual returns exhibit negative skewness and a moderately positive kurtosis.

	Period	Mean	Std dev	Skew	Kurt	Min	Max
Annual returns							
VW nominal	1927-2019	11.9%	19.9%	-0.446	0.098	-45.5%	53.3%
VW nominal	1946-2019	12.4%	17.0%	-0.351	0.109	-36.6%	52.8%
VW nominal	1990-2019	11.5%	17.5%	-0.767	0.623	-36.6%	37.7%
EW nominal	1927-2019	14.3%	24.1%	0.042	1.081	-53.0%	95.7%
EW nominal	1946-2019	13.9%	19.0%	-0.176	0.193	-40.1%	58.0%
EW nominal	1990-2019	13.1%	18.5%	-0.637	1.236	-40.1%	47.4%
Inflation	1946-2019	3.69%	3.36%	1.942	4.960	-2.07%	18.1%
VW real	1946-2019	8.60%	17.4%	-0.301	0.144	-36.7%	53.9%
EW real	1946-2019	10.1%	19.1%	-0.204	0.146	-40.2%	55.2%
1Y riskfree	1946-2019	3.99%	3.12%	0.919	0.886	0.02%	14.7%
VW excess	1946-2019	8.41%	17.4%	-0.320	0.126	-38.2%	51.9%
EW excess	1946-2019	9.96%	19.3%	-0.125	0.267	-41.7%	56.4%
Monthly returns							
VW nominal	1927-2019	0.94%	5.41%	0.356	9.794	-28.7%	41.4%
VW nominal	1946-2019	0.96%	4.13%	-0.425	1.654	-21.6%	16.8%
VW nominal	1990-2019	0.88%	4.09%	-0.612	1.260	-16.7%	11.4%
EW nominal	1927-2019	1.14%	6.74%	1.472	18.574	-31.0%	68.0%
EW nominal	1946-2019	1.08%	4.69%	-0.325	2.508	-25.6%	23.1%
EW nominal	1990-2019	1.01%	4.65%	-0.492	2.125	-20.9%	18.5%
Inflation	1946-2019	0.30%	0.45%	2.577	29.203	-1.92%	5.88%
VW real	1946-2019	0.66%	4.17%	-0.421	1.458	-21.8%	15.6%
EW real	1946-2019	0.78%	4.73%	-0.320	2.310	-25.8%	22.7%
1M riskfree	1946-2019	0.32%	0.25%	0.976	1.174	0.00%	1.35%
VW excess	1946-2019	0.64%	4.15%	-0.446	1.627	-22.2%	16.3%
EW excess	1946-2019	0.76%	4.70%	-0.345	2.485	-26.2%	22.5%

Table 6.2: Summary statistics for the S&P 500 index.

Data on the value-weighted (VW) and equally-weighted (EW) returns on the stocks in the S&P 500 index as well as the inflation rate were downloaded from CRSP through WRDS on June 22, 2020. Data on the 1 month and the 1 year riskfree rate were downloaded from the homepage of Kenneth French on June 22, 2020. The mean shown in the table is the arithmetic average.

The larger mean and standard deviation for the equally-weighted portfolio relative to the value-weighted portfolio shows that the smaller stocks in the index (measured by market capitalization) have offered a larger and more volatile return than larger stocks. The equally-weighted portfolio also exhibits a larger kurtosis and more extreme minimum and maximum returns than the value-weighted portfolio. The different performance of small and large stocks is discussed in more depth in a later section. The histograms for this sample period seem roughly consistent with a normal distribution, and mostly so for the annual returns of the value-weighted portfolio. The kurtosis tends to be larger for monthly returns than for annual returns, whereas the skewness has roughly the same magnitude for monthly as for annual returns. Given the analysis of Section 3.6, we would probably expect to see the skewness and kurtosis increase when going from monthly to annual returns. Apparently, relatively extreme monthly price movements occur rather frequently, but are often “corrected” by price movements in the opposite direction in subsequent months, so that the annual returns are less extreme.

Investors should generally care about the returns adjusted for inflation. Given the peri-

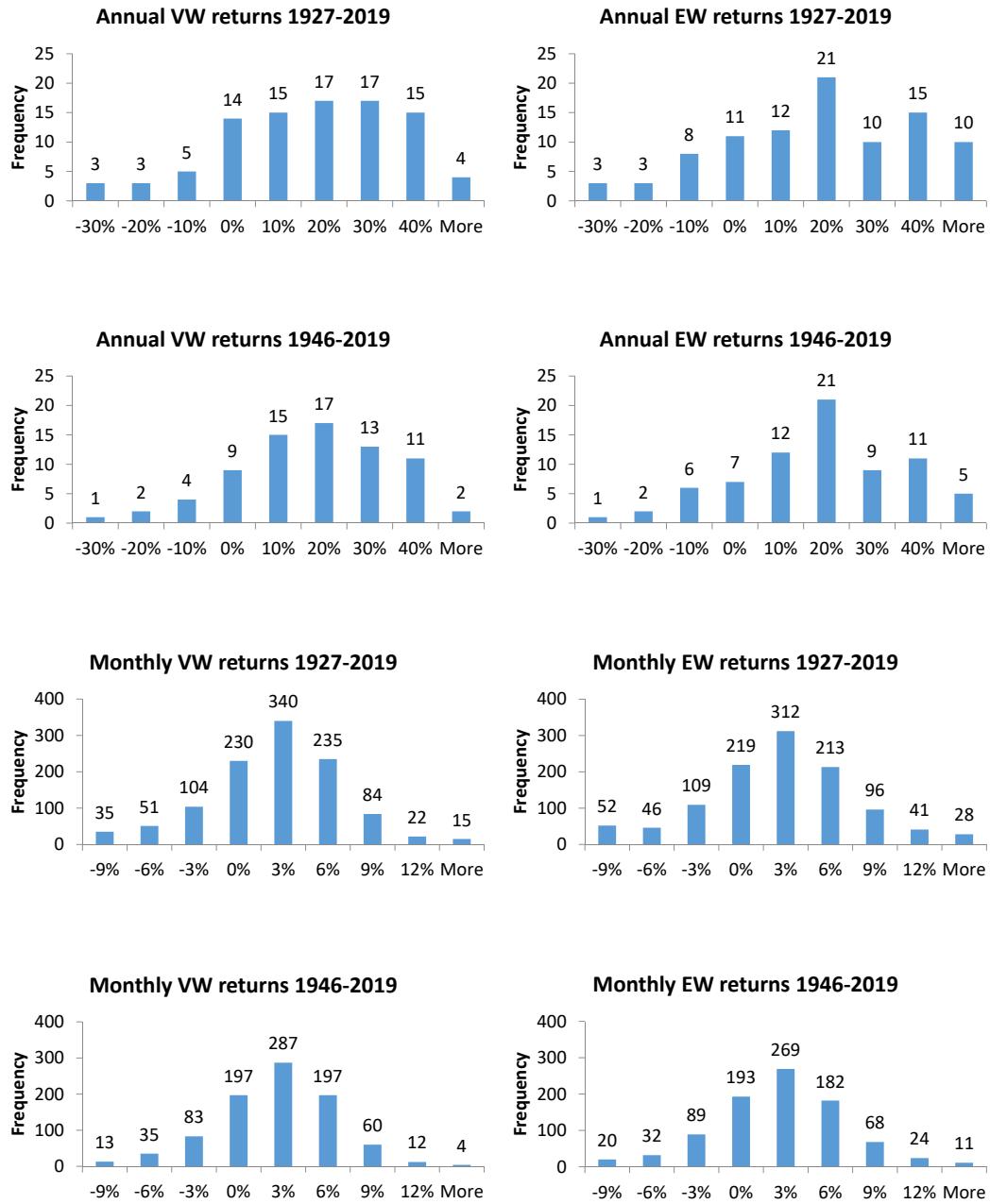


Figure 6.3: Stock return distributions.

Data on the value-weighted (VW) and equally-weighted (EW) returns on the stocks in the S&P 500 index were downloaded from CRSP through WRDS on June 22, 2020. The frequency shown above a given number is the frequency of returns in the return interval ending at that number. For example, the histogram in the upper left corner exhibits 3 observations less than -30% , 3 observations between -30% and -20% , 5 observations between -20% and -10% , etc.

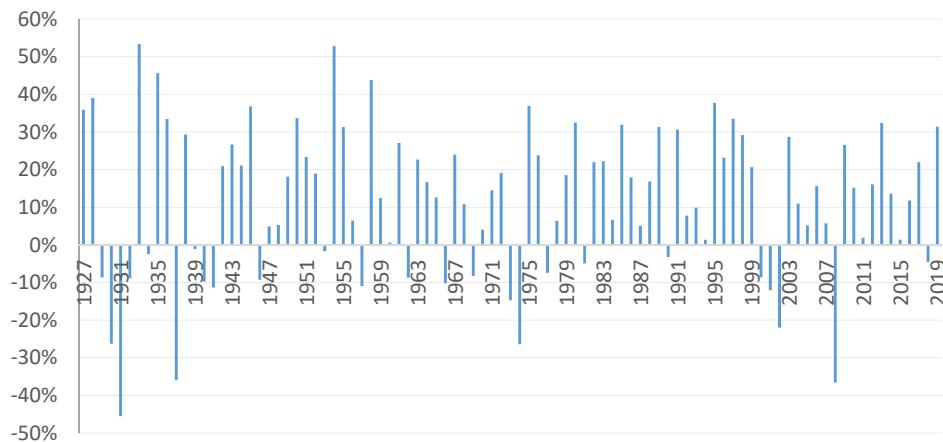


Figure 6.4: Time series of U.S. stock market returns.

The graph shows annual returns on the value-weighted portfolio of stocks in the S&P 500 index over the period 1927-2019. The data were downloaded from CRSP through WRDS on June 22, 2020.

odic nominal return and inflation rate, the real return each period is calculated from (2.15). The average real return is roughly equal to the average nominal return less the average inflation rate. In the post-1946 sample, the arithmetic average real annual return is 8.6% for the value-weighted portfolio and 10.1% for the equally-weighted portfolio. For each portfolio, the standard deviation, skewness, and kurtosis are very similar for the nominal and the real returns.

The stock market volatility of 17-20% per year seems large. First, as we discuss in more detail below, stock returns are considerably more volatile than bond returns. Second, the volatility of stock market returns is considerably larger than the variability of the dividends paid by the companies. Based on a certain data sample, Campbell (2003) estimates the standard deviation of the annual growth rate of real stock market dividends to be around 6%, which is much lower than the return standard deviation of 15.6% in his data.² Since a stock is a right to future dividends of the company, you might expect the volatility of stock prices and dividends to be of a similar magnitude. The high return volatility reflects large variations in stock prices, that is large variations in expected discounted future dividends. The low volatility of dividends suggests that the discount rates involved in the valuation must vary substantially over time. A discount rate for a future dividend consists of a riskfree rate plus a risk premium, and, since riskfree rates are quite stable, the risk premium apparently varies a lot over time. Another indication that stock prices are somewhat detached from dividends is that the correlation between quarterly real dividend growth and real stock returns is only 0.03, but the correlation increases with the measurement period up to a correlation of 0.47 at a 4-year horizon.

Table 6.3 shows statistics for the real returns on the S&P 500 over investment horizons varying from one month to 15 years. The underlying data consists of 1800 monthly real returns from early 1871 to early 2021 downloaded in August 2021 from the homepage of Professor Robert Shiller. The upper part of the table shows statistics for real rates of return, the lower part for real log-returns. Note that the statistics are always calculated

² At much shorter horizons dividend volatility is considerably higher because of seasonality in dividend payments.

Horizon	#Obs	Mean	Std dev	Skew	Kurt	Min	Max	Mean/ T	Std/ \sqrt{T}
<i>Rates of return</i>									
1m	1800	0.65%	4.10%	0.58	18.20	-26.2%	52.4%	7.79%	14.20%
3m	600	2.03%	8.14%	0.55	5.60	-26.2%	58.6%	8.11%	16.27%
6m	300	4.17%	12.22%	0.22	2.29	-36.3%	60.8%	8.33%	17.28%
1y	150	8.54%	17.87%	-0.11	0.01	-38.0%	53.1%	8.54%	17.87%
5y	30	47.97%	48.43%	0.53	0.02	-41.4%	158.4%	9.59%	21.66%
10y	15	127.13%	120.24%	0.52	-0.91	-31.8%	353.1%	12.71%	38.02%
15y	10	190.75%	96.20%	0.37	-0.45	42.7%	349.7%	12.72%	24.84%
<i>Log-returns</i>									
1m	1800	0.56%	4.08%	-0.36	11.36	-30.4%	42.2%	6.77%	14.12%
3m	600	1.69%	7.96%	-0.15	3.14	-30.4%	46.1%	6.77%	15.92%
6m	300	3.39%	11.92%	-0.45	1.90	-45.1%	47.5%	6.77%	16.86%
1y	150	6.77%	17.23%	-0.64	0.61	-47.9%	42.6%	6.77%	17.23%
5y	30	33.85%	33.81%	-0.33	0.16	-53.4%	94.9%	6.77%	15.12%
10y	15	67.71%	57.31%	-0.27	-0.84	-38.2%	151.1%	6.77%	18.12%
15y	10	101.56%	34.58%	-0.39	0.18	35.6%	150.4%	6.77%	8.93%

Table 6.3: Index returns and the investment horizon.

The table shows S&P 500 real return statistics for different investment horizons. The underlying data consist of monthly real returns from February 1871 to January 2021 derived from real total return prices downloaded in August 2021 from the homepage of Professor Robert Shiller. The mean reported is the arithmetic average.

from realizations over non-overlapping periods. For example, the 1800 monthly realized returns lead to 300 realized returns over non-overlapping 6-month periods. Even with 150 years of data, we only have few realizations of returns over non-overlapping 10- or 15-year periods, so the statistics for such long horizons are not that reliable.

The analysis in Section 3.6 showed that, under the assumption that returns in different periods are independent of each other, the expected log-return should grow proportionally with the horizon length T and the standard deviation of the log-return with \sqrt{T} . We can see that average log-returns satisfy this perfectly (at least with the number of decimals shown here), and the pattern of the standard deviation is roughly as it should be. The average rate of return grows faster than T and the standard deviation of the rate of return faster than \sqrt{T} , which is also in line with the expected patterns. The statistics for 15-year returns deviate somewhat from the patterns, but they are based on only 10 observations. Based on Section 3.6 we could expect the skewness and kurtosis to increase with the horizon, but this is not what the data indicate. The skewness shows no clear pattern, whereas the kurtosis appears to decrease with the horizon. The large kurtosis reported for short-term returns are partly due to the older parts of the dataset. For example, the kurtosis of one-month rates of return is 18.20 for the full sample, but only 3.32 for the period from 1946 to 2021, which is close to the kurtosis estimates for monthly real returns in Table 6.2. Still, the fact that the skewness is roughly invariant to the horizon and the kurtosis is decreasing in the horizon indicate that we have not experienced any long periods with extremely large returns. Given the relatively few observations of long-run returns, this could just be a coincidence. The patterns in long-run skewness and kurtosis predicted by Section 3.6 are to a large extent driven by a decent chance of extremely high long-run returns which will materialize from long sequences of substantial short-run returns.

An alternative explanation of the skewness and kurtosis patterns is that returns in reality

Period	Periods lagged							
	1	2	3	4	5	6	9	12
1m	0.26	0.00	-0.05	0.01	0.08	0.04	0.03	-0.03
3m	0.11	0.09	0.08	-0.12	-0.06	-0.05	-0.04	0.04
6m	0.21	-0.14	-0.16	-0.07	0.07	0.04	-0.02	-0.04
1y	0.01	-0.18	0.10	-0.08	-0.11	0.07	-0.06	-0.11
5y	-0.04	-0.15	-0.48	0.00	-0.01	0.37		
10y	-0.48	-0.23	0.40					
15y	-0.33							

Table 6.4: Index return autocorrelations.

The table shows autocorrelations of S&P 500 returns for different investment horizons. The underlying data consist of monthly real returns from February 1871 to January 2021 derived from real total return prices downloaded in August 2021 from the homepage of Professor Robert Shiller. The autocorrelations shown are based on rates of return. The column headings show the number of lags. Some autocorrelations are not shown due to the low number of observations.

are not independent over time as we assumed in Section 3.6. Maybe there is a tendency in the stock market that high returns are followed by low returns and vice versa, which would reduce the skewness and kurtosis of long-run returns. In order to investigate that, Table 6.4 reports autocorrelations for selected investment horizons and lags. For a time series r_1, r_2, \dots, r_T of returns over the same horizon, the autocorrelation with lag k is an empirical estimate of $\rho_k = \text{Corr}[r_t, r_{t+k}]$.

Most of the autocorrelations shown are close to zero, but there are exceptions. For example, the correlation between two subsequent monthly returns is 0.26, so a high return one month is followed by another month with a high return more often than by a month with a low return. And conversely: a low return one month is often followed by another low return next month. Six-month returns exhibit a positive autocorrelation of 0.21 for lag 1, but then negative autocorrelations of -0.14 and -0.16 for lags 2 and 3. Hence, a large six-month return is relatively often followed by another large six-month return and then by low returns the next two six-month periods. A large 5-year return is often followed by low returns 6-15 years later as indicated by the autocorrelations of -0.15 for lag 2 and -0.48 for lag 3. A large 10-year return is often followed by a low return over the next 10-year period. Similarly, for 15-year returns. Recall again that we only have few non-overlapping observations of returns over periods of 10 or 15 years, so the estimates of their autocorrelations are highly uncertain. Nevertheless, the predominantly negative autocorrelations of long-run returns fit well with the relatively few observations of very high long-run returns and lower-than-expected estimates of the skewness and the kurtosis compared to the case of serially independent returns.

Table 6.5 shows the monthly rates of return in and around the six months with the lowest real returns and the six months with the highest real returns since 1871. The right-most columns show the rate of return in the six-month and 12-month period after the month with the extreme return. Two of the worst six months were followed by a year with a return above 30%, whereas the most unfavorable month was followed by a bad year with a rate of return of -10.8%. Two of the best six months were followed by an annual return above 40%, whereas the other four great months were followed by mediocre annual returns. Probably the only conclusion to be drawn from this table is that there is no clear pattern that an exceptionally good month is always or almost always followed by more high returns or by low returns. Similarly, exceptionally bad months are not always

Month	-3	-2	-1	0	1	2	3	1-6	1-12
<i>Worst months</i>									
Nov 1929	6.0	4.2	-10.3	-26.2	5.0	2.4	7.3	21.6	-10.8
Apr 1932	1.2	1.3	1.8	-22.7	-10.1	-11.8	6.1	24.7	32.2
Jan 2008	-6.6	2.5	-4.7	-19.4	-6.8	0.7	-1.5	-9.6	13.7
Mar 2020	2.6	2.9	-0.2	-18.7	5.0	5.9	5.9	27.0	46.2
Dec 1931	-13.8	-12.2	3.4	-17.5	1.2	1.3	1.8	-35.9	-0.9
Sep 1946	-1.4	-8.0	-3.6	-15.3	-3.8	-2.4	2.4	-4.2	-7.1
<i>Best months</i>									
Aug 1932	-10.1	-11.8	6.1	52.4	11.2	-12.6	0.4	-8.5	54.6
May 1933	-9.9	1.1	11.2	29.3	16.7	5.1	-5.4	7.6	9.4
Jul 1938	-4.1	2.3	2.9	20.5	1.0	-4.1	12.4	5.3	2.4
Jun 1933	1.1	11.2	29.3	16.7	5.1	-5.4	-0.5	-5.6	-5.3
Jan 1938	20.5	1.0	-4.1	12.4	0.4	-2.6	-1.2	-14.1	3.2
Apr 2009	-1.5	-7.2	-5.9	12.0	6.3	2.0	1.4	25.8	41.2

Table 6.5: Returns around worst and best months of the S&P 500.

The underlying data consist of monthly real returns from February 1871 to January 2021 derived from real total return prices downloaded in August 2021 from the homepage of Professor Robert Shiller. Month 0 is the month listed in the leftmost column. The table shows the returns in each of the three months before and the three months after that month and, in the rightmost columns, the return in the six and twelve months after the listed month. All returns are real rates of return in percent.

or almost always followed by more low returns or by high returns.

6.5.2 U.S. stocks versus bonds

Investors often focus on the return that the stock market can provide in excess of the riskfree return they could obtain over the same period in terms of the yield of a government bond. In fact, this view is consistent with the leading asset pricing theories covered in later chapters.

Both the average and the standard deviation of stock returns are high compared to returns on government bonds. In the post-1946 sample, the average one-year riskfree return in the U.S. was 3.99%, and subtracting this from the average nominal stock return, we get an average annual excess return or risk premium of 8.41% for the value-weighted and 9.96% for the equally-weighted portfolio. The standard deviation, skewness, and kurtosis are roughly the same for the annual excess returns as for the annual raw returns. The annual returns on 10-year U.S. government bonds exhibited an arithmetic average of 5.59%, corresponding to a risk premium of 6.80% [8.36%] on the value-weighted [equally-weighted] stock market portfolio. The 17-19% volatility on the stock market is considerably higher than the volatility of 3.12% of one-year Treasury bills and 8.39% of 10-year Treasuries. For Baa-rated corporate bonds, the annual returns over 1946-2019 have an arithmetic average of 7.30% and a standard deviation of 7.36%, both numbers significantly lower than for stocks.³

The significant differences in annual returns compound into extreme differences over long investment horizons. Figure 6.5 shows (on a logarithmic scale) how much a \$100 investment in either the stock market, 10-year government bonds, 3-month government

³The data for the 10-year Treasury bonds and the Baa-rated corporate bonds were downloaded on June 24, 2020 from the homepage of Professor Aswath Damodaran at New York University, see <http://pages.stern.nyu.edu/~adamodar/>.

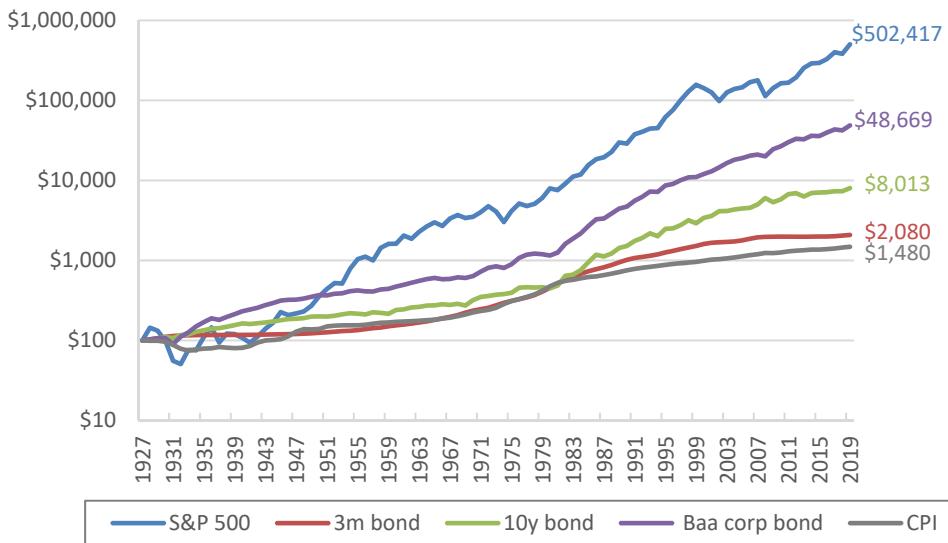


Figure 6.5: Stocks look great in the long run.

The graph shows cumulative returns on U.S. asset classes from 1927 to 2019. The data were downloaded on June 24, 2020 from the homepage of Professor Aswath Damodaran at New York University, see <http://pages.stern.nyu.edu/~adamodar/>.

bills, or Baa-rated corporate bonds at the end of year 1927 grew until the end of year 2019, assuming that returns are reinvested in the same asset class. The stock market transformed \$100 into roughly half a million over the 92-year period, compared to about 49,000 in the corporate bond market and 8,000 in the 10-year government bond market. In the same period, consumer price rose by a factor 14.8. The stock market investment has increased the purchasing power by a factor $502,417/1,480 \approx 339$, whereas the factor is only 5.4 for the 10-year and 1.4 for the three-month government bonds.

6.5.3 Some statistical issues

Based on Figure 6.5, many people would probably conclude that if you want to invest for a long period of time, you should definitely invest in stocks, not bonds. However, this conclusion is premature. Even if you had, say, a 92-year investment horizon, you should not base your decision on the single realization of 92-year returns represented by Figure 6.5 but also take into account that returns over the next 92 years might be different. And, as discussed in the preceding subsection, we only have few non-overlapping observations of returns over, say, 20 years, and we should consider the possibility that the next 20 years are not repeating any of those past 20-year periods.

Furthermore, in some fairly long periods, bonds have in fact outperformed stocks in terms of returns. That was the case in the U.S. between 1929 and 1949, where the geometric average rate of return was 3.21% for 10-year government bonds and only 2.37% for the stock market index. Over the more recent period from 1991 to 2011, the average stock market return of 7.73% barely beat the 6.97% average return on government bonds. So over any given 20-year period, you cannot be sure that stocks will significantly outperform bonds. And over shorter periods, it is even less clear which asset class will dominate the other. Figure 6.6 depicts the 10-year trailing geometric average returns on U.S. stocks, bills, and bonds. For example, the value of 13.44% for stocks in 2019 is the geometric

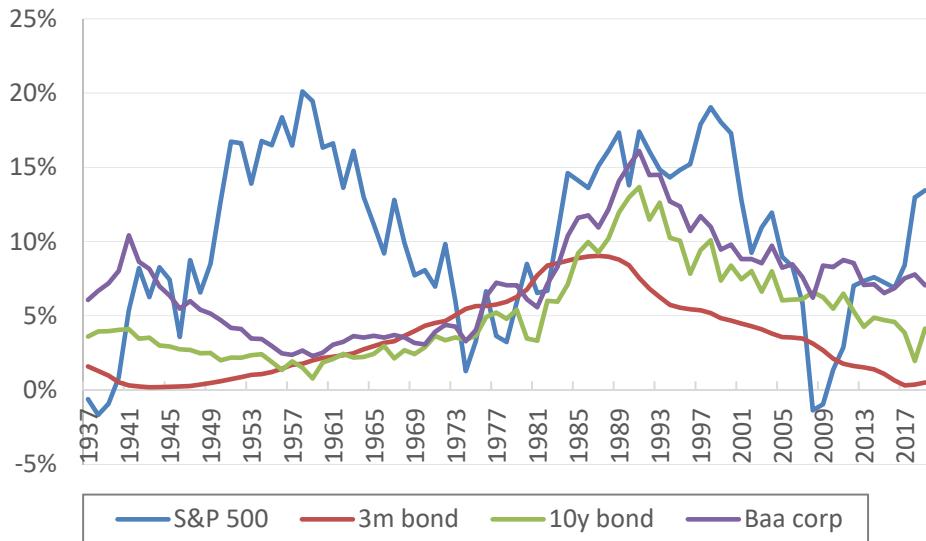


Figure 6.6: Trailing 10-year average returns.

The graphs show the trailing 10-year geometric average rate of return for the S&P 500 stock market index, 3-month Treasury bills, 10-year Treasury bonds, and Baa-rated corporate bonds over the period 1937-2019. The data were downloaded on June 24, 2020 from the homepage of Professor Aswath Damodaran at New York University, see <http://pages.stern.nyu.edu/~adamodar/>.

average return on the S&P 500 index from the end of 2009 to the end of 2019. Note that stocks fall behind all three bonds in several 10-year windows, most recently from 2000 to 2010, and in some 10-year periods the stock market had a negative return, most recently from 1999 to 2009.

The average historical return is not necessarily indicative of the expected returns on the stock market in the future. There are two potential *survivorship biases*. One refers to the possibility that the stock market in the data period has simply performed better than could be expected in the beginning of the period. There are several reasons why this may have happened. In the second half of the 20th Century, the United States and other economies experienced relatively high growth rates. Around 1950, investors in these countries were probably not so sure that the economy the next 50 years would avoid major financial and political crises and outperform numerous other countries. Brown, Goetzmann, and Ross (1995) estimate that, due to this effect, the realized stock returns overstate the ex-ante expected rate of returns significantly by as much as 2-4 percentage points. Note however that in many crises in which stocks do badly, also bonds and deposits tend to provide low returns so it is not clear how big the effect on the expected *excess* stock return is.

Another reason for the better-than-expected performance of the stock market is due to changes in the investment environment beneficial to stocks. Mehra and Prescott (2003) note two significant changes in the U.S. tax system in the period between 1960 and 2000. First, the marginal tax rate for stock dividends has dropped from 43% to 17%. Second, stock returns in most pension savings accounts are now tax-exempt, which was not so in the 1960s, whereas bond returns in savings accounts have been tax-exempt throughout the period. Both changes have led to increased demands for stocks with stock price increases as a result. These changes in the tax rules were hardly predicted by investors and, hence,

they can partly explain the large realized stock returns. Similarly, it can be argued that the reductions in direct and indirect transaction costs and the liberalizations of international financial markets experienced over the last decades have increased the demands for stocks and driven up stock returns above what could be expected *ex ante*. The high transaction costs and restrictions on particularly international investments in the past may have made it impossible or at least very expensive for investors to obtain the optimal diversification of their investments so that even unsystematic risks may have been priced with higher required returns as a consequence. In the same vein, many markets are more liquid so that, most often, investors can easily sell off an asset, which may have been more difficult many years ago. An increase in liquidity tends to drive up the price of the asset as well. Without similar changes in tax codes, trading costs, etc., in the future, the average return in the past is most likely exaggerating the return you can expect in future periods.

The other survivorship bias potentially arises if the average stock market return is based on the performance of an index. Most indices are regularly revised so that some stocks are excluded from the index, while others enter. The stocks excluded from the index are typically stocks that have performed badly in recent periods and often continue to do so in subsequent periods. Conversely, stocks added to the index are typically recent high performers that often continue to provide large returns in subsequent periods. By eliminating low-performing stocks and adding high-performing stocks, the index may overestimate the average return across all stocks. A similar bias is present if you calculate average historical stock returns based only on the stocks traded today, as this would ignore the very low realizations of the return on a stock in a liquidated company (the return is -100% if the stock ends up completely worthless).

6.5.4 International evidence

Both the high average stock returns and the big difference between average stock returns and average bond returns are found consistently across countries. Table 6.6 lists arithmetic averages and standard deviations of real annual returns in percent of stocks, bonds, and bills in selected countries over the period from 1900 to 2018. The table is based on Dimson, Marsh, and Staunton (2019)—updating their earlier book Dimson, Marsh, and Staunton (2002)—which provides an abundance of information about the long-run performance of financial markets around the World. Over that period, the average real U.S. stock return was 8.3% with a standard deviation of 19.9%. The table shows that in any of the listed countries the average return on stocks is much higher than the average return on long-term bonds, which again is higher than the average short-term interest rate. The asset classes are ranked in the same order in terms of their standard deviations.

The international data also contains examples of long periods of poor stock market performance. The blue curve in Figure 6.7 shows the history of the leading Japanese stock index, the Nikkei 225, between January 1984 and September 2015. The index peaked at 38,950 in December 1989 and was at 19,033 at the end of 2015, and it even dipped below 8,000 during 2003. The Nikkei 225 index ignores dividends, but by reinvesting the dividends the index would have followed the red curve. In spite of the index doubling over the last 3 years in the sample, the geometric average annual total rate of return from 1989 to 2015 is a meager -1.6% . In mid-2022, the Nikkei index was still only at around 26,000, 33% below the peak more than 30 years earlier. After a steep increase through 2023 and early 2024, the Nikkei index finally surpassed its December 1989 value in February 2024 and even broke the 40,000 level in March 2024.

Anarkulova, Cederburg, and O'Doherty (2022) use stock index returns from 39 developed countries from 1841 to 2019 to study return distributions over different investment

Country	Stocks		Bonds		Bills	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Australia	8.2	17.5	2.5	13.0	0.8	5.2
Belgium	5.1	23.4	1.6	14.9	0.5	12.5
Canada	7.0	16.9	2.7	10.2	1.5	4.8
Denmark	7.2	20.7	2.8	12.9	2.2	5.9
France	5.7	22.0	1.2	12.9	-1.6	9.4
Germany	8.0	31.4	1.3	15.6	-0.5	12.9
Ireland	6.7	22.9	2.6	14.9	0.9	6.5
Italy	5.8	28.3	0.2	14.6	-2.4	11.1
Japan	8.6	29.3	1.7	19.4	-0.3	13.5
The Netherlands	7.0	21.2	2.2	9.7	0.6	4.8
South Africa	9.2	22.0	2.4	10.4	1.2	6.0
Spain	5.7	21.7	2.6	12.4	0.4	5.7
Sweden	7.9	21.0	3.4	12.6	1.9	6.4
Switzerland	6.2	19.4	2.7	9.3	0.8	4.8
United Kingdom	7.2	19.7	2.7	13.5	1.2	6.3
United States	8.3	19.9	2.4	10.3	0.9	4.6

Table 6.6: Risk and return around the World.

The table shows (arithmetic) means and standard deviations (StdDev) of annual real rates of return in different countries over the period 1900-2018. Numbers are in percentage terms. Bonds mean long-term (approximately 20-year) government bonds, while bills mean short-term (approximately 1-month) government bonds. The bond and bill statistics for Germany exclude the year 1922-23, where the hyperinflation resulted in a loss of 100 percent for German bond investors. For Swiss stocks the data series begins in 1911. Source: Tables 1, 3, and 4 in [Dimson, Marsh, and Staunton \(2019\)](#) that extend the data period covered by [Dimson, Marsh, and Staunton \(2002\)](#) (Copyright 2019 Elroy Dimson, Paul Marsh and Mike Staunton).

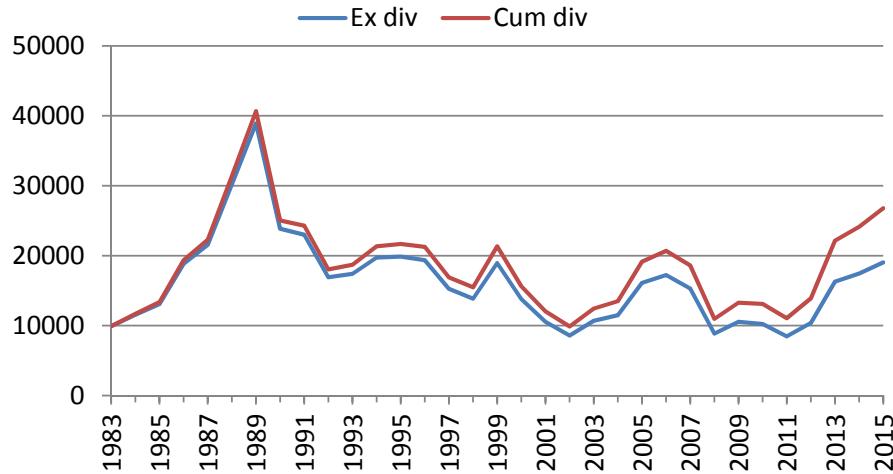


Figure 6.7: Stocks may disappoint.

The blue curve shows the Nikkei 225 index at the end of each year from 1983 to 2015, retrieved from Yahoo Finance at August 16, 2016. The red curve includes dividend payments on the Nikkei index stocks and is constructed by adding for each year the difference between the annual return of the Nikkei 225 Total Return and the Nikkei 225 read off a graph published in the fact sheet `nikkei_factsheet.pdf` which is included in the supplementary material for these lecture notes.

horizons. The long-run return distributions they find are similar to those that we generated in Section 3.6 assuming serially uncorrelated normally distributed monthly log-returns. For example, over a 30-year horizon the real rate of return on a country's stock market index has an expectation of 6.38 or 638% and a standard deviation of 13.76 or 1376%. There is a 12.1% risk of a negative real return over 30 years. The 1st and 5th percentiles are -86% and -53% , showing a non-negligible risk of highly negative real returns in the long run. On the other hand, the return distribution has a long right tail that reflects a good chance of very large long-run returns. The 95th and 99th percentiles are 2230% and 5245%, respectively.

6.5.5 Time-varying moments and predictability of returns

Stock return moments are not constant over time. For example, the stock market volatility varies over time and tends to cluster so that there are periods with low volatility and periods with high volatility, a phenomenon apparently first noted by [Mandelbrot \(1963\)](#). The volatility peaks in periods with high political or macroeconomic uncertainty, as documented by [Bloom \(2009\)](#), among others. Stock volatility exhibits positive autocorrelation over several days. Stock volatility tends to be negatively correlated with the return so that volatility is high in periods of low returns and *vice versa*, as originally observed by [Black \(1976\)](#). The volatility is far from perfectly correlated with the price so it has a separate stochastic component not linked to the stochastic price.

Let us now focus on variations in average returns. Several empirical studies indicate that average stock market returns seem to vary counter-cyclically, i.e. that average returns are higher in bad times than in good times. Such a counter-cyclical pattern is also seen in average excess stock returns and realized Sharpe ratios (the average excess return divided by the standard deviation of the return), see [Fama and French \(1989\)](#) and [Lettau and](#)

Ludvigson (2010).

Intuitively, counter-cyclical average stock returns can be explained by investors being more reluctant to hold stocks in bad times due to increased risk aversion or financial and regulatory constraints. Here, bad times could mean recessions and good times booms as measured by the GDP growth rates, but bad and good times could also be defined from other indicators of the business cycle, macroeconomic variables, or variables that seem to indicate whether stock prices are currently high or low. Of course, it can be highly valuable to identify variables that can predict whether we can expect high or low stock market returns over some future period as such variables can be used as signals for entering or exiting the market. Saying that stock returns are predictable by some variable X means that you can predict next period's return r_{t+1} more precisely if you condition on X_t than if you do not.

We should, however, realize that it seems unlikely that we can find very strong predictors of the stock market, i.e. variables that with a large precision predict the future stock market returns. First, there is a statistical issue. Even if we assume that the expected return is constant over time, it is impossible to estimate its value from past returns with a large precision due to the large variability of returns and the limited length of our data sample, cf. the discussion in Section 3.7.2. Intuitively, it is even harder to estimate a time-varying expected return. For example, if we divide all the periods for which we have return observations into good times and bad times, we have fewer observations in each category than the total number of observations, which further reduces the precision of the estimates of expected returns. Secondly, if a strong predictor was easy to find, many investors would profit immensely from timing the stock market by trading on the predictor, and we do not observe that in practice.

As a first try, let us check whether the realized return in the most recent period is a good predictor of next period's return. The first-order autocorrelation captures any such pattern. Based on the annual observations over 1927-2019, the autocorrelation is a tiny 0.01 for the value-weighted returns on the S&P 500 stocks and -0.06 for the equally-weighted returns. For the post-1946 period, the numbers are -0.08 and -0.21 , respectively. For the monthly observations since 1927 [1946], the autocorrelation is 0.08 [0.03] for the value-weighted returns and 0.14 [0.09] for the equally-weighted returns. These numbers suggest short run *momentum* in the stock market, i.e. a tendency that good months are followed by good months and bad months are followed by bad months. In contrast, in the slightly longer run, the market shows signs of *reversal* in the sense that good years tend to be followed by bad years and vice versa. However, the autocorrelations are of a small magnitude. A good month in the stock market is quite often followed by a bad month, even though it is slightly more likely to be followed by another good month. Of course, you could also allow the length of the past period used in the prediction to differ from the length of the period you try to forecast the return in. For example, Moskowitz, Ooi, and Pedersen (2012) show that the return over the past 12 months positively predicts the return over the next month, both for stock indices in nine countries as well as for various bonds, currencies, and commodities. Further examples of estimates and discussions hereof can be found in DeBondt and Thaler (1985), Fama and French (1988), Jegadeesh and Titman (1993, 2001), Campbell, Lo, and MacKinlay (1997, Sec. 2.8), and Cochrane (2005, Sec. 20.1). In sum, the data shows signs of time-series predictability in stock index returns in the form of momentum in the short run and reversal in the medium-long run.

As discussed in previous sections, analysts often study valuation ratios such as the price-dividend ratio or the price-earnings ratio. If you believe that the valuation ratio varies around a certain long-term level, then deviations from that level should predict

future stock returns. A low current valuation ratio predicts high future returns. A high current valuation ratio predicts low future returns. However, valuation ratios can deviate substantially from the historical average over an extended period of time. As shown in Figure 6.1 the S&P 500 price-dividend ratio has been far above the historical average from 2010 and until the time of writing (March 2022). This can be partly explained by low interest rates and the shift from cash dividends to share buybacks. As long as the low interest rates continue and companies do not shift back to cash dividends, it is not clear that the price-dividend ratio should fall to a level near the historical average.

Extensive research has investigated whether future stock returns and dividend growth can be predicted by the current price-dividend ratio or the dividend yield. The evidence is mixed as the conclusion depends on the country and time period used in the study, cf. Campbell and Shiller (1988a), Campbell and Ammer (1993), Kothari and Shanken (1997), Ang and Bekaert (2007), Cochrane (2008), Chen (2009), van Binsbergen and Kooijen (2010), Engsted and Pedersen (2010), and Rangvid, Schmeling, and Schrimpf (2014). Motivated by the growing importance of share buybacks as a method for companies to compensate their shareholders, several studies have replaced cash dividends by a measure of total payouts or net total payouts to shareholders and focus on the price-payout ratio or its reciprocal, the payout yield. These variables do seem to predict future stock returns better than the traditional price-dividend ratio and the cash dividend yield, cf. Boudoukh, Michaely, Richardson, and Roberts (2007) and Eaton and Payne (2017).

The price-earnings ratio is also frequently used as an indicator of whether stocks are cheap or expensive and may therefore predict the direction of stock price movements in the near future. Let us look at the CAPE for the U.S. stock market as introduced towards the end of Section 6.2.1 with the observed time series shown in Figure 6.2. Does the CAPE predict stock returns?

Figure 6.8 illustrates the relation between the CAPE for the S&P 500 index at the beginning of each year and the realized real index return over that year. Each point corresponds to a year in the 141-year period from 1881 to 2021. The line superimposed is the best straight line and the equation for the line is shown in the figure. The line slopes downwards so a high CAPE tends to be followed by a lower-than-average stock return next year. The slope is estimated to be -0.44% and suggests that for every one-point increase in the CAPE, the expectation of next year's return is lowered by 0.44 percentage points. The 95% confidence interval of the slope is $[-0.86\%, -0.02\%]$, and the slope estimate has a p-value of 4.2% and is thus only borderline significantly different from zero in a statistical sense. The R^2 of this regression is a meager 2.9% reflecting the fact that many of the observation points are located far away from the line. Prediction errors are sizeable. For example, the orange square represents the year 1935, where the predicted excess return was 10.72% but the realized excess return was a stunning 52.71%. Conversely, the green square represents 1931, where the realized excess return was -38.03% , but the CAPE-based prediction was 8.44%. In fact, the average yearly miss is 14.28 percentage points! If, instead, we use the 1692 monthly observations and regress monthly real stock returns on the beginning-of-month CAPE, the estimated slope is -0.02% with a p-value of 15%, but the R^2 is only 0.12% and, on average, the predicted monthly return misses the realized return by 3.08%. Implementing market timing strategies based on CAPE seems to be very risky business. The CAPE is a weak predictor of stock market returns, both in statistical and economical terms.

Other variables that have been claimed to predict stock market returns—based on a specific data sample—include

- the short-term interest rate (Ang and Bekaert 2007);

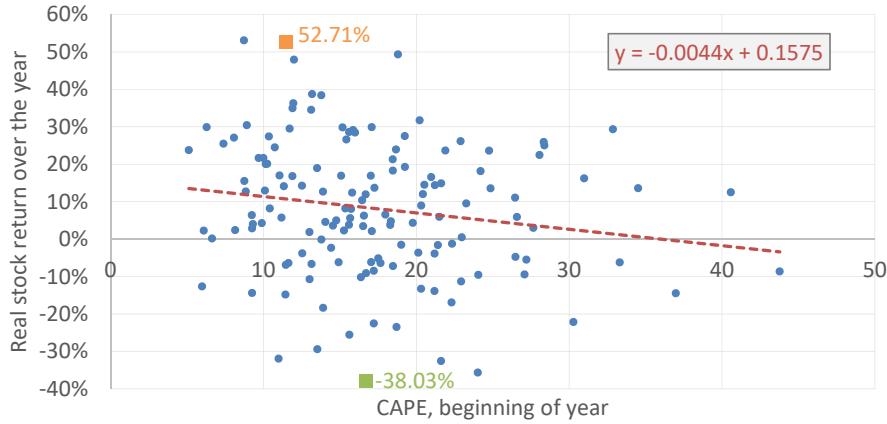


Figure 6.8: Cape-able of predicting stock returns?

Each data point represents a year between 1881 and 2021. The horizontal axis shows the CAPE of the U.S. stock market at the beginning of the year, i.e. the real value of the stocks in the index divided by their average real earnings over the past ten years. The vertical axis shows the real return on the stock market in each year. The data were downloaded on March 4, 2022 from the homepage of Professor Robert Shiller at Yale University, see <http://www.econ.yale.edu/~shiller/data.htm>. The dashed line indicates the regression line that has the equation shown in the box.

- the consumption-wealth ratio (Lettau and Ludvigson 2001);
- the fourth-quarter growth rate in personal consumption expenditures (Møller and Rangvid 2015);
- the housing collateral ratio (Lustig and van Nieuwerburgh 2005);
- the ratio of stock prices to GDP (Rangvid 2006);
- the ratio of aggregate labor income to aggregate consumption (Santos and Veronesi 2006);
- the output gap, that is the difference between actual GDP and the potential GDP (Cooper and Priestley 2009).

Most of the predictors seem relatively weak even in the original paper that claims predictive power, and many of the predictors show even weaker predictive power in other data sets. Note that predictors are not necessarily independent of each other. A variable highly correlated with a predictor or with a combination of predictors is also likely to predict returns. There are various statistical challenges in measuring predictability, and academics are still debating whether predictability is there or not, see for example Ang and Bekaert (2007), Goyal and Welch (2008), Campbell and Thompson (2008), Cochrane (2008), Boudoukh, Richardson, and Whitelaw (2008), and Lettau and van Nieuwerburgh (2008), and, if so, which predictors work and why. Koijen and van Nieuwerburgh (2011) and Goyal, Welch, and Zafirov (2021) survey the recent research on return and dividend predictability.

6.6 The cross section of stock returns

There are systematic differences in the average returns of different stocks. Over the last four decades, a large empirical literature has documented that average stock returns

depend on various characteristics of the stock or the issuing company. This section gives an updated view of some of these return patterns. We focus on U.S. stocks where the homepage of Professor Kenneth French at Dartmouth College supplies relevant return data based on all stocks traded at the NYSE, AMEX, and NASDAQ exchanges. However, similar return patterns have been shown to exist in many other markets.

The documented return patterns are evidence of *cross-sectional predictability* in the sense that we can predict the returns on individual stocks relative to other stocks based on information on how the stock ranks relative to the other stocks according to some currently observable criteria. Cross-sectional predictability patterns are potentially useful for determining the weights of different assets in a portfolio. Based on each asset's rank relative to the other assets, you may hope to identify assets with high expected returns (should have high weights, other things equal) and assets with low expected returns (low weights). Maybe you even want to set up a long-short investment strategy with long positions in the assets with high expected returns and short position in the assets with low expected returns.

Before we present some statistical evidence, a number of caveats should be acknowledged. First, there might be rational explanations for such differences in expected returns. As shown below, the historical evidence points to stocks of small companies (measured by market capitalization) having larger average returns than stocks of large companies. But this might be due to investors perceiving small stocks to be riskier than large stocks since they would then demand being compensated for the higher risk by getting higher average returns. More generally, if the predictor somehow tracks variations in risks or in risk premiums, expected stock returns—and thus average realized returns—*should* vary with the value of the predictor. This issue is discussed further in Chapters 10 and 11.

Secondly, if you want to implement a trading strategy where you, say, every month hold stocks with high expected returns over the next month and short stocks with low expected returns, you have to rebalance the portfolio regularly as some stocks may move from one category to the other over time and because you want specific weights on the different stocks. The portfolio is rebalanced by trading stocks and trades involve transaction costs. The return differences we show below do not take costs into account. Net of transaction costs, the profitability of the trading strategies is lower and, for some strategies, maybe even insignificant. Also, some strategies may involve trading in relatively illiquid stocks that might be difficult to sell when your strategy commands you to do so.

Thirdly, just because a certain variable or stock characteristic is statistically significant for explaining return differences in a given sample, there is no guarantee it is useful in the future. Data mining is a concern here. If, for a given data set, researchers try enough variables, eventually one will seem to be a useful predictor, even if the returns are completely unpredictable. If you look long enough for a pattern, you will see one. Also, a variable may really predict stock returns in a given data set without any rational risk-based explanation, but simply due to some systematic mispricing. Once such a predictive pattern becomes publicly known, many investors will presumably try to exploit it through appropriate trading strategies (long undervalued, short overvalued assets), and eventually prices should change such that the mispricing disappears and the variable no longer predicts returns. [McLean and Pontiff \(2016\)](#) consider 97 predictors of the cross section of stock returns that have all been shown statistically significant for some particular data set. They find that the returns on portfolios designed to profit from the predictability are, on average, significantly lower out of sample. Furthermore, the profitability of such portfolios drops substantially after publication of the study identifying the predictor. But, while weaker out of sample and post-publication, the predictive power does not completely

disappear. Similar results are reported by Linnainmaa and Roberts (2018).

6.6.1 Value vs. growth stocks

The book-to-market ratio of a listed company is simply the ratio of the book value of the stocks of the company to the market value of these stocks. Stocks in companies with high book-to-market ratios are called *value stocks*. Stocks in companies with low book-to-market ratios are called *growth stocks*; if the market value of equity is high relative to the book value, it is probably because the company has substantial and valuable options to grow over time and boost future earnings and dividends.⁴ Rosenberg, Reid, and Lanstein (1985) and Fama and French (1992) showed that value stocks provide a higher average return than growth stocks. The difference in average returns is the so-called *value premium*.

At the end of June each year, a book-to-market ratio is calculated for each stock using the book equity for the most recent fiscal year and the prevailing market value of the shares outstanding at the end of the previous calendar year. Then stocks are divided into portfolios based on their book-to-market value. A frequently applied split involves the following three portfolios:

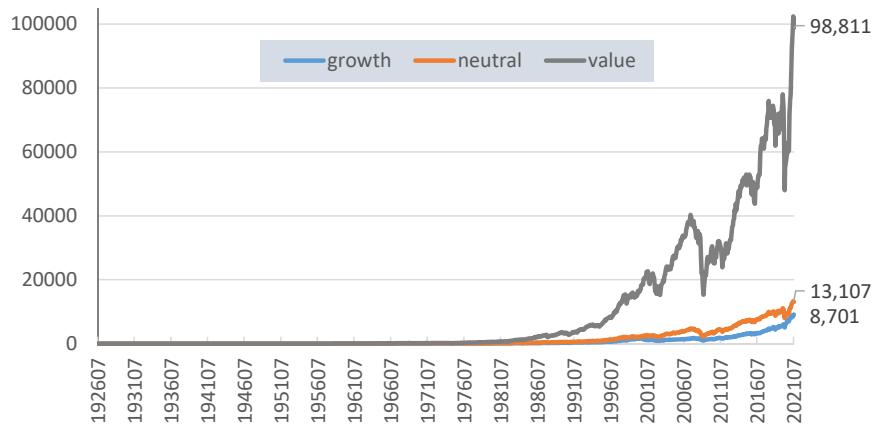
- The value stock portfolio consisting of the stocks with 30% largest book-to-market values.
- The growth stock portfolio comprised by the stocks with 30% smallest book-to-market values.
- The neutral stock portfolio with the remaining stocks, i.e., the stocks having a book-to-market in the middle 40% interval.

Stocks are re-categorized at the end of June each year. For each portfolio monthly returns are calculated as the value-weighted average of returns on the stocks in the portfolio. Professor French offers a time series of returns on the portfolios going back to July 1926.

Figure 6.9 shows the cumulative returns from investing \$1 on July 1, 1926 in the different portfolios until the end of June 2021. Over this 95-year period an investment in value stocks grew by a factor of 62,831 and was thus much more profitable than an investment in growth stocks or neutral stocks. Growth stocks have underperformed relative to neutral stocks. Of course, obtaining these returns demands a lot of trading in a lot of stocks. While this is cumbersome and potentially costly due to trading fees, there is no reason to believe that the rebalancing the value portfolio would be much more costly than rebalancing the growth portfolio. And nowadays investors can buy exchange-traded funds mimicking value and growth portfolios at a low cost and with very little hassle.

Table 6.7 lists statistics based on monthly returns in percent from 1926/07 to 2021/06. In terms of the average monthly return, value stocks clearly beat neutral stocks which in turn beat growth stocks. If we annualize the average return by compounding the arithmetic monthly mean as in (3.83), we get $(1.0095)^{12} - 1 \approx 0.1196 = 11.96\%$ for the growth portfolio, 12.57% for the neutral portfolio, and 16.24% for the value portfolio. The same ordering holds for the standard deviations so in that sense value stocks are also riskier than the other stocks. If we annualize the standard deviations using (3.84), we get 20.62% for growth, 22.10% for neutral, and 29.11% for value stocks. Still, value stocks have a higher Sharpe ratio than neutral and growth stocks. Both the minimum and maximum observations as well as the skewness and kurtosis estimates also reveal that value returns tend to be more extreme and “less normally” distributed than growth returns. These

⁴Using the book-to-market ratio to define value and growth stocks is common in academic research, but many practitioners use other measures such as the earnings yield (the reciprocal of the price-earnings ratio).

**Figure 6.9: Value vs. growth stocks: cumulative returns.**

The figure is based on monthly returns on three book-to-market based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

claims are backed by Figure 6.10 which shows histograms of the 1140 monthly returns for each of the three stock portfolios. As shown in the boxes in the figure, value stock returns are more frequently extremely large or extremely small than returns on growth or neutral stocks.

While value stocks have significantly outperformed growth stocks since 1926, the relative performance of the different stock types have varied substantially within this period. Table 6.8 shows the cumulative returns in percent within each decade. In some decades value stocks have offered much larger returns than growth stocks, but both in the 1930's, the 1990's, and the 2010's growth stocks have been a better investment than value stocks. Table 6.9 shows the estimates of the average return and the standard deviation, both for the full 95-year period and for the most recent 30 and 10 years. The dominance of value stocks is clearly less convincing over the more recent decades. In fact, in the 2011/07-

	Growth	Neutral	Value
Mean, arithmetic, in %	0.95	0.99	1.26
Mean, geometric, in %	0.80	0.84	1.01
Standard deviation, in %	5.33	5.67	7.22
Sharpe ratio	0.127	0.127	0.137
Skewness	-0.15	1.19	1.63
Kurtosis	5.39	17.20	18.79
Minimum, in %	-28.85	-28.25	-34.68
Maximum, in %	33.80	52.03	67.04

Table 6.7: Value vs. growth stocks: return statistics.

The table is based on monthly returns on three book-to-market based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

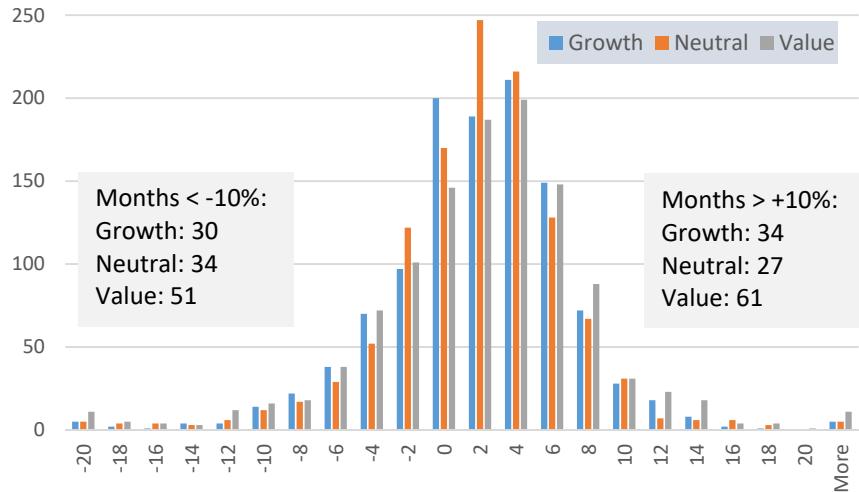


Figure 6.10: Value vs. growth stocks: return distribution.

The figure is based on monthly returns in percent on three book-to-market based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

2021/06 period growth stocks have delivered both a larger average return and a lower standard deviation than value stocks.⁵

6.6.2 Small vs. large stocks

Another well-investigated question is whether small or large stocks tend to deliver higher returns. Again, small and large refer to the market capitalization of the company, i.e., the market value of all stocks issued by the company. The pioneering study by Banz (1981) and many subsequent studies draw a clear conclusion: small stocks offer higher average returns than large stocks.

To get an updated view, we can again use the data supplied by Professor French. First, we consider quintile portfolios, i.e., stocks are divided into five portfolio each covering 20% of the stocks. The portfolio Lo-20 (or Quint1) consists of the 20% stocks with the lowest stock market capitalization and so forth. The classification of stocks is revised once a year. Figure 6.11 depicts the cumulative returns from investing \$1 on July 1, 1926 until the end of June 2021 in the five size portfolios. The portfolio of large stocks has shown a much worse performance than the other portfolios, whereas the best performing portfolio is the Quint2 portfolio of “small, but not very small” stocks followed by the very small stocks in Quint1 and the medium-sized stocks in Quint3.

Table 6.10 lists statistics based on monthly percentage returns from 1926/07 to 2021/06. In line with the figure, the large-stock portfolio (Hi-20 or Quint5) stands out by having a lower average return than the other quintile portfolios. But it also has a lower standard deviation and generally less extreme returns. The Sharpe ratio of the large-stock portfolio is actually larger than that of the much more risky small-stock portfolio, but the largest

⁵ Arnott, Harvey, Kalesnik, and Linnainmaa (2021) argue that the recent underperformance of value stocks is partly due to the book-to-market ratio failing to capture the increasing importance of intangible assets.

	Growth	Neutral	Value
1931-40	19	-27	-28
1941-50	95	204	391
1951-60	409	450	633
1961-70	116	118	217
1971-80	45	138	262
1981-90	309	396	555
1991-00	480	295	348
2001-10	-12	56	73
2011-20	308	228	188

Table 6.8: Value vs. growth stocks: decade returns.

The table shows cumulative returns per decade on three book-to-market based portfolios on the U.S. stock market. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>. The numbers shown are cumulative returns in percent.

	Growth	Neutral	Value
<i>1926/07-2021/06</i>			
Mean, arithmetic, in %	0.95	0.99	1.26
Standard deviation, in %	5.33	5.67	7.22
Sharpe ratio	0.127	0.127	0.137
<i>Most recent 30 years</i>			
Mean, arithmetic, in %	1.02	0.97	1.08
Standard deviation, in %	4.33	4.29	5.26
Sharpe ratio	0.189	0.180	0.167
<i>Most recent 10 years</i>			
Mean, arithmetic, in %	1.44	1.00	1.16
Standard deviation, in %	3.99	4.30	5.69
Sharpe ratio	0.350	0.222	0.195

Table 6.9: Value vs. growth stocks: risk and return estimates.

The table is based on monthly returns on three book-to-market based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

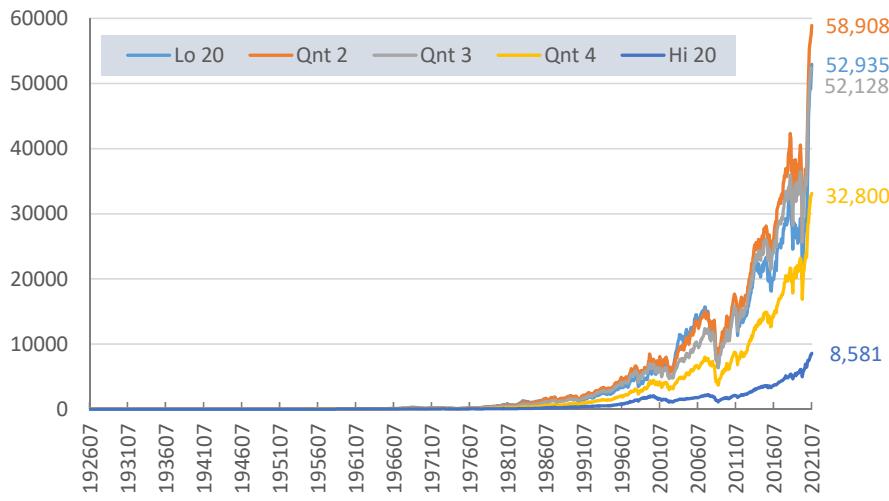


Figure 6.11: Size portfolios: cumulative returns.

The figure is based on monthly returns on five size-based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

Sharpe ratio is obtained for the portfolios of medium-sized stocks in Quint3 and Quint4. Figure 6.12 shows histograms for returns on three portfolios of the smallest 30%, the largest 30%, and the remaining 40% of stocks (with five quintile portfolios, the figure would be too messy). While some resemblance with the normal distribution is obvious, the number of extreme return observations is large especially for the small-stock portfolio.

Table 6.11 shows the percentage returns over each decade. Again notice the large variations across decades. In some decades small stocks have substantially outperformed large stocks (e.g., the 1940's), whereas the opposite is seen in other decades (e.g., the 1990's). Table 6.12 contains estimates of the average return and the standard deviation, both for the full 95-year period and for the most recent 30 and 10 years. For all sample periods, small stocks seem more risky than large stocks, whereas the ranking of the estimates of

	Lo 20	Quint 2	Quint 3	Quint 4	Hi 20
Mean, arithmetic, in %	1.32	1.24	1.18	1.11	0.93
Mean, geometric, in %	0.89	0.91	0.91	0.87	0.76
Standard deviation, in %	8.90	7.57	6.84	6.19	5.12
Sharpe ratio	0.118	0.128	0.134	0.135	0.129
Skewness	2.61	1.56	0.96	0.68	0.11
Kurtosis	26.50	16.79	12.32	10.79	7.35
Minimum return, in %	-33.07	-31.36	-31.67	-29.77	-28.69
Maximum return, in %	95.92	70.46	58.36	52.74	36.95

Table 6.10: Size portfolios: return statistics.

The table is based on monthly returns on five size-based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

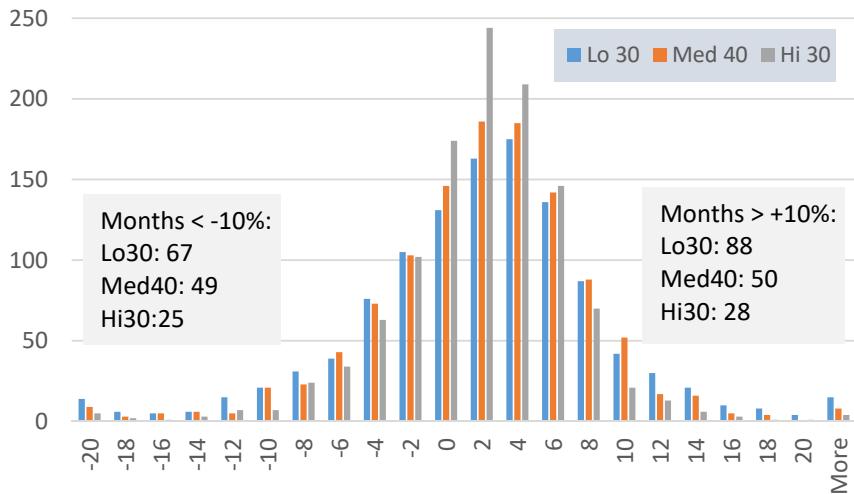


Figure 6.12: Size portfolios: return distribution.

The figure is based on monthly returns in percent on three size-based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

the mean return varies somewhat. The old perception that small stocks outperform large stocks is not backed by the recent evidence, especially considering the high standard deviation of small stock returns. And recall that trading small stocks is relatively more costly than trading large stocks.

6.6.3 Portfolios sorted both on size and book-to-market

Now let us consider double-sorted portfolios where stocks are divided into portfolios based on both the book-to-market value and the market capitalization. Table 6.13 shows statistics on 25 such portfolios over the full sample period from 1926/07 to 2021/06 in the left part and over the most recent 10-year period in the right part. Here ‘BM’ refers to the book-to-market value of the company and ‘ME’ to the market value of the equity. For example, the ‘Grow/Large’ portfolio consists of the stocks having a book-to-market value among the 20% lowest values of all stocks—thus being growth stocks—and a market capitalization among the 20% largest values of all stocks.⁶ As above, stocks are reallocated to portfolios once a year.

The upper panel of Table 6.13 depicts the arithmetic average return in percent, and the lower panel the standard deviation in percent. For example, over the full sample period the ‘Grow/Large’ portfolio has delivered an average monthly return of 0.95% with a standard deviation of 5.32%. In the full sample the largest average return (1.64%) is for the small-value stock portfolio in the upper-right corner, and the smallest average return (0.91%) is for the small-growth portfolio in the upper-left corner. To some extent the standard deviation follows the average return so that higher risk is rewarded, but especially the

⁶With this procedure, the number of stocks differs across portfolios. An alternative would be a conditional sort where you first sort along one dimension, say the book-to-market value. Then within each book-to-market portfolio you sort along the other dimension—size—putting 20% in each quintile. This produces 25 portfolios with roughly the same number of stocks in each, but then for example some of the stocks in the ‘Grow/Small’ portfolio could be larger than some of the stocks in the ‘Value/ME2’ portfolio.

	Lo 20	Quint 2	Quint 3	Quint 4	Hi 20
1931-40	37	42	46	15	-5
1941-50	520	378	278	235	127
1951-60	473	485	477	433	440
1961-70	322	257	195	165	103
1971-80	133	143	132	113	64
1981-90	216	375	395	400	389
1991-00	270	264	319	349	458
2001-10	95	70	63	58	-12
2011-20	162	227	243	277	254

Table 6.11: Size portfolios: decade returns.

The table shows cumulative returns per decade on five size-based portfolios on the U.S. stock market. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>. The numbers shown are cumulative returns in percent.

	Lo 20	Quint 2	Quint 3	Quint 4	Hi 20
<i>1926/07-2017/08</i>					
Mean, arithmetic, in %	1.32	1.24	1.18	1.11	0.93
Standard deviation, in %	8.90	7.57	6.84	6.19	5.12
Sharpe ratio	0.118	0.128	0.134	0.135	0.129
<i>Most recent 30 years</i>					
Mean, arithmetic, in %	1.19	1.11	1.11	1.12	0.95
Standard deviation, in %	6.27	5.83	5.33	4.96	4.19
Sharpe ratio	0.159	0.156	0.171	0.186	0.180
<i>Most recent 10 years</i>					
Mean, arithmetic, in %	1.26	1.21	1.17	1.26	1.27
Standard deviation, in %	6.17	5.86	5.38	4.89	3.94
Sharpe ratio	0.197	0.199	0.209	0.248	0.311

Table 6.12: Size portfolios: risk and return estimates.

The table is based on monthly returns on five size-based portfolios on the U.S. stock market from 1926/07 to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

	Full sample: 1926/07-2021/06					Recent sample: 2011/07-2021/06				
	Grow	BM2	BM3	BM4	Value	Grow	BM2	BM3	BM4	Value
<i>Arithmetic mean, in %</i>										
Small	0.91	1.02	1.28	1.45	1.64	1.23	1.44	1.10	1.14	1.40
ME2	0.95	1.24	1.27	1.32	1.52	1.54	1.40	1.15	1.00	1.13
ME3	1.03	1.19	1.21	1.30	1.40	1.27	1.36	1.05	1.17	0.95
ME4	1.04	1.06	1.14	1.24	1.32	1.47	1.26	1.15	1.01	1.09
Large	0.95	0.92	1.00	0.93	1.22	1.53	1.20	1.14	0.87	1.21
<i>Standard deviation, in %</i>										
Small	12.08	9.78	8.91	8.30	9.28	7.23	6.45	5.86	5.96	7.02
ME2	7.95	7.47	7.23	7.40	8.68	6.56	5.46	5.76	5.82	6.82
ME3	7.37	6.46	6.47	6.90	8.49	5.77	5.14	5.26	5.80	6.69
ME4	6.21	6.06	6.36	6.84	8.66	4.96	4.93	5.03	5.50	6.33
Large	5.32	5.26	5.60	6.59	8.54	4.11	3.68	4.11	4.39	6.80

Table 6.13: Estimates for size-B/M portfolios. The upper panel shows the mean and the lower panel the standard deviation of the monthly returns on 25 double-sorted portfolios. ‘BM’ refers to the book-to-market value of the equity of the company and ‘ME’ to the market capitalization. The left part of the table is based on the full sample, the right part only on the 10 year period up to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

small growth stocks do not obey this rule as they deliver low average returns with a high standard deviation. For the most recent 10-year period from 2011/07 to 2021/06, the patterns are less pronounced, but the best performers were probably medium-sized stocks with a not-too-high book-to-market value.

6.6.4 Momentum: portfolios sorted on prior returns

In Section 6.5.5 we reported some evidence on time-series momentum in the stock index. Momentum also plays a key role in cross-sectional predictability. An asset that has outperformed similar assets in the recent past tends to outperform the same assets in the near future as documented for U.S. stocks by Jegadeesh and Titman (1993), Rouwenhorst (1998), and Asness, Moskowitz, and Pedersen (2013) among others. This suggest implementing a *winners-minus-losers strategy* (sometimes called the up-minus-down strategy) in which you take a long position in a portfolio of recent relative winners and a short position in a portfolio of recent relative losers. Here, a winner (loser) is a stock that has provided a higher (lower) return than a value-weighted index of all stocks over a certain recent period, typically chosen somewhere between 3 and 12 months. Of course, both portfolios are to be revised from time to time.

The winner and loser portfolios can be constructed using equal weights on all winners and losers, respectively, or by using weights reflecting the degree to which the stocks have over- or underperformed in the recent past, see, e.g., Asness, Moskowitz, and Pedersen (2013). Some suggest limiting the strategy to the more extreme recent winners and losers and thus not trade in all stocks. Similar cross-sectional patterns were found in non-U.S. stock markets and markets for other financial assets. While a winners-minus-losers strategy seems to be profitable in most periods, there are also periods in which the strategy performs really badly, often following markets decline and when markets are highly volatile. Such,

apparently rare, momentum crashes are documented by Daniel and Moskowitz (2016).

Also note that the data suggest that the winners-minus-losers strategy and the high-minus-low strategy complement each other. Their returns tend to be negatively correlated so that a combination of the two strategy can generate a high expected return even at a low risk, cf. Asness, Moskowitz, and Pedersen (2013).

To document the performance of recent winners and losers, we again use data from Professor French's homepage. At the beginning of every month, the return over the prior 12 months is calculated for each stock, excluding the most recent month in order to avoid the one-month reversal often seen in stock returns and which may be caused by liquidity or market microstructure effects. By double-sorting stocks on market capitalization and the prior return, 25 portfolios are formed. The portfolios are rebalanced every month.

Table 6.14 shows the mean and the standard deviation of the monthly returns on these 25 portfolios. The left part of the table is based on the long sample from 1927/01 to 2021/06, whereas the right part is based on only the most recent 10 years. In the column headings, 'PRI' refers to the prior return with 'Lose' ['Win'] indicating the 20% worst [best] performing stocks. In the full sample, the largest average return (1.87%) was delivered by the portfolio of small stocks with the best recent performance. The smallest return (0.23%) is seen for the portfolio of large stocks with the worst recent performance. In each size bucket, average returns are increasing in the past returns. This is a clear illustration of return momentum: Winners tend to stay winners, losers tend to stay losers. Within each size bucket, the standard deviations exhibit more or less the opposite pattern of the average returns so the portfolio of recent winners have both the largest average return and a small standard deviation. The recent losers have a low average return and a high standard deviation.

In the more recent sample, the patterns are less clear. Large, recent losers still seem to perform poorly, but recent winners have been outperformed by stocks with average prior returns. Recent losers still seem to be more risky than other stocks.

6.6.5 Portfolios sorted on variance

If the variance or standard deviation of an asset's return is an appropriate measure of its risk, you might expect that assets with a high variance offer a high average return to attract investors. However, this relation does not seem to hold up empirically. The homepage of Professor Kenneth French provides time series of returns of portfolios formed on size and recent return variance dating back to 1963/07. More specifically, at the end of each month the variance of the most recent 60 daily returns is calculated for each stock. By double-sorting on market capitalization and variance, 25 portfolios are formed and held until next month where realized returns are recorded and the portfolios are rebalanced.

Table 6.15 shows means and standard deviations of the monthly returns on the 25 size-variance portfolios with results for the full sample to the left and for the most recent 10 years to the right. In the column headings, 'VAR' refers to the prior return variance with 'Low' ['High'] indicating the 20% stocks with the lowest [highest] recent return variance. For the full sample, the average return in each size bucket is lowest for the high-variance portfolio, except for the large stocks. In particular, the small and medium-sized stocks with the 20% highest variances stand out by delivering a much lower average return than all other stocks. Moreover, the stocks with the high recent variance tend to exhibit a high variance also in the coming month, so the poor average returns are accompanied by a high return variance. While these patterns certainly seem puzzling, we cannot conclude that stock markets are not working properly or that traders make systematic mistakes. First of all, the return variance might not be the only relevant quantity in the assessment of an

	Full sample: 1926/07-2021/06					Recent sample: 2011/07-2021/06				
	Lose	PRI2	PRI3	PRI4	Win	Lose	PRI2	PRI3	PRI4	Win
<i>Arithmetic mean, in %</i>										
Small	0.95	1.41	1.63	1.64	1.89	1.12	1.15	1.36	1.44	1.45
ME2	0.72	1.17	1.25	1.45	1.69	1.20	1.35	1.25	1.30	1.57
ME3	0.68	1.08	1.14	1.22	1.58	1.02	1.25	1.19	1.06	1.32
ME4	0.68	0.95	1.09	1.21	1.54	1.02	1.20	1.32	1.20	1.28
Large	0.25	0.80	0.90	1.03	1.24	0.99	1.30	1.38	1.08	1.32
<i>Standard deviation, in %</i>										
Small	10.76	9.41	8.81	8.79	8.73	8.51	5.86	5.67	5.50	6.48
ME2	9.79	8.10	7.18	7.18	7.76	8.00	6.12	5.51	5.54	6.45
ME3	9.42	7.56	6.78	6.18	6.83	8.08	6.11	5.23	5.10	5.69
ME4	9.45	7.21	6.28	6.03	6.39	8.09	5.82	4.95	4.41	5.03
Large	10.36	6.41	5.70	5.21	5.76	7.61	4.83	4.09	3.80	4.31

Table 6.14: Estimates for size-momentum portfolios.

The upper panel shows the mean and the lower panel the standard deviation of the monthly returns on 25 double-sorted portfolios. ‘PRI’ refers to the return over the prior period from 12 to 2 months before portfolio formation and ‘ME’ to the market capitalization. The left part of the table is based on the full sample, the right part only on the 10 year period up to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

asset’s riskiness. In the recent 10-year sample the differences in average returns are less pronounced, but the stocks with low past variance still seem attractive due to their low standard deviation of future returns.

6.6.6 Perspectives

The above analysis outlines some patterns in which stocks offer high average returns and which stocks offer low average returns. Most patterns are less clear for more recent data samples, which might indicate that investors have started trading on the patterns in the older sample. In any case, we cannot at this point conclude whether any such characteristics-based patterns are due to mispricing in the market or due to the characteristics proxying for risks that investors demand compensation for bearing. We will return to these questions in Chapters 10 and 11.

6.7 Individual stocks

In this section we discuss historical evidence on individual stock returns. Since a broad stock market index is some average across stocks, you might expect that the return patterns of a typical stock is similar to the patterns of the market index, but there are notable differences.

First, individual stocks are much more volatile than the market index. Goyal and Santa-Clara (2003) report that the average volatility on individual stocks is four times the volatility of an equally-weighted stock market index, which also demonstrates that shocks to individual stocks can be diversified away to a large degree by forming portfolios. According to Professor Aswath Damodaran at New York University, returns are generally more volatile for small stocks than for large stocks: among the 10% smallest stocks issued

	Full sample: 1963/07-2021/06					Recent sample: 2011/07-2021/06				
	Low	VAR2	VAR3	VAR4	High	Low	VAR2	VAR3	VAR4	High
<i>Arithmetic mean, in %</i>										
Small	1.40	1.55	1.49	1.22	0.40	1.45	1.32	1.38	1.23	1.13
ME2	1.28	1.42	1.45	1.34	0.79	1.15	1.32	1.34	1.37	1.44
ME3	1.14	1.23	1.36	1.27	0.91	1.22	1.08	1.27	1.24	1.21
ME4	1.09	1.16	1.21	1.17	0.95	1.28	1.29	1.32	1.23	1.15
Large	0.87	0.99	0.96	0.90	0.93	1.18	1.37	1.19	1.25	1.38
<i>Standard deviation, in %</i>										
Small	4.18	5.76	6.64	7.69	9.47	4.75	6.04	7.07	8.14	10.15
ME2	4.10	5.32	6.00	6.88	8.82	4.52	5.60	6.29	7.04	8.75
ME3	3.75	4.87	5.44	6.24	8.10	4.06	5.18	5.63	6.22	7.86
ME4	3.75	4.48	5.13	5.79	7.60	3.66	4.22	5.03	5.81	6.83
Large	3.45	3.99	4.48	5.04	6.59	3.31	3.68	4.39	4.68	5.53

Table 6.15: Estimates for size-variance portfolios.

The upper panel shows the mean and the lower panel the standard deviation of the monthly returns on 25 double-sorted portfolios. ‘VAR’ refers to the prior return variance and ‘ME’ to the market capitalization. The left part of the table is based on the full sample, the right part only on the 10 year period up to 2021/06. The data were downloaded on September 15, 2021 from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>.

by U.S. companies, the median standard deviation of the stock return is 50%, whereas it is only 20% among the 10% largest stocks.⁷

Secondly, the shape of the return distribution is different for individual stocks than indices. Daily returns on stock market indices tend to be negatively skewed so that the peak of the return distribution exceeds the mean and the left tail is longer than for a normal distribution. In contrast, daily returns on individual stocks are roughly symmetric (zero skewness) or slightly positively skewed. See [Albuquerque \(2012\)](#) for an empirical documentation and an attempt to reconcile the differences between individual stocks and indices.

Thirdly, it is more difficult to predict returns on individual stocks than returns on characteristics-sorted portfolios or the entire market index, which can be explained by the higher variance of individual stocks leading to even lower precision in estimates of expected returns. Nevertheless, [Goyal and Jegadeesh \(2018\)](#) report that momentum is also present in the time series of individual stocks.

[Bessembinder \(2018\)](#) has performed a comprehensive analysis of monthly individual stock returns from the leading U.S. exchanges over the period from 1926 to 2016. The analysis documents a number of remarkable results:

1. Only 48.4% of all the monthly returns on individual stocks are positive and only 47.8% of them exceed the one-month riskfree return in the same month, measured by the one-month Treasury bill rate at the beginning of the month. Even though the average monthly stock return of 1.13% is positive, the median stock return is not. The distribution of all monthly stock returns has a positive skew of 6.955.
2. Many individual stocks are only exchange-listed for a relatively short period. The median listing period in the 90-year data set is only 7.5 years. Some stocks disappear

⁷See <http://pages.stern.nyu.edu/~adamodar/> under “Risk Measures by Market Cap Class,” accessed on July 5, 2022.

- due to mergers or because majority stockholders, for whatever reason, prefer to take the stock off the exchange. Other stocks are delisted by the exchange in response to delinquency, insufficient capital, or very poor performance. Over the full “lifetime” (i.e. listing period), only 9.8% of these delisted stocks generated a positive return, and the median lifetime return of the delisted stocks is -91.95% . And about 12% of all stocks listed at some point in time lost 95% or more of their value before leaving the exchange.
3. Only 49.5% of the stocks exhibit positive lifetime returns and only 42.6% deliver a lifetime return above the riskfree return over the same period as measured by a roll-over strategy in one-month Treasury bills. The median lifetime stock return is -2.29% , whereas the mean is as high as 18,747% corresponding to the value being multiplied by a factor 187.47. The distribution of lifetime returns has an extreme positive skewness of 154.8, and only 2.2% of stocks in the database have a lifetime return above the mean but many of them much higher than the mean.
 4. The total lifetime wealth creation of all stocks in the data set is about \$35 trillion in excess of the wealth gained by investing the same capital in Treasury bills. Only 1092 stocks (about 4% of all stocks) account for all of this wealth creation. Just 90 stocks have delivered half of the total wealth creation. Only five stocks account for 10% of the overall wealth creation, namely Exxon Mobile, Apple, Microsoft, General Electric, and IBM.

In sum, the U.S. data shows that, in the long run, a small number of stocks perform extremely well, most stocks do rather poorly, and a large number of stocks do very poorly losing most of their value. At first, these observations might seem to conflict with the traditional assumptions of either rates of return or log-returns being normally distributed. However, as shown in Section 3.6, if these assumptions are made on monthly returns, the distribution of long-run rates of return does have a highly positive skewness, a median far below the mean, and a large probability of negative outcomes. [Bessembinder, Chen, Choi, and Wei \(2019\)](#) find the same patterns in the 1990-2018 returns on stocks both in the United States and in the other 40 countries included in their study, and [Fang, Marshall, Nguyen, and Visaltanachoti \(2021\)](#) provide similar results. These findings highlight the potentially large value to be gained from stock selection: if you can predict which of the stocks currently traded will be among the future long-run winners, you can obviously make significant profits. On the other hand, if you are not able to identify the future winners, you are probably best off by investing in a broad and well-diversified portfolio of stocks since then you have a good chance of including one or more of the future stars.

6.8 Correlations

First, let us consider correlations among individual stocks. The correlation at a given point in time between the returns on two assets can be estimated by the empirical correlation of recent return observations. For example, the current correlation between two stocks could be measured by the empirical correlation between the most recent 52 realized weekly returns on the two stocks. Next week the correlation estimate would again be based on the then most recent 52 observations, so a rolling window is applied.

[Jones and Kincaid \(2014\)](#) report an average pairwise correlation of around 0.36 between the 30 stocks in the Dow Jones Industrial Average index based on rolling monthly returns over the period 1950-2008. Whether 6, 12, or 18 months are used in the rolling window has little effect on the correlation estimate. The correlation estimate varies from around 0.09 to 0.67 through the data period, typically with an inverse relation between the correlation and the direction of the stock market index. This illustrates a general finding that stock

correlations typically (not always) rise in bad times and fall in good times. An obvious implication is that diversification of risk becomes more difficult in bad times, which is presumably exactly when you are most keenly interested in reducing risk.

The correlation between groups or portfolios of stocks is typically larger than between individual stocks due to the diversification of firm-specific risk. The cross-country correlations have increased in recent decades ([Quinn and Voth 2008](#)), reducing (but not eliminating) the benefits from international diversification. The increase in correlation can be explained by the globalization trend leading more, especially larger, corporations to sell products and buy various production inputs in other countries. Hence, corporations located in different countries are increasingly exposed to the same shocks. This trend is supported by the reduction in trade barriers and the gradually increasing openness of formerly isolated countries and markets. Countries are becoming more interconnected and so are large corporations around the world. More on stock correlations later in this chapter.

Next, we focus on the correlation between the stock market and government bonds. Price changes (and thus returns) of stocks and long-term bonds tend to be positively correlated as shown by [Shiller and Beltratti \(1992\)](#) and [Campbell and Ammer \(1993\)](#) among others. This makes sense if we think of the price of an asset being the sum of the expected future dividends discounted by an appropriately risk-adjusted discount rate. If the discount rates of the stock and the long-term bonds move together, then, assuming expected dividends do not change, the prices should move together.

However, the riskfree discount rate of the long-term bond and the risk-adjusted discount rate of the stock do not have to move in lockstep as the equity risk premium might vary with the riskfree interest rates. Recent studies show that the stock-bond correlation varies considerably over time and is even negative in some periods, see for example [Ilmanen \(2003\)](#), [Cappiello, Engle, and Sheppard \(2006\)](#), and [Andersson, Krylova, and Vähämaa \(2008\)](#). Figure 6.13 shows a picture of the variations in the stock-bond correlation. Here, the correlation at a given point in time is computed as the sample correlation between the most recent 36 monthly returns on the S&P 500 index and on U.S. Treasury bills and bonds of maturities of either 1 year (blue) or 10 years (orange). The stock-bond correlation was mostly positive from 1965 to around 2000 and then mostly negative until another sign shift in 2022-23.

Empirical research suggests that the stock-bond correlation is partly driven by expectations about economic growth and inflation expectations. When growth expectations are positive, companies are expected to increase their earnings, leading to higher stock prices. In periods of strong growth, the central bank tends to tighten monetary policy to avoid a too strong economic activity, and this is implemented by increasing interest rates which lead to reduced bond prices. Of course, the increasing stock prices and decreasing bond prices produce a negative correlation. Low positive or even negative stock-bond correlations are often seen around stock market crashes. Increased stock market uncertainty may induce flight-to-safety portfolio adjustments from stocks to bonds that simultaneously drive stock prices down and bond prices up. This indicates that diversification across stocks and government bonds works better in bad times than in good times. On the other hand, high inflation uncertainty or unexpected inflation increases often lead to lower prices of bonds and stocks, generating a positive stock-bond correlation. It should be noted, though, that the determinants of the level and the dynamics of the stock-bond correlation are yet not fully understood.

Finally, we consider the correlations among a broader range of asset classes. Various financial institutions regularly publish their estimates of expected returns and volatilities of major asset classes, as well as their pairwise return correlations. As an example, Table 6.16

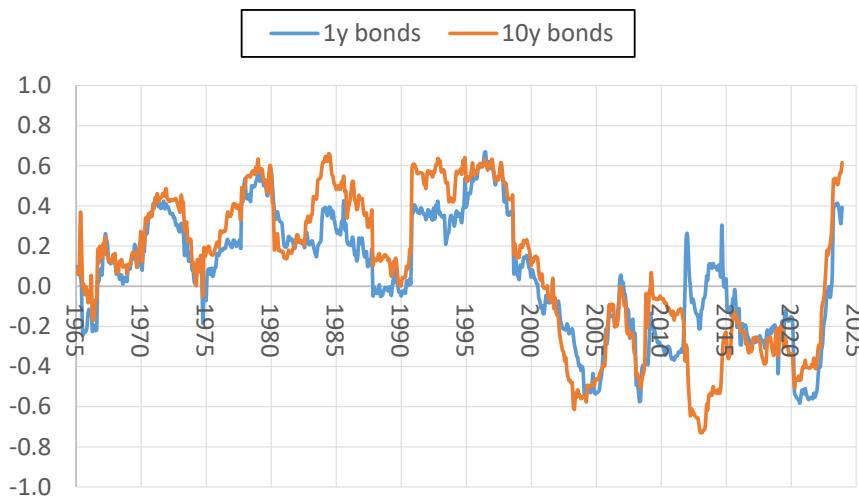


Figure 6.13: The correlation between stock and bond returns.

Each curve shows for each month an estimate of the correlation between stock returns and bond returns calculated as the sample correlation of the returns over the past 36 months. The stock returns are based on the value-weighted returns on the S&P 500 index including dividends. The bond returns are based on Treasury bills and bonds of maturities of 1 year (blue) or 20 years (orange). The return data spans the period from January 1962 to December 2023 and were downloaded from CRSP (the Center for Research in Security Prices) on May 17, 2024, with the bond returns coming from the CRSP Fixed Term Indexes.

shows the estimates for 20 major asset classes published in late 2023 by J.P. Morgan Asset Management in their report “2024 Long-Term Capital Market Assumptions,” available on their homepage. In fact, their report contains estimates for 59 asset classes, but for clarity we focus on 20 classes. The estimates are intended to apply to the subsequent 10 to 15 years and are apparently based on a mix of statistical estimates based on past returns and an assessment of the macro-economic and political outlook. The 20 asset classes consists of 8 fixed-income classes (including inflation), 5 equity classes, and 7 alternative asset classes. Most names should be self-explanatory. ‘Cash’ refers to short-term deposits at the prevailing short-term interest rates. High-yield bonds which are bonds issued by companies with poor credit rating, in contrast to investment grade corporate bonds. EAFE is short for Europe, Australasia, and the Far East and refers to the MSCI EAFE stock index that covers around 800 stocks from 21 developed countries in these regions (16 from Western Europe plus Australia, Hong Kong, Japan, New Zealand, Singapore, and Israel).

First, we see that the expected annual return is generally higher for equity than for alternatives and bonds, maybe with the exception of private (non-listed) equity which is traditionally categorized as an alternative investment. The same ordering tends to hold for the volatility, i.e., the standard deviation of the annual return. The high volatility of long-term Treasury bonds may seem surprising as such bonds are considered safe, but over a one-year horizon long-term bond prices can fluctuate substantially because of the high duration.

Next, turning to the correlations we see that the different stock classes are highly positively correlated with estimates 0.68 or higher, and in particular the three classes of U.S. stocks have pairwise correlations of at least 0.90. Within the fixed income category, the pairwise correlations of different asset classes are typically smaller and some corre-

		Annualized Volatility (%)									
		Expected Return 2024 (%)					Inflation				
		U.S. Inflation		U.S. Cash		U.S. Long Treasuries		TIPS		U.S. Inv Grade Corp Bonds	
Alternatives	Equities	U.S. High Yield Bonds	6.83	8.36	0.01	-0.07	-0.06	-0.06	0.46	0.64	1.00
	World Government Bonds	5.03	6.91	-0.15	0.10	0.74	0.63	0.67	0.67	0.30	1.00
Equities	U.S. Large Cap	8.19	16.19	0.01	-0.04	-0.11	-0.10	0.29	0.46	0.75	0.26
	U.S. Mid Cap	9.08	18.13	0.01	-0.05	-0.16	-0.12	0.28	0.47	0.78	0.22
Equities	U.S. Small Cap	9.07	20.44	-0.02	-0.07	-0.19	-0.19	-0.18	0.36	0.71	0.15
	EAFE Equity	10.58	17.64	-0.01	0.01	-0.08	-0.09	0.30	0.52	0.77	0.39
Equities	Emerging Markets Equity	10.77	21.20	0.00	0.02	-0.07	-0.06	0.32	0.52	0.72	0.38
	Private Equity	11.46	20.06	0.09	0.00	-0.36	-0.42	0.17	0.36	0.73	0.07
Alternatives	U.S. Real Estate	8.02	10.60	0.31	-0.17	-0.25	-0.16	0.11	0.05	0.38	-0.13
	European Real Estate	8.06	12.84	0.31	-0.16	-0.35	-0.31	0.17	0.14	0.53	-0.03
Alternatives	Global Infrastructure	7.38	11.24	0.20	0.01	-0.25	-0.28	0.23	0.25	0.57	0.16
	Diversified Hedge Funds	5.16	5.80	0.09	0.01	-0.33	-0.25	0.19	0.34	0.61	0.02
Alternatives	Commodities	5.31	18.00	0.27	-0.04	-0.17	-0.23	0.27	0.21	0.46	0.23
	Gold	5.43	16.93	-0.01	0.09	0.37	0.31	0.48	0.38	0.14	0.52

Table 6.16: Return moments for major asset classes.

The table shows estimates of expected returns, volatilities (standard deviations), and correlation for a range of major asset classes in U.S. dollar terms. The values are taken from the 2024 Long-Term Capital Market Assumptions published by J.P. Morgan Asset Management. See <https://am.jpmorgan.com/dk/en/asset-management/institutional/insights/portfolio-insights/lrcma/>.

lations are even negative, e.g. between U.S. high-yield bonds and Treasury bonds. Also, assets in the alternative category have mostly modest pairwise correlations. Looking across asset classes, stock returns are expected to be negatively correlated with Treasury bond returns but positively correlated with corporate bonds, especially the high-yield bonds as these bonds—like stocks—are very sensitive to shocks to the corporate sector. Stocks are expected to have near-zero correlations with gold prices, moderate correlations with real estate, infrastructure, and commodities, and sizeable correlations with private equity and diversified hedge funds. Most alternative classes have low or even negative correlations with cash and Treasury bonds but modestly positive correlations with corporate bonds. Overall, the correlation matrix suggests that substantial diversification gains can be obtained by carefully combining various asset classes into a portfolio.

6.9 Diversification of stock portfolios: an example

Chapter 4 showed the idea of diversification: by combining risky assets into a portfolio, we can reduce the risk as measured by the variance or standard deviation of the rate of return—at least if the returns of these assets are not perfectly correlated. In other words, we can diversify risk away by spreading our investments. In particular, Figure 4.10 on page 142 illustrated how the portfolio standard deviation declines with the number of assets in an equally-weighted portfolio, assuming that all assets have identical standard deviations and all pairs of assets have identical correlations. Here we consider a related and more practically oriented exercise.

Table 6.17 lists the 30 stocks in the Dow Jones Industrial Average index as of May 2024 ordered alphabetically by their ticker code. Based on the 120 monthly returns on each stock from January 2014 to December 2023, we have calculated the standard summary statistics, i.e., the (arithmetic) mean, standard deviation, skewness, and kurtosis. These moments are annualized without taking compounding into account by applying Eqs. (3.55) and (3.56) with $n = 12$. The annualized means range between 3.0% (for 3M) and 27.3% (for Microsoft), whereas the annualized standard deviations range between 15.4% (for Johnson & Johnson) and 38.0% (for Boeing). The skewness is negative for 15 and positive for 15 stocks. The kurtosis is negative for 7 and positive for 23 stocks. In addition, we have calculated the sample correlation matrix of the 30 stocks, and for each stock calculated the simple average of its 29 correlation coefficients with the other stocks. The average correlation ranges from 0.253 (for Walmart) to 0.526 (for Honeywell).

Now, we explore the properties of equally-weighted portfolios of these 30 stocks. We focus first on the standard deviation of the portfolio return. If you invest only in the first stock, namely Amazon, of course the standard deviation of your return will be 31.4% since this is the standard deviation of Amazon's stock returns. Now consider adding the second stock, American Express, and investing 50% in each of them. With a standard deviation of 25.7%, American Express stocks is less volatile than Amazon stocks, but the standard deviation of your portfolio return also depends on the correlation between the two stocks, which is estimated to be 0.322. Applying Eq. (4.4), the portfolio standard deviation is

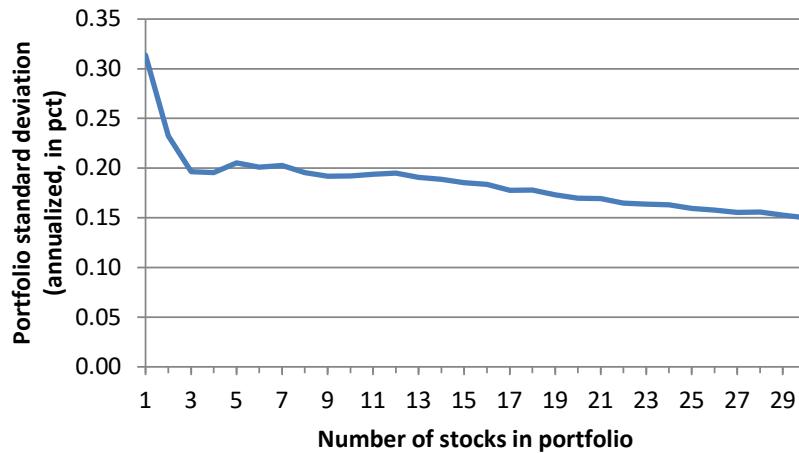
$$\sigma_p = \sqrt{\left(\frac{1}{2}\right)^2 \times (31.4)^2 + \left(\frac{1}{2}\right)^2 \times (25.7)^2 + 2 \times \frac{1}{2} \times \frac{1}{2} \times 0.322 \times 31.4 \times 25.7} \approx 23.3,$$

which is considerably smaller than the standard deviation of Amazon. Continuing this way, we sequentially add one more stock following alphabetic ordering and calculate the return standard deviation of an equally-weighted portfolio of 3, 4, . . . , 30 Dow Jones stocks. Figure 6.14 shows how the portfolio standard deviation varies with the number of stocks in the portfolio. Also this figure indicates a substantial diversification gain (i.e., reduction in

Ticker	Company	Mean	Std Dev	Avg Corr	Skewness	Kurtosis
AMZN	Amazon.com Inc	0.252	0.314	0.296	0.085	0.052
AXP	American Express Co	0.120	0.257	0.445	-0.027	0.217
AMGN	Amgen Inc	0.148	0.237	0.339	0.062	-0.002
AAPL	Apple Ince	0.280	0.279	0.372	-0.038	-0.042
BA	Boeing Co	0.154	0.380	0.380	0.014	0.335
CAT	Caterpillar Inc	0.189	0.294	0.397	0.058	0.060
CRM	Salesforce Inc	0.204	0.315	0.323	0.230	0.180
CSCO	Cisco Systems Inc	0.140	0.235	0.387	0.016	-0.012
CVX	Chevron Corp	0.097	0.275	0.400	0.176	0.167
DIS	Walt Disney Co	0.063	0.278	0.423	0.125	0.094
DOW	Dow Inc	0.102	0.281	0.441	-0.109	0.158
GS	Goldman Sachs Group Inc	0.135	0.281	0.442	0.053	0.025
HD	Home Depot Inc	0.191	0.216	0.421	-0.007	0.031
HON	Honeywell International Inc	0.128	0.202	0.526	0.143	0.235
IBM	Int'l Business Machines Corp	0.058	0.225	0.379	-0.031	0.103
INTC	Intel Corp	0.133	0.278	0.332	0.037	0.112
JNJ	Johnson & Johnson	0.093	0.154	0.401	-0.010	0.018
JPM	JPMorgan Chase & Co	0.164	0.242	0.440	-0.034	0.078
KO	Coca-Cola Co	0.081	0.161	0.402	-0.189	0.086
MCD	McDonald's Corp	0.153	0.168	0.422	0.064	0.113
MMM	3M Co	0.030	0.211	0.482	-0.062	-0.024
MRK	Merck & Co Inc	0.132	0.189	0.271	-0.024	0.064
MSFT	Microsoft Corp	0.273	0.213	0.384	0.052	0.025
NKE	Nike Inc	0.144	0.251	0.379	-0.061	0.007
PG	Proctor & Gamble Co	0.100	0.159	0.306	-0.005	-0.017
TRV	Travelers Companies Inc	0.117	0.195	0.391	-0.010	0.055
UNH	UnitedHealth Group Inc	0.230	0.198	0.307	0.079	0.001
V	Visa Inc	0.183	0.205	0.478	0.019	-0.026
VZ	Verizon Communications Inc	0.036	0.171	0.283	-0.002	-0.041
WMT	Walmart Inc	0.107	0.181	0.253	-0.071	0.062

Table 6.17: Annualized return statistics for the Dow Jones 30 stocks.

For each of the 30 stocks included in the Dow Jones Industrial Average index on May 15, 2024, the table shows summary statistics based on the time series of observed monthly rates of return from January 2014 to December 2023. From left to right, the columns show the ticker code, the company name, the annualized mean rate of return (12 times the monthly mean), the annualized standard deviation ($\sqrt{12}$ times the monthly standard deviation), the average correlation with the rates of return on the other 29 index constituents, the annualized skewness (monthly skewness divided by $\sqrt{12}$), and the annualized kurtosis (monthly kurtosis divided by 12). The monthly rates of return were downloaded from CRSP on May 15, 2024.

**Figure 6.14: Dow diversification.**

The table shows the portfolio standard deviation with an increasing number of stocks. The stocks are those included in the Dow Jones Industrial Average stock index in May 2024. The portfolios are equally weighted and formed as explained in the text. The standard deviations are annualized but based on monthly returns over the period from January 2014 to December 2023.

standard deviation) when adding a stock to a portfolio with few stocks, whereas adding a stock to a portfolio that already consists of many stocks has a smaller effect on the standard deviation. Furthermore, the figure confirms that there is a lower bound on the standard deviation we can obtain, which we know from Section 4.4 is related to the covariances among the assets. Compared to the curves in Figure 4.10, the curve in Figure 6.14 is less smooth because the next stock added might be more or less risky than the stocks already included in the portfolio and might have high or low correlations with these stocks.

Together with the standard deviations, Table 6.18 also lists the mean, the skewness, and the kurtosis of the different equally-weighted portfolios, with all moments being annualized as explained above. The portfolio mean stabilizes rather quickly as additional stocks are included. All the portfolios of two or more stocks have negative skewness, even though half of the stocks have positive skewness. All the portfolios have positive kurtosis. Also note that there seems to be no clear pattern in how the portfolio skewness and kurtosis depend on the number of stocks in the portfolio. These observations confirm that the higher-order moments of portfolios are hard to predict from the higher-order moments of the stocks in the portfolio, as discussed in Section 4.3.

6.10 Exercises

Exercise 6.1. Hypothetics Inc. has just paid the annual dividend to shareholders based on \$5 earnings per share over the past year. It is a firm policy of Hypothetics to use a plowback ratio of 40%. The company's return on investment is 15%. You have estimated the risk-adjusted discount rate to be 12%.

- (a) What is the forward price-earnings ratio?
- (b) What are next year's expected earnings per share?
- (c) Based on your answers to the two preceding questions, what is the fair stock price per share?
- (d) How big is the dividend just paid per share?
- (e) What is the price-dividend ratio?

Stocks	Mean	Std dev	Skew	Kurt	Stocks	Mean	Std dev	Skew	Kurt
1	0.252	0.314	0.085	0.052	16	0.150	0.184	-0.044	0.069
2	0.186	0.233	-0.019	0.071	17	0.146	0.178	-0.039	0.064
3	0.174	0.196	-0.060	0.027	18	0.147	0.178	-0.044	0.071
4	0.200	0.196	-0.063	0.013	19	0.144	0.173	-0.049	0.075
5	0.191	0.205	-0.063	0.050	20	0.144	0.170	-0.045	0.081
6	0.191	0.201	-0.054	0.026	21	0.139	0.169	-0.040	0.073
7	0.192	0.203	-0.037	0.026	22	0.138	0.165	-0.031	0.070
8	0.186	0.195	-0.058	0.019	23	0.144	0.164	-0.037	0.060
9	0.176	0.192	-0.036	0.045	24	0.144	0.163	-0.043	0.058
10	0.165	0.192	-0.031	0.059	25	0.143	0.159	-0.044	0.057
11	0.159	0.194	-0.041	0.066	26	0.142	0.158	-0.044	0.060
12	0.157	0.195	-0.039	0.072	27	0.145	0.156	-0.040	0.058
13	0.160	0.191	-0.044	0.063	28	0.146	0.156	-0.038	0.059
14	0.157	0.189	-0.034	0.072	29	0.142	0.153	-0.037	0.057
15	0.151	0.185	-0.039	0.076	30	0.141	0.150	-0.032	0.052

Table 6.18: Annualized moments for portfolios of DJ30 stocks.

The row with the number n in the leading column shows summary statistics of the monthly returns on the equally-weighted portfolio of the first n stocks in the Dow Jones index according to the list in Table 6.17. The monthly rates of return from January 2014 to December 2023 were used. The moments are annualized as explained in the text.

- (f) Based on your answers to the two preceding questions, what is the fair stock price per share?
Check that you get the same price as in question (c)!
- (g) Estimate the present value of the firm's growth opportunities.

Exercise 6.2. You are trying to value the stocks of Imaginary Inc. The company is currently involved in a very risky, but potentially very profitable project. In the preceding year, the company had earnings of \$10 per share. You expect the earnings per share to grow to \$15, \$20, and \$25 in the next three years, after which you expect a growth rate of 2%. The company always applies a plowback ratio of 60%. You estimate that a risk-adjusted discount rate of 10% is appropriate.

- (a) What are the expected dividends per share in the next three years?
- (b) What is the expected price per share three years from now?
- (c) What is the fair stock price per share today?
- (d) What is the expected annual rate of return on the stock?

Exercise 6.3. Illuminati Inc. has just paid a dividend of \$2 per share. The company's annual dividends are expected to grow by 5% each year indefinitely. The appropriate risk-adjusted discount rate is 13%.

- (a) What is the fair stock price per share today?
- (b) Compute the stock price if the growth rate immediately would change to 2%, 4%, 6%, 8%, 10%, or 15%. Discuss the sensitivity of the stock price to the growth rate.
- (c) Returning to a growth rate of 5%, compute the stock price if the risk-adjusted discount rate immediately would change to 8%, 10%, 12%, 14%, 16%, or 18%. Discuss the sensitivity of the stock price to the risk-adjusted discount rate.

Exercise 6.4. Go to the homepage of Professor Kenneth French and download time series of monthly returns on 10 industry portfolios. For each portfolio, calculate the average, the standard deviation, the skewness, and the kurtosis of the time series of returns. Discuss your findings.

Exercise 6.5. The price-earnings ratio of a stock is often seen as an indicator for the expected growth rate of the company. In the infinite-horizon analysis of Section 6.2.1, the price-earnings ratio is the constant

$$P/e = \frac{(1-b)(1+g)}{r-g},$$

where b is the plowback ratio, g is the annual growth rate in earnings, and r is the appropriate discount rate.

- (a) Show that the above equation implies that

$$g = \frac{r \times P/e - (1 - b)}{1 - b + P/e},$$

which can be interpreted as the growth rate required to justify a given price-earnings ratio.

- (b) Assume $b = 0.4$ and $r = 0.08$. Calculate the required growth rate for price-earnings ratios of $1, 2, \dots, 30$, and illustrate graphically how the required growth rate varies with the price-earnings ratio.
(c) In order to investigate the sensitivity of the findings in the previous question to the values of b and r , answer the same question in each of the following cases:
- $b = 0.2, r = 0.08$
 - $b = 0.6, r = 0.08$
 - $b = 0.4, r = 0.04$
 - $b = 0.4, r = 0.12$

Discuss your results.

- (d) Some investors pay attention to the so-called PEG ratio, which is the price/earnings ratio divided by the percentage expected annual growth rate of earnings. For example, with a price/earnings ratio of 20 and an expected growth rate of 10% in earnings, the PEG is 2. According to an old rule-of-thumb, a stock is overvalued if the PEG exceeds 1. Based on your findings in the previous questions, discuss this rule-of-thumb.

CHAPTER 7

One-period portfolio choice

Investors can invest in a large number of assets. How can an investor optimally form a portfolio of many assets? To answer this question, we first need to formulate the objective of the investor. We expect that investors are greedy in the sense that, other things equal, they prefer a large expected return, but also that they are risk-averse and prefer low risk. This chapter presents the *mean-variance analysis* introduced by Markowitz (1952, 1959) as a tool for determining optimal portfolios over a given period of time when the investor cares about the mean, i.e. the expected return, and the variance which is a measure of the return risk. The mean-variance analysis is the most well-known and most frequently applied framework for optimal portfolio choice.

First, Section 7.1 considers portfolios of only risky assets. Here we determine a mean-variance efficient frontier of risky assets. The frontier represents portfolios that has the lowest variance among all portfolios with the same mean return. In particular, we identify two specific frontier portfolios, the minimum-variance portfolio and the maximum-slope portfolio, from which the entire frontier is easily generated. Next, Section 7.2 adds a riskfree asset to the analysis and concludes that the only mean-variance efficient portfolio are then combinations of the riskfree asset and a specific portfolio of risky assets, namely the so-called tangency portfolio. Section 7.3 moves on to find the optimal combination of the riskfree asset and the tangency portfolio for a given investor, which depends on how risk averse the investor is. Finally, Sections 7.4 and 7.5 provide a thorough discussion of potential practical and theoretical issues with the mean-variance framework.

7.1 Mean-variance analysis with only risky assets

The main assumption of mean-variance analysis is that, when the investor is choosing among different portfolios, she considers only the expectation and the variance of the return over a fixed, future period of time. Furthermore, she prefers as high an expected return as possible and as low a return variance as possible. An investor behaving this way is said to have *mean-variance preferences* or to be a *mean-variance optimizer*.

In this section we assume that the investor can invest in N risky assets and no riskfree asset. We use the same notation as introduced in Section 4.2: μ is the vector of expected rates of return and $\underline{\Sigma} = (\Sigma_{ij})$ is the variance-covariance matrix of the rates of return. Recall that π_i denotes the fraction of the total portfolio value which is invested in asset i

and, hence, a portfolio vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_N)^\top$ must satisfy

$$\boldsymbol{\pi} \cdot \mathbf{1} = \pi_1 + \pi_2 + \cdots + \pi_N = 1, \quad (7.1)$$

where $\mathbf{1}$ is a vector of ones. The expectation, variance, and standard deviation of the portfolio return are given by

$$\mu(\boldsymbol{\pi}) = \boldsymbol{\pi} \cdot \boldsymbol{\mu} = \sum_{i=1}^N \pi_i \mu_i, \quad (7.2)$$

$$\sigma^2(\boldsymbol{\pi}) = \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} = \sum_{i=1}^N \sum_{j=1}^N \pi_i \pi_j \Sigma_{ij}, \quad (7.3)$$

$$\sigma(\boldsymbol{\pi}) = \sqrt{\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}} = \left(\sum_{i=1}^N \sum_{j=1}^N \pi_i \pi_j \Sigma_{ij} \right)^{1/2}. \quad (7.4)$$

We assume that the variance-covariance matrix $\underline{\Sigma}$ is non-singular, which is the case if none of the assets are redundant, i.e., if it is impossible to replicate one asset by a portfolio of the other assets. Equivalently, it is impossible to form a riskfree portfolio from these risky assets. The inverse of $\underline{\Sigma}$ is denoted by $\underline{\Sigma}^{-1}$. By construction $\underline{\Sigma}$ is symmetric, and then $\underline{\Sigma}^{-1}$ is also symmetric.

7.1.1 Mean-variance efficient portfolios

A portfolio is said to be **mean-variance efficient** if it has the minimum return variance among all the portfolios with the same mean return. Any mean-variance optimizer will choose some mean-variance efficient portfolio. Assuming that there are no further portfolio constraints, we can find a mean-variance efficient portfolio with expected return $\bar{\mu}$ by solving the quadratic minimization problem

$$\begin{aligned} & \min_{\boldsymbol{\pi}} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \\ & \text{s.t. } \boldsymbol{\pi} \cdot \boldsymbol{\mu} = \bar{\mu}, \\ & \quad \boldsymbol{\pi} \cdot \mathbf{1} = 1. \end{aligned} \quad (7.5)$$

The solution is stated in the theorem below. While [Markowitz \(1952, 1959\)](#) presented an intuitive understanding and a graphical illustration of this theorem as well as some of the following theorems, the formal mathematical proofs of the theorems for the case with many risky assets are due to [Merton \(1972\)](#).

Before stating the theorem, we introduce four auxiliary constants

$$A = \boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu}, \quad B = \boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \mathbf{1} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu}, \quad C = \mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1}, \quad D = AC - B^2, \quad (7.6)$$

where the equality $\boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \mathbf{1} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu}$ follows from (4.38). It can be shown that A , C , and D are positive, whereas B can be positive or negative.¹ We do require, however, that B is different from zero.

¹Here is an explanation. A symmetric $N \times N$ matrix $\underline{\Sigma}$ is said to be *positive definite* if $\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} > 0$ for any non-zero N -vector $\boldsymbol{\pi}$. Since in our case $\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi}$ equals the variance of the portfolio $\boldsymbol{\pi}$ and all portfolios of risky assets have a return with positive variance, the variance-covariance matrix $\underline{\Sigma}$ is indeed a positive definite matrix. A result in linear algebra says that the inverse $\underline{\Sigma}^{-1}$ is then also positive definite, i.e.,

Theorem 7.1

The mean-variance efficient portfolio with expected rate of return $\bar{\mu}$ is given by the portfolio weight vector

$$\boldsymbol{\pi}(\bar{\mu}) = \frac{C\bar{\mu} - B}{D} \underline{\Sigma}^{-1} \boldsymbol{\mu} + \frac{A - B\bar{\mu}}{D} \underline{\Sigma}^{-1} \mathbf{1}. \quad (7.7)$$

The variance of the return on this portfolio is equal to

$$\sigma^2(\bar{\mu}) = \boldsymbol{\pi}(\bar{\mu}) \cdot \underline{\Sigma} \boldsymbol{\pi}(\bar{\mu}) = \frac{C\bar{\mu}^2 - 2B\bar{\mu} + A}{D}. \quad (7.8)$$

so the standard deviation is

$$\sigma(\bar{\mu}) = \sqrt{\frac{C\bar{\mu}^2 - 2B\bar{\mu} + A}{D}}. \quad (7.9)$$

Proof

We solve the problem (7.5) by the Lagrange optimization technique.² Letting α and β denote the Lagrange multipliers of the two constraints, the Lagrangian is

$$L(\boldsymbol{\pi}) = \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} + \alpha (\bar{\mu} - \boldsymbol{\pi} \cdot \boldsymbol{\mu}) + \beta (1 - \boldsymbol{\pi} \cdot \mathbf{1}). \quad (7.10)$$

We need to differentiate the Lagrangian with respect to the choice variables which in our case are the portfolio weights, i.e., the portfolio vector $\boldsymbol{\pi}$. Here we apply the differentiation rules (4.40) and (4.41) which imply that

$$\frac{\partial(\boldsymbol{\pi} \cdot \boldsymbol{\mu})}{\partial \boldsymbol{\pi}} = \boldsymbol{\mu}, \quad \frac{\partial(\boldsymbol{\pi} \cdot \mathbf{1})}{\partial \boldsymbol{\pi}} = \mathbf{1}, \quad \frac{\partial(\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi})}{\partial \boldsymbol{\pi}} = 2\underline{\Sigma} \boldsymbol{\pi}.$$

In total, the first-order condition for maximizing L with respect to $\boldsymbol{\pi}$ is

$$\frac{\partial L}{\partial \boldsymbol{\pi}} = 2\underline{\Sigma} \boldsymbol{\pi} - \alpha \boldsymbol{\mu} - \beta \mathbf{1} = 0, \quad (7.11)$$

which implies that

$$\boldsymbol{\pi} = \frac{1}{2} \alpha \underline{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{2} \beta \underline{\Sigma}^{-1} \mathbf{1}. \quad (7.12)$$

Applying the rules for dot products in (4.34), we get

$$\begin{aligned} \boldsymbol{\pi} \cdot \boldsymbol{\mu} &= \boldsymbol{\mu} \cdot \boldsymbol{\pi} = \boldsymbol{\mu} \cdot \frac{1}{2} \alpha \underline{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu} \cdot \frac{1}{2} \beta \underline{\Sigma}^{-1} \mathbf{1} \\ &= \frac{1}{2} \alpha \boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{2} \beta \boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{2} \alpha A + \frac{1}{2} \beta B, \end{aligned}$$

$\mathbf{x} \cdot \underline{\Sigma}^{-1} \mathbf{x} > 0$ for any non-zero N -vector \mathbf{x} . This implies that $A > 0$ and $C > 0$. Also

$$AD = A(AC - B^2) = (B\boldsymbol{\mu} - A\mathbf{1}) \cdot \underline{\Sigma}^{-1} (B\boldsymbol{\mu} - A\mathbf{1}) > 0$$

and since $A > 0$ we must have $D > 0$.

$$\begin{aligned}\pi \cdot \mathbf{1} &= \mathbf{1} \cdot \pi = \mathbf{1} \cdot \frac{1}{2} \alpha \Sigma^{-1} \boldsymbol{\mu} + \mathbf{1} \cdot \frac{1}{2} \beta \Sigma^{-1} \mathbf{1} \\ &= \frac{1}{2} \alpha \mathbf{1} \cdot \Sigma^{-1} \boldsymbol{\mu} + \frac{1}{2} \beta \mathbf{1} \cdot \Sigma^{-1} \mathbf{1} = \frac{1}{2} \alpha B + \frac{1}{2} \beta C.\end{aligned}$$

Hence, we can write the two constraints to the optimization problem as

$$\frac{1}{2} A\alpha + \frac{1}{2} B\beta = \bar{\mu}, \quad \frac{1}{2} B\alpha + \frac{1}{2} C\beta = 1,$$

which have the solution

$$\alpha = 2 \frac{C\bar{\mu} - B}{D}, \quad \beta = 2 \frac{A - B\bar{\mu}}{D}. \quad (7.13)$$

Substituting this into (7.12) we obtain the expression (7.7) for π . The variance expression (7.8) follows from tedious computations—Exercise 7.3 asks for a proof.

If you invest in three or more assets, many portfolios share the same expected rate of return. Among all portfolios with an expected return of $\bar{\mu}$, the portfolio in (7.8) is the one with the lowest variance.

Example 7.1

If you have three assets, the portfolio weight of asset 3 is given in terms of the weights of the other assets as $\pi_3 = 1 - \pi_1 - \pi_2$. The expected rate of return on the portfolio is

$$\begin{aligned}\pi_1 \mu_1 + \pi_2 \mu_2 + \pi_3 \mu_3 &= \pi_1 \mu_1 + \pi_2 \mu_2 + (1 - \pi_1 - \pi_2) \mu_3 \\ &= \mu_3 + \pi_1(\mu_1 - \mu_3) + \pi_2(\mu_2 - \mu_3).\end{aligned}$$

²Suppose you want to maximize a function $f(\mathbf{x})$ of a vector variable \mathbf{x} among all the vectors that satisfy a given equality constraint $g(\mathbf{x}) = b$, where b is a number. In mathematical notation the problem is

$$\max f(\mathbf{x}) \quad \text{s.t. } g(\mathbf{x}) = b,$$

where “s.t.” is short for “such that” or “subject to.” Such a constrained maximization problem can be solved by Lagrange’s method. First, set up the Lagrangian function

$$L(\mathbf{x}) = f(\mathbf{x}) + \lambda(b - g(\mathbf{x})),$$

where the constant λ is the so-called Lagrange multiplier. Then solve the equations

$$\frac{\partial L(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}, \quad g(\mathbf{x}) = b$$

for \mathbf{x} . If \mathbf{x} solves these equations and $L(\mathbf{x})$ is concave, then \mathbf{x} solves the constrained maximization problem. For a constrained minimization problem

$$\min f(\mathbf{x}) \quad \text{s.t. } g(\mathbf{x}) = b,$$

the procedure is similar. Solve the same equations as for the maximization problem above. If \mathbf{x} solves these equations and $L(\mathbf{x})$ is convex, then \mathbf{x} solves the constrained minimization problem. More than one constraint can be handled by having a Lagrange multiplier and a corresponding term in the Lagrangian function for each constraint.

Any choice of π_1 and π_2 that satisfies

$$\mu_3 + \pi_1(\mu_1 - \mu_3) + \pi_2(\mu_2 - \mu_3) = \bar{\mu}$$

or, equivalently,

$$\pi_1(\mu_1 - \mu_3) + \pi_2(\mu_2 - \mu_3) = \bar{\mu} - \mu_3$$

gives a portfolio with expected return $\bar{\mu}$. As long as $\mu_1 \neq \mu_2$, there are infinitely many such portfolios.

Equation (7.8) shows that the efficient combinations of variance and mean will form a parabola in a (mean, variance)-diagram. Traditionally the portfolios are depicted in a (standard deviation, mean)-diagram. Equation (7.8) can also be written as

$$\frac{\sigma^2(\bar{\mu})}{1/C} - \frac{(\bar{\mu} - B/C)^2}{D/C^2} = 1, \quad (7.14)$$

from which it follows that the optimal combinations of standard deviation and mean form a hyperbola in the (standard deviation, mean)-diagram. This hyperbola is called the **mean-variance frontier** or the **efficient frontier** of risky assets. Note that some authors reserve the term efficient frontier for only the upward-sloping part of the hyperbola since no rational investors would choose a portfolio corresponding to a point on the downward-sloping part of the hyperbola. Figure 7.1 below shows an example of a typical mean-variance frontier. The mean-variance efficient portfolios are sometimes called *frontier portfolios*.

7.1.2 The minimum-variance portfolio

The **minimum-variance portfolio** is the portfolio that has the minimum variance among all portfolios. This portfolio is sometimes called the *global* minimum-variance portfolio to emphasize that it is not just the portfolio with the minimum variance for a given expected return, but the portfolio with the minimum variance among all portfolios. In mathematical terms, the minimum-variance portfolio is the solution to the constrained minimization problem

$$\begin{aligned} & \min_{\boldsymbol{\pi}} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \\ & \text{s.t. } \boldsymbol{\pi} \cdot \mathbf{1} = 1, \end{aligned} \quad (7.15)$$

where there is no constraint on the expected portfolio return. Note that the minimum-variance portfolio obviously is on the efficient frontier. The next theorem characterizes the minimum-variance portfolio.

Theorem 7.2

The (global) minimum-variance portfolio is given by

$$\boldsymbol{\pi}_{\min} = \frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{\mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1}} \underline{\Sigma}^{-1} \mathbf{1} \quad (7.16)$$

and has an expected rate of return of

$$\mu_{\min} = \frac{B}{C}, \quad (7.17)$$

a return variance of

$$\sigma_{\min}^2 = \sigma^2(\bar{\mu}_{\min}) = \frac{1}{C}, \quad (7.18)$$

and a standard deviation of

$$\sigma_{\min} = \frac{1}{\sqrt{C}}. \quad (7.19)$$

If r_{\min} denotes the return on the (global) minimum-variance portfolio and r denotes the return on any risky asset or portfolio of risky assets, efficient or not, then

$$\text{Cov}[r, r_{\min}] = \text{Var}[r_{\min}]. \quad (7.20)$$

Proof

We can solve the constrained optimization problem (7.15) with the Lagrange technique used in the proof of Theorem 7.1. The Lagrangian is now

$$L(\boldsymbol{\pi}) = \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} + \delta(1 - \boldsymbol{\pi} \cdot \mathbf{1}),$$

where δ is the Lagrange multiplier associated with the constraint. The first-order condition is

$$\frac{\partial L}{\partial \boldsymbol{\pi}} = 2\underline{\Sigma} \boldsymbol{\pi} - \delta \mathbf{1} = 0 \quad \Rightarrow \quad \boldsymbol{\pi} = \frac{\delta}{2} \underline{\Sigma}^{-1} \mathbf{1}.$$

By substituting this into the constraint, we obtain

$$1 = \boldsymbol{\pi} \cdot \mathbf{1} = \mathbf{1} \cdot \boldsymbol{\pi} = \mathbf{1} \cdot \left(\frac{\delta}{2} \underline{\Sigma}^{-1} \mathbf{1} \right) = \frac{\delta}{2} \mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \frac{\delta}{2} C \quad \Rightarrow \quad \frac{\delta}{2} = \frac{1}{C}.$$

Hence, the minimum-variance portfolio is $\boldsymbol{\pi} = \frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1}$, confirming (7.16). The expected rate of return on this portfolio is

$$\boldsymbol{\mu} \cdot \boldsymbol{\pi} = \boldsymbol{\mu} \cdot \left(\frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1} \right) = \frac{1}{C} \boldsymbol{\mu} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \frac{B}{C},$$

and the variance is

$$\boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} = \frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1} \cdot \underline{\Sigma} \left(\frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1} \right) = \frac{1}{C^2} \underline{\Sigma}^{-1} \mathbf{1} \cdot \mathbf{1} = \frac{1}{C^2} \mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{C^2} C = \frac{1}{C},$$

which shows (7.17) and (7.18), while (7.19) follows immediately from (7.18).

Here is an alternative proof. Since the minimum-variance portfolio is on the efficient frontier, the variance and the expected return are related as in (7.8). We can therefore identify the minimum-variance portfolio by minimizing $\sigma^2(\bar{\mu})$ in (7.8) over $\bar{\mu}$ and then substituting that value into Eq. (7.7). We want to minimize the function

$$f(\bar{\mu}) = \frac{C\bar{\mu}^2 - 2B\bar{\mu} + A}{D}.$$

It is a quadratic function and since C is positive, the minimum is obtained when $f'(\bar{\mu}) = 0$. Differentiating we get

$$f'(\bar{\mu}) = \frac{2C\bar{\mu} - 2B}{D}$$

and then $f'(\bar{\mu}) = 0$ implies $2C\bar{\mu} = 2B$ and thus $\bar{\mu} = \frac{B}{C}$. The minimum variance is

$$\begin{aligned} f\left(\frac{B}{C}\right) &= \frac{C\left(\frac{B}{C}\right)^2 - 2B\frac{B}{C} + A}{D} = \frac{\frac{B^2}{C} - 2\frac{B^2}{C} + A}{D} = \frac{A - \frac{B^2}{C}}{D} \\ &= \frac{AC - B^2}{CD} = \frac{D}{CD} = \frac{1}{C}. \end{aligned}$$

Now we substitute the expression for $\bar{\mu}$ into (7.7). First look at the fractions:

$$\begin{aligned} \frac{C\bar{\mu} - B}{D} &= \frac{C\frac{B}{C} - B}{D} = \frac{B - B}{D} = 0, \\ \frac{A - B\bar{\mu}}{D} &= \frac{A - B\frac{B}{C}}{D} = \frac{AC - B^2}{CD} = \frac{D}{CD} = \frac{1}{C}. \end{aligned}$$

Therefore, the minimum-variance portfolio is

$$\boldsymbol{\pi} = 0 \times \underline{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{C} \underline{\Sigma}^{-1} \mathbf{1}.$$

The fact that the minimum-variance portfolio has the property (7.20) is to be shown in Exercise 7.4.

Note that $\underline{\Sigma}^{-1} \mathbf{1}$ in (7.16) is a vector of dimension N . The division by $C = \mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1}$ just scales the portfolio weights so that they sum up to one as required.

We expect that assets with low standard deviation have relatively large weights in the minimum-variance portfolio. However, the correlation structure of the assets is also important. For example, the minimum-variance portfolio might have a substantial weight on an asset with a relatively large standard deviation if that asset has low correlation with some low-variance assets and therefore is useful for diversifying away risk.

Let us briefly reconsider the case with only two risky assets. Here we check that if we apply the results derived in this section to the two-asset case, we obtain the exact same results that we already found in Section 4.1. With two assets the variance-covariance matrix and its inverse are

$$\underline{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad \underline{\Sigma}^{-1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}, \quad (7.21)$$

where σ_1 and σ_2 are the standard deviations of the returns of the two assets and ρ is the correlation between them. Now

$$\underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 - \rho\sigma_1\sigma_2 \\ \sigma_1^2 - \rho\sigma_1\sigma_2 \end{pmatrix} \quad (7.22)$$

and therefore

$$C = \mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) \quad (7.23)$$

so that the minimum-variance portfolio is

$$\boldsymbol{\pi}_{\min} = \begin{pmatrix} \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \\ \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \end{pmatrix}, \quad (7.24)$$

which confirms (4.6). The variance of the minimum-variance portfolio is given by

$$\sigma^2(\bar{\mu}_{\min}) = \frac{1}{C} = \frac{(1 - \rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \quad (7.25)$$

which is identical to (4.7).

7.1.3 The maximum-slope portfolio

Any portfolio of the risky assets corresponds to a point (σ, μ) in a diagram with standard deviation along the horizontal axis and expected return along the vertical axis, as we have it in Figure 7.1 below for example. By definition of the mean-variance frontier any such point is on the frontier or to the right of the frontier. We can connect any such point with the origin $(0, 0)$ by a straight line. The slope of that line will be μ/σ . It turns out to be interesting to find the portfolio for which this slope is maximized. This portfolio is naturally called the **maximum-slope portfolio**. The following theorem provides an expression for this portfolio and the expectation and variance of its rate of return.

Theorem 7.3

Assume $B \neq 0$. The maximum-slope portfolio is given by

$$\boldsymbol{\pi}_{\text{slope}} = \frac{1}{B} \underline{\Sigma}^{-1} \boldsymbol{\mu} = \frac{1}{\mathbf{1} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu}} \underline{\Sigma}^{-1} \boldsymbol{\mu} \quad (7.26)$$

and has expected return, variance, and standard deviation given by

$$\mu_{\text{slope}} = \frac{A}{B}, \quad (7.27)$$

$$\sigma_{\text{slope}}^2 = \frac{A}{B^2}, \quad (7.28)$$

$$\sigma_{\text{slope}} = \frac{\sqrt{A}}{|B|}. \quad (7.29)$$

Proof

We want to maximize the ratio μ/σ over all portfolios. We could again apply the Lagrange technique, but let us exploit that the maximum-slope portfolio corresponds to a point on the efficient frontier so that its variance and expected return must satisfy the relation (7.8). Hence, we maximize the function

$$f(\bar{\mu}) = \frac{\bar{\mu}}{\sqrt{\frac{C\bar{\mu}^2 - 2B\bar{\mu} + A}{D}}} = \frac{\bar{\mu}\sqrt{D}}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}.$$

By using the rules for differentiating a quotient and the square root, we obtain

$$\begin{aligned} f'(\bar{\mu}) &= \frac{\sqrt{D}\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} - \bar{\mu}\sqrt{D}\frac{2C\bar{\mu} - 2B}{2\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}}{C\bar{\mu}^2 - 2B\bar{\mu} + A} \\ &= \sqrt{D}\frac{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} - \bar{\mu}\frac{C\bar{\mu} - B}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}}{C\bar{\mu}^2 - 2B\bar{\mu} + A}. \end{aligned}$$

Now we see that $f'(\bar{\mu}) = 0$ when

$$\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} = \bar{\mu}\frac{C\bar{\mu} - B}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}},$$

which is equivalent to

$$C\bar{\mu}^2 - 2B\bar{\mu} + A = \bar{\mu}(C\bar{\mu} - B) \quad \Leftrightarrow \quad B\bar{\mu} = A.$$

Hence, the maximum-slope portfolio has an expected return of $\mu_{\text{slope}} = A/B$. We obtain its variance by substituting this expected return into (7.8):

$$\begin{aligned} \sigma_{\text{slope}}^2 &= \frac{C\left(\frac{A}{B}\right)^2 - 2B\frac{A}{B} + A}{D} = \frac{\frac{CA^2}{B^2} - A}{D} \\ &= \frac{CA^2 - AB^2}{DB^2} = \frac{A(AC - B^2)}{DB^2} = \frac{AD}{DB^2} = \frac{A}{B^2}. \end{aligned}$$

The portfolio weights are obtained by substituting the expected return of A/B into (7.7). First consider the fractions:

$$\frac{C\bar{\mu} - B}{D} = \frac{C\frac{A}{B} - B}{D} = \frac{CA - B^2}{BD} = \frac{D}{BD} = \frac{1}{B}, \quad \frac{A - B\bar{\mu}}{D} = \frac{A - B\frac{A}{B}}{D} = 0.$$

Therefore, the maximum-slope portfolio is

$$\boldsymbol{\pi}_{\text{slope}} = \frac{1}{B} \underline{\Sigma}^{-1} \boldsymbol{\mu} + 0 \times \underline{\Sigma}^{-1} \mathbf{1} = \frac{1}{B} \underline{\Sigma}^{-1} \boldsymbol{\mu}$$

as stated in the theorem.

It seems natural to expect that the maximum-slope portfolio corresponds to a point on the upward-sloping branch of the curved frontier, but this is in fact only the case if the expected return on the minimum-variance portfolio is positive, which is true whenever $B > 0$. However, we cannot rule out that B is negative by any valid mathematical or economical arguments. In the case $B < 0$, the maximum-slope portfolio is located on the downward-sloping branch of the curved frontier and is really the portfolio giving the most negative slope of all lines between $(0,0)$ and a point on or to the right of the frontier.

The case in which $B = 0$ is weird. While there is an upper bound on the slope you can obtain, there is no portfolio attaining that maximum. However, B will be non-zero unless the values of the expected returns, variances, and covariances are carefully selected

to obtain $B = 0$ so, in practice, the case $B = 0$ is uninteresting.

7.1.4 Properties of the efficient frontier

Theorem 7.1 gives us a basic way of generating the mean-variance frontier of risky assets: Consider a range of values of $\bar{\mu}$ and then compute the corresponding standard deviation from Eq. (7.9). Obviously you need to compute the auxiliary constants A, B, C, D . The portfolio generating each point is determined from (7.7).

However, it is clear from (7.16) and (7.26) that $\underline{\Sigma}^{-1}\mathbf{1} = C\boldsymbol{\pi}_{\min}$ and $\underline{\Sigma}^{-1}\boldsymbol{\mu} = B\boldsymbol{\pi}_{\text{slope}}$ so (7.7) implies that any frontier portfolio is a combination of the maximum-slope portfolio and the minimum-variance portfolio. This is a **two-fund separation** result: if the investors can only form portfolios of the N risky assets, any mean-variance optimizing investor chooses a combination of two special portfolios or funds, namely the minimum-variance portfolio and the maximum-slope portfolio. These two portfolios are said to generate the mean-variance frontier of risky assets.

In fact, the next theorem shows that *any* two frontier portfolios generate the entire frontier, i.e. investors are satisfied if they can trade in any two frontier portfolios.

Theorem 7.4

Assume $B \neq 0$. Suppose $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are any two different frontier portfolios. A portfolio $\boldsymbol{\pi}$ is then a frontier portfolio if and only if a number w exists so that $\boldsymbol{\pi} = w\boldsymbol{\pi}_1 + (1 - w)\boldsymbol{\pi}_2$. Hence the entire efficient frontier of risky assets is generated by all such combinations of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$.

Proof

A portfolio $\boldsymbol{\pi}$ is a frontier portfolio if it is of the form (7.7) for some expected return $\bar{\mu}$. If we combine this with (7.16) and (7.26) we have

$$\boldsymbol{\pi} = \frac{(C\bar{\mu} - B)B}{D}\boldsymbol{\pi}_{\text{slope}} + \frac{(A - B\bar{\mu})C}{D}\boldsymbol{\pi}_{\min},$$

where the two multipliers of the portfolios sum to one,

$$\frac{(C\bar{\mu} - B)B}{D} + \frac{(A - B\bar{\mu})C}{D} = \frac{-B^2 + AC}{D} = 1,$$

so that $\boldsymbol{\pi}$ indeed is a weighted average of the maximum-slope portfolio and the minimum-variance portfolio. Varying $\bar{\mu}$ between $-\infty$ and ∞ , we vary the weights of the two portfolios between $-\infty$ and ∞ , so the frontier consists exactly of all such combinations of the maximum-slope and the minimum-variance portfolio.

Since $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are frontier portfolio, weights w_1 and w_2 exist so that

$$\boldsymbol{\pi}_1 = w_1\boldsymbol{\pi}_{\text{slope}} + (1 - w_1)\boldsymbol{\pi}_{\min}, \quad \boldsymbol{\pi}_2 = w_2\boldsymbol{\pi}_{\text{slope}} + (1 - w_2)\boldsymbol{\pi}_{\min}.$$

Now consider any combination of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, say with a weight of w on $\boldsymbol{\pi}_1$ and thus a

weight of $1 - w$ on $\boldsymbol{\pi}_2$. We can write this portfolio as

$$\begin{aligned}\boldsymbol{\pi} &= w\boldsymbol{\pi}_1 + (1 - w)\boldsymbol{\pi}_2 \\ &= w(w_1\boldsymbol{\pi}_{\text{slope}} + (1 - w_1)\boldsymbol{\pi}_{\text{min}}) + (1 - w)(w_2\boldsymbol{\pi}_{\text{slope}} + (1 - w_2)\boldsymbol{\pi}_{\text{min}}) \\ &= (ww_1 + (1 - w)w_2)\boldsymbol{\pi}_{\text{slope}} + (w(1 - w_1) + (1 - w)(1 - w_2))\boldsymbol{\pi}_{\text{min}}.\end{aligned}$$

Since $\boldsymbol{\pi}$ is a combination of $\boldsymbol{\pi}_{\text{slope}}$ and $\boldsymbol{\pi}_{\text{min}}$, it is indeed mean-variance efficient.

The frontier is often generated from the minimum-variance portfolio and the maximum-slope portfolio as we have nice formulas and interpretations for these portfolios. If we let w denote the weight of the minimum-variance portfolio and $1 - w$ the weight of the maximum-slope portfolio, the expected return on the combined portfolio is simply

$$\mu(w) = w\bar{\mu}_{\text{min}} + (1 - w)\bar{\mu}_{\text{slope}}. \quad (7.30)$$

To compute the variance of the combined portfolio, we need the covariance between the minimum-variance and the maximum-slope portfolios, but according to (7.20) this equals the variance of the minimum-variance portfolio. Hence the variance is

$$\begin{aligned}\sigma^2(w) &= w^2\sigma_{\text{min}}^2 + (1 - w)^2\sigma_{\text{slope}}^2 + 2w(1 - w)\sigma_{\text{min}}^2 \\ &= w(2 - w)\sigma_{\text{min}}^2 + (1 - w)^2\sigma_{\text{slope}}^2,\end{aligned} \quad (7.31)$$

and the standard deviation follows by taking the square root. This gives us one point $(\sigma(w), \mu(w))$ on the frontier of risky assets. By repeating this for a wide range of values for w , we get a good picture of the frontier. You need to include values of w above 1 to get frontier points below the minimum-variance portfolio and values of w below 0 to get points above the maximum-slope portfolio.

In the next section we introduce the tangency portfolio which is another frontier portfolio with a simple formula and interpretation. We can then generate the frontier of risky assets by combining, e.g., the minimum-variance portfolio and the tangency portfolio.

Example 7.2

Let us consider an example with ten risky assets. The expectation, standard deviation, and variance of their returns are shown in the left part of Table 7.1 (the column with the header ‘‘Tangency portfolio’’ is explained below). Table 7.2 shows the variance-covariance matrix and Table 7.3 the corresponding correlation matrix.

Given these inputs, the minimum-variance portfolio and the maximum-slope portfolio can be computed, and they are shown in the right part of Table 7.1. Do the portfolio weights make intuitive sense? The minimum-variance portfolio has by far the largest weights on assets 3 and 6, which is natural as these assets have the smallest standard deviations. The correlations also matter: despite having the third-smallest standard deviation, asset 10 has only a tiny weight, probably because of its fairly large correlation with asset 6. Note that the minimum-variance portfolio does not give large negative weights to assets with large standard deviations, as that will also contribute to a large portfolio variance due to the terms $\pi_i^2\sigma_i^2$. The maximum-slope portfolio puts by far the largest weight on

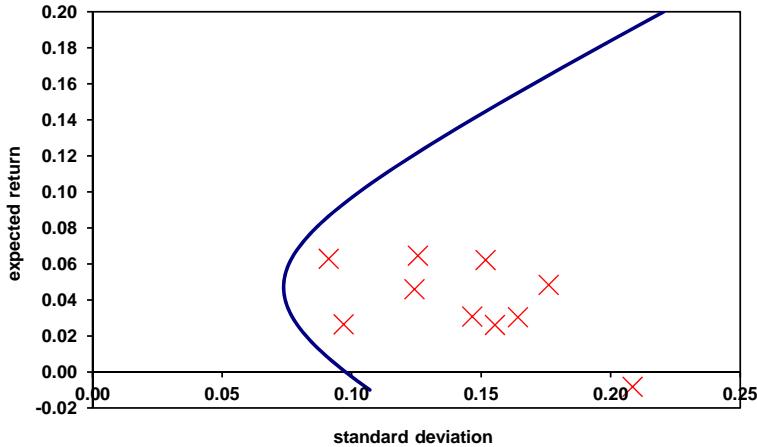


Figure 7.1: An example of the mean-variance frontier.

The figure refers to Example 7.2. The curve shows the mean-variance frontier generated from the 10 individual assets corresponding to the x's.

asset 3 due to its large expected return and low risk. Compared to the minimum-variance weights, the maximum-slope weights are much more dispersed and also include substantial negative values in order to optimize the risk-return tradeoff.

Figure 7.1 shows the mean-variance frontier generated from these ten assets. Each individual asset corresponds to an “x”. As you can see from the graph, by combining the risky assets we can reduce the standard deviation substantially without reducing the expected return. This is because of the diversification of risk you obtain by investing in different assets with returns that are not perfectly correlated with each other.

The following theorem establishes another property of the mean-variance portfolios. Exercise 7.6 asks for a proof hereof.

Theorem 7.5

For any mean-variance efficient portfolio π different from the minimum-variance portfolio, another mean-variance efficient portfolio $\tilde{\pi}$ exists with the property that the covariance between the returns on the two portfolios is zero, i.e.,

$$\text{Cov}[r(\pi), r(\tilde{\pi})] = 0. \quad (7.32)$$

The expected rate of return on the portfolio $\tilde{\pi}$ is

$$E[r(\tilde{\pi})] = \frac{A - B E[r(\pi)]}{B - C E[r(\pi)]}, \quad (7.33)$$

where A , B , and C are the constants defined in Eq. (7.6). In the (σ, μ) -diagram the tangent to the mean-variance frontier at the point corresponding to π intersects the vertical axis exactly in $E[r(\tilde{\pi})]$.

Asset	Return moment			Percentage weights		
	expect	std dev	variance	min-var pf	max-slope pf	tangency pf
1	0.0307	0.1465	0.0215	-2.83	-1.20	-0.44
2	0.0304	0.1642	0.0270	0.04	5.67	8.30
3	0.0629	0.0912	0.0083	32.63	126.23	169.97
4	0.0483	0.1761	0.0310	-3.76	-11.75	-15.48
5	0.0645	0.1256	0.0158	18.52	12.76	10.06
6	0.0264	0.0969	0.0094	38.56	-31.31	-63.95
7	-0.0082	0.2085	0.0435	4.30	-32.11	-49.12
8	0.0261	0.1554	0.0241	-0.51	-45.13	-65.98
9	0.0621	0.1518	0.0230	10.10	52.29	72.01
10	0.0460	0.1244	0.0155	2.96	24.54	34.62

Table 7.1: Example with ten risky assets.

The table refers to Example 7.2. Columns 2-4 shows the expectation, standard deviation, and variance of the returns of the ten assets. Columns 5 and 6 show the portfolio weights of the different assets in the minimum-variance portfolio and the maximum-slope portfolio, respectively. Column 7 shows the portfolio weights of the tangency portfolio which is explained in Section 7.2.

	Asset 1	Asset 2	Asset 3	Asset 4	Asset 5	Asset 6	Asset 7	Asset 8	Asset 9	Asset 10
Asset 1	0.0215	0.0088	0.0068	0.0025	0.0030	0.0048	0.0174	0.0055	0.0075	0.0011
Asset 2	0.0088	0.0270	0.0043	0.0074	0.0085	0.0046	0.0061	0.0131	0.0084	0.0056
Asset 3	0.0068	0.0043	0.0083	0.0058	0.0047	0.0044	0.0064	0.0038	0.0021	0.0029
Asset 4	0.0025	0.0074	0.0058	0.0310	0.0073	0.0057	0.0072	0.0022	0.0067	0.0088
Asset 5	0.0030	0.0085	0.0047	0.0073	0.0158	0.0013	-0.0005	0.0115	0.0080	0.0048
Asset 6	0.0048	0.0046	0.0044	0.0057	0.0013	0.0094	0.0027	0.0021	0.0020	0.0060
Asset 7	0.0174	0.0061	0.0064	0.0072	-0.0005	0.0027	0.0435	0.0020	0.0107	0.0078
Asset 8	0.0055	0.0131	0.0038	0.0022	0.0115	0.0021	0.0020	0.0241	0.0140	0.0051
Asset 9	0.0075	0.0084	0.0021	0.0067	0.0080	0.0020	0.0107	0.0140	0.0230	0.0089
Asset 10	0.0011	0.0056	0.0029	0.0088	0.0048	0.0060	0.0078	0.0051	0.0089	0.0155

Table 7.2: Example with ten risky assets.

The table refers to Example 7.2 and shows the variance-covariance matrix of the ten assets.

	Asset 1	Asset 2	Asset 3	Asset 4	Asset 5	Asset 6	Asset 7	Asset 8	Asset 9	Asset 10
Asset 1	1.000	0.365	0.508	0.096	0.165	0.337	0.568	0.242	0.338	0.063
Asset 2	0.365	1.000	0.284	0.256	0.414	0.289	0.178	0.513	0.338	0.276
Asset 3	0.508	0.284	1.000	0.361	0.407	0.503	0.334	0.265	0.153	0.254
Asset 4	0.096	0.256	0.361	1.000	0.329	0.334	0.196	0.079	0.250	0.401
Asset 5	0.165	0.414	0.407	0.329	1.000	0.103	-0.021	0.587	0.418	0.307
Asset 6	0.337	0.289	0.503	0.334	0.103	1.000	0.135	0.137	0.133	0.496
Asset 7	0.568	0.178	0.334	0.196	-0.021	0.135	1.000	0.061	0.338	0.299
Asset 8	0.242	0.513	0.265	0.079	0.587	0.137	0.061	1.000	0.593	0.265
Asset 9	0.338	0.338	0.153	0.250	0.418	0.133	0.338	0.593	1.000	0.474
Asset 10	0.063	0.276	0.254	0.401	0.307	0.496	0.299	0.265	0.474	1.000

Table 7.3: Example with ten risky assets.

The table refers to Example 7.2 and shows the correlation matrix of the ten assets.

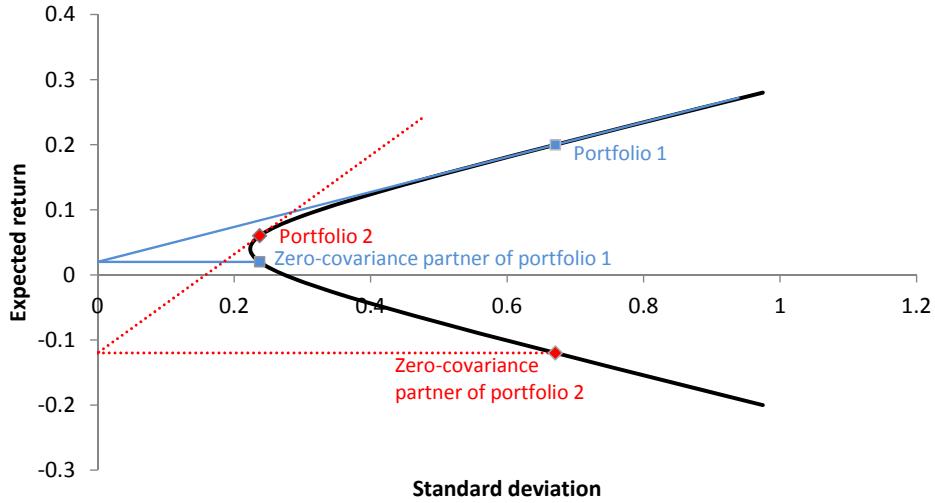


Figure 7.2: Uncorrelated frontier portfolios.

The figure illustrates Theorem 7.5. The two frontier portfolios marked with squares have zero covariance (and thus zero correlation). Similarly, the two frontier portfolios marked with diamonds have zero covariance.

The covariance can also be written as $\pi \cdot \underline{\Sigma} \tilde{\pi}$, cf. Eq. (4.33). Figure 7.2 illustrates the theorem. While this theorem is interesting in itself, it is also important in the derivation of the so-called zero-beta CAPM in Chapter 10.

7.2 Mean-variance analysis with both risky assets and a riskfree asset

A riskfree asset corresponds to the point $(0, r_f)$ in the diagram with standard deviation along the horizontal axis and expected return along the vertical axis. The investors can combine any portfolio of risky assets with an investment in the riskfree asset. As explained in Section 4.1.4, the (standard deviation, mean)-pairs that can be obtained by such a combination form a straight line between the point $(0, r_f)$ and the point (σ, μ) corresponding to the portfolio of risky assets. The slope of the line is exactly the Sharpe ratio of the risky portfolio, $(\mu - r_f)/\sigma$. This is true for combinations with a positive weight on the risky portfolio. If the weight of the risky portfolio is negative, the point will be on the straight line starting at $(0, r_f)$ and having a slope equal to minus the Sharpe ratio. The latter straight line is below the former as long as the Sharpe ratio is positive, i.e., $\mu > r_f$.

The assumption is that the investors want a high expected return and a low standard deviation so they prefer points to the “north-west” in the (standard deviation, mean)-diagram. Figure 7.3 shows that it is important whether the riskfree rate r_f is smaller or greater than the expected return on the minimum-variance portfolio, which is $\mu_{\min} = B/C$ according to Theorem 7.2. In either case the mean-variance efficient portfolios of all assets are combinations of the riskfree asset and a **tangency portfolio** of the risky assets. The tangency portfolio is the portfolio corresponding to the point where the straight line starting at $(0, r_f)$ is tangent to the mean-variance frontier of risky assets. We distinguish between two cases:

1. $B > C r_f$: then the riskfree rate is smaller than μ_{\min} as in the left panel of Figure 7.3. The tangency portfolio is on the upward-sloping branch of the mean-variance frontier

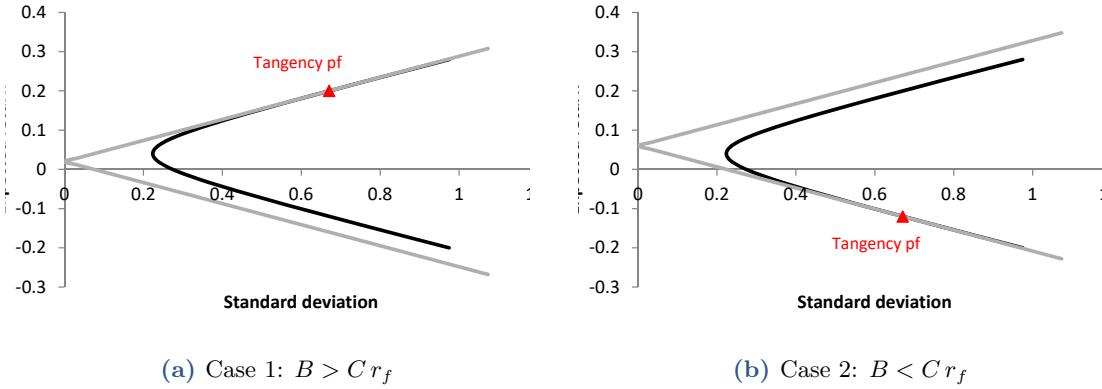


Figure 7.3: Two cases of efficient frontiers.

The graphs show the mean-variance frontier of all assets and the location of the tangency portfolio (indicated by the triangle). Both diagrams assume $A = 0.096$, $B = 0.8$, and $C = 20$ which, among other things, imply that the expected return on the minimum-variance portfolio is $B/C = 0.04 = 4\%$. The riskfree rate is 2% in the left panel and 6% in the right panel.

of risky assets. The highest Sharpe ratio is obtained on the tangency line, which is sometimes called the *capital allocation line*. Points on this line are reached by combining a long position in the tangency portfolio with the riskfree asset. Points between $(0, r_f)$ and the tangency point correspond to combinations with positive weights in both the riskfree asset and the tangency portfolio. Points on the tangency line above and to the right of the tangency point correspond to combinations with a short position in the riskfree asset, i.e., a leveraged position in the tangency portfolio.

2. $B < Cr_f$: then the riskfree rate is greater than μ_{\min} as in the right panel of Figure 7.3. The tangency portfolio is on the downward-sloping branch of the mean-variance frontier of risky assets. The highest Sharpe ratio is obtained on the upward-sloping straight line. Points on this line are reached by combining a short position in the tangency portfolio with a position of more than 100% in the riskfree asset.

In either case the wedge consisting of the two straight lines represents the mean-variance frontier or efficient frontier of *all* assets. Obviously, mean-variance optimizers would never pick a portfolio corresponding to a point on the downward-sloping straight line so the upward-sloping line is the interesting part of the frontier.

Of the two cases, Case 1 appears to be the most natural, but it seems impossible to rule out Case 2 based on mathematical or economical arguments. If the necessary inputs (expected returns, variances, and covariances) are estimated using historical returns, you will most often end up in Case 1 but sometimes in Case 2.

Obviously, there is also a third case in which $B = Cr_f$ so that the riskfree rate equals the expected return on the minimum-variance portfolio. This case is even weirder than Case 2 as no tangency portfolio exists and the mean-variance efficient portfolios has 100% invested in the riskfree asset and then a position in a specific zero-investment portfolio of the risky assets. The latter is a portfolio with long positions in some assets and short positions in other assets so that the net investment is zero. By scaling up both the long and the short positions, you change both the expected return and the standard deviation of your total, combined portfolio. However, $B \neq Cr_f$ unless the values of the expected returns, variances, and covariances are carefully selected to obtain $B = Cr_f$ so, in practice, the case $B = Cr_f$ is uninteresting.

Let us return to Case 1 and Case 2. The next theorem characterizes the tangency portfolio.

Theorem 7.6

Suppose that $B \neq C r_f$. Then the tangency portfolio is the portfolio of risky assets represented by the portfolio weight vector

$$\boldsymbol{\pi}_{\tan} = \frac{\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} = \frac{1}{B - C r_f} \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) \quad (7.34)$$

and has expected return, variance, and standard deviation given by

$$\mu_{\tan} = \frac{\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} = \frac{A - B r_f}{B - C r_f}, \quad (7.35)$$

$$\sigma_{\tan}^2 = \frac{(\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})}{(\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}))^2} = \frac{A - 2B r_f + C r_f^2}{(B - C r_f)^2}, \quad (7.36)$$

$$\sigma_{\tan} = \frac{((\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}))^{1/2}}{|\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})|} = \frac{\sqrt{A - 2B r_f + C r_f^2}}{|B - C r_f|}. \quad (7.37)$$

The absolute value and the square of the Sharpe ratio are

$$\left| \frac{\mu_{\tan} - r_f}{\sigma_{\tan}} \right| = \left((\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) \right)^{1/2} = \sqrt{A - 2B r_f + C r_f^2}, \quad (7.38)$$

$$\left(\frac{\mu_{\tan} - r_f}{\sigma_{\tan}} \right)^2 = (\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) = A - 2B r_f + C r_f^2. \quad (7.39)$$

Moreover, the ratio of the excess expected return to the variance is

$$\frac{\mu_{\tan} - r_f}{\sigma_{\tan}^2} = \mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) = B - C r_f. \quad (7.40)$$

The tangency portfolio has the property that

$$\frac{\mathbb{E}[r_i] - r_f}{\text{Cov}[r_i, r_{\tan}]} = \frac{\mathbb{E}[r_{\tan}] - r_f}{\text{Var}[r_{\tan}]}, \quad \text{for all } i = 1, 2, \dots, N, \quad (7.41)$$

where r_{\tan} is the return on the tangency portfolio.

Proof

The Sharpe ratio of a portfolio $\boldsymbol{\pi}$ is the ratio $(\mu(\boldsymbol{\pi}) - r_f)/\sigma(\boldsymbol{\pi})$. Since the tangency portfolio is on the efficient frontier, we know that its mean and variance satisfy the rela-

tion (7.8). The Sharpe ratio for the efficient portfolio with expected return $\bar{\mu}$ is therefore

$$f(\bar{\mu}) = \frac{\bar{\mu} - r_f}{\sigma(\bar{\mu})} = \frac{(\bar{\mu} - r_f)\sqrt{D}}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}.$$

In case 1 we want to maximize $f(\bar{\mu})$ and in case 2 we want to minimize $f(\bar{\mu})$. In both cases, the problem thus boils down to solving $f'(\bar{\mu}) = 0$. Applying standard rules for differentiation of quotients and square roots, we find that

$$\begin{aligned} f'(\bar{\mu}) &= \frac{\sqrt{D}\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} - (\bar{\mu} - r_f)\sqrt{D}\frac{2C\bar{\mu} - 2B}{2\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}}{C\bar{\mu}^2 - 2B\bar{\mu} + A} \\ &= \frac{\sqrt{D}}{C\bar{\mu}^2 - 2B\bar{\mu} + A} \left(\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} - \frac{(\bar{\mu} - r_f)(C\bar{\mu} - B)}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}} \right). \end{aligned}$$

Next, observe that $f'(\bar{\mu}) = 0$ when

$$\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A} = \frac{(\bar{\mu} - r_f)(C\bar{\mu} - B)}{\sqrt{C\bar{\mu}^2 - 2B\bar{\mu} + A}}$$

or, equivalently,

$$C\bar{\mu}^2 - 2B\bar{\mu} + A = (\bar{\mu} - r_f)(C\bar{\mu} - B) = C\bar{\mu}^2 - (B + Cr_f)\bar{\mu} + Br_f,$$

which implies that

$$\bar{\mu} = \frac{A - Br_f}{B - Cr_f}.$$

This verifies the right-most expression in (7.35), whereas the first part is clear once (7.34) is established since $\mu_{\tan} = \boldsymbol{\mu} \cdot \boldsymbol{\pi}_{\tan}$.

Substituting the above expression for $\bar{\mu}$ into (7.7), we get

$$\begin{aligned} \boldsymbol{\pi} &= \frac{C \frac{A - Br_f}{B - Cr_f} - B}{D} \underline{\Sigma}^{-1} \boldsymbol{\mu} + \frac{A - B \frac{A - Br_f}{B - Cr_f}}{D} \underline{\Sigma}^{-1} \mathbf{1} \\ &= \frac{[C(A - Br_f) - B(B - Cr_f)] \underline{\Sigma}^{-1} \boldsymbol{\mu} + [A(B - Cr_f) - B(A - Br_f)] \underline{\Sigma}^{-1} \mathbf{1}}{D(B - Cr_f)} \\ &= \frac{(AC - B^2) \underline{\Sigma}^{-1} \boldsymbol{\mu} - (AC - B^2)r_f \underline{\Sigma}^{-1} \mathbf{1}}{D(B - Cr_f)} \\ &= \frac{\underline{\Sigma}^{-1} \boldsymbol{\mu} - r_f \underline{\Sigma}^{-1} \mathbf{1}}{B - Cr_f} = \frac{1}{B - Cr_f} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}), \end{aligned}$$

which shows (7.34). Note that

$$\mathbf{1} \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) = [\mathbf{1} \cdot \underline{\Sigma}^{-1} \boldsymbol{\mu}] - r_f [\mathbf{1} \cdot \underline{\Sigma}^{-1} \mathbf{1}] = B - Cr_f. \quad (7.42)$$

Substituting the above value of $\bar{\mu}$ into (7.8), we find the variance of the tangency port-

folio:

$$\begin{aligned}\sigma_{\tan}^2 &= \frac{C \left(\frac{A-B r_f}{B-C r_f} \right)^2 - 2B \frac{A-B r_f}{B-C r_f} + A}{D} \\ &= \frac{C (A - B r_f)^2 - 2B(A - B r_f)(B - C r_f) + A(B - C r_f)^2}{D(B - C r_f)^2} \\ &= \frac{(AC - B^2)(A - 2B r_f + C r_f^2)}{D(B - C r_f)^2} = \frac{A - 2B r_f + C r_f^2}{(B - C r_f)^2},\end{aligned}$$

where the third equality involves some tedious, but straightforward computations. It can be shown that $(\mu - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\mu - r_f \mathbf{1}) = A - 2B r_f + C r_f^2$, and then we have confirmed both the variance expressions in (7.36). The standard deviation in (7.37) follows immediately.

The Sharpe ratio is

$$\begin{aligned}\frac{\mu_{\tan} - r_f}{\sigma_{\tan}} &= \frac{\frac{A-B r_f}{B-C r_f} - r_f}{\sqrt{\frac{A-2B r_f+C r_f^2}{|B-C r_f|}}} = \frac{\frac{A-B r_f}{B-C r_f} - r_f \frac{B-C r_f}{B-C r_f}}{\sqrt{\frac{A-2B r_f+C r_f^2}{|B-C r_f|}}} \\ &= \frac{A - 2B r_f + C r_f^2}{\sqrt{A - 2B r_f + C r_f^2}} \frac{|B - C r_f|}{B - C r_f} = \sqrt{A - 2B r_f + C r_f^2} \frac{|B - C r_f|}{B - C r_f},\end{aligned}$$

from which (7.38) and (7.39) follow.

Finally, to show (7.41) first recall from Eq. (4.42) that $\underline{\Sigma} \boldsymbol{\pi}_{\tan}$ is the vector of covariances of the individual assets with the portfolio $\boldsymbol{\pi}_{\tan}$. Because of (7.34), the vector of covariances of the individual assets with the tangency portfolio is

$$\underline{\Sigma} \boldsymbol{\pi}_{\tan} = \frac{1}{B - C r_f} (\mu - r_f \mathbf{1}).$$

This means that for any asset $i = 1, 2, \dots, N$, we have

$$\text{Cov}[r_i, r_{\tan}] = \frac{\text{E}[r_i] - r_f}{B - C r_f} \Leftrightarrow \frac{\text{E}[r_i] - r_f}{\text{Cov}[r_i, r_{\tan}]} = B - C r_f. \quad (7.43)$$

From (7.35) and (7.36) it follows that

$$\begin{aligned}\mu_{\tan} - r_f &= \frac{A - B r_f}{B - C r_f} - r_f = \frac{A - B r_f - r_f (B - C r_f)}{B - C r_f} \\ &= \frac{A - 2B r_f + C r_f^2}{B - C r_f} = (B - C r_f) \sigma_{\tan}^2,\end{aligned}$$

which verifies (7.40) and implies that

$$B - C r_f = \frac{\text{E}[r_{\tan}] - r_f}{\text{Var}[r_{\tan}]} \quad (7.44)$$

By combining (7.43) and (7.44) we obtain (7.41).

Since the tangency portfolio is the one maximizing (in Case 1) the Sharpe ratio, we expect that individual assets with large Sharpe ratios have a relatively large weight in the tangency portfolio. However, the correlations are also important. In order to diversify risk away, the tangency portfolio might give a large weight to an asset with a low Sharpe ratio if that asset has low correlation with assets having a large Sharpe ratio.

Let us turn to the property (7.41). It implies that the portfolio weights of the tangency portfolio are selected so that the ratio of the expected excess return on each asset to the covariance of the asset with the tangency portfolio is the same for all assets. Does this make sense? Consider what happens to the variance of a portfolio if you change one of the weights, say π_i , a little bit. This is reflected by the partial derivative of the variance with respect to π_i . The portfolio variance is

$$\text{Var}[r(\boldsymbol{\pi})] = \sum_{i=1}^N \pi_i^2 \text{Var}[r_i] + 2 \sum_{i=1}^N \sum_{\substack{j>i \\ j=1}}^N \pi_i \pi_j \text{Cov}[r_i, r_j],$$

cf. Eq. (4.27). The partial derivative $\partial \text{Var}[r(\boldsymbol{\pi})]/\partial \pi_i$ for a fixed i involves only the terms in $\text{Var}[r(\boldsymbol{\pi})]$ where π_i enters, which is $\pi_i^2 \text{Var}[r_i]$ from the first term and, for every $j \neq i$, $2\pi_i \pi_j \text{Cov}[r_i, r_j]$ from the second term. Hence, the partial derivative is

$$\begin{aligned} \frac{\partial \text{Var}[r(\boldsymbol{\pi})]}{\partial \pi_i} &= 2\pi_i \text{Var}[r_i] + 2 \sum_{\substack{j \neq i \\ j=1}}^N \pi_j \text{Cov}[r_i, r_j] = 2 \sum_{j=1}^N \pi_j \text{Cov}[r_i, r_j] \\ &= 2 \text{Cov} \left[r_i, \sum_{j=1}^N \pi_j r_j \right] = 2 \text{Cov} [r_i, r(\boldsymbol{\pi})]. \end{aligned} \quad (7.45)$$

Here we have used that $\text{Var}[r_i] = \text{Cov}[r_i, r_i]$ and the covariance property (3.49). The covariance of the asset's return with the portfolio return measures how the portfolio variance reacts to a marginal change in the weight of the asset, so the covariance is measuring the marginal risk of the asset. When considering whether or not to increase the weight of asset i , the relevant tradeoff is between the excess expected return $E[r_i] - r_f$ and the marginal effect on portfolio risk, i.e., $\text{Cov}[r_i, r(\boldsymbol{\pi})]$. Therefore it is reasonable that the tangency portfolio is set up so that this marginal risk-return tradeoff is the same for all assets. If one asset had a higher ratio $(E[r_i] - r_f)/\text{Cov}[r_i, r(\boldsymbol{\pi})]$ than the others, we should have invested a larger fraction of wealth in that asset. As we shall see in Section 10.1, the relation (7.41) is a key ingredient in the Capital Asset Pricing Model (the CAPM).

The considerations above Theorem 7.6 show that, whenever $B \neq Cr_f$, we have a two-fund separation result again:

Theorem 7.7

Assume $B \neq Cr_f$. Then any mean-variance optimal portfolio of a riskfree and one or more risky assets can be written as a combined portfolio of the riskfree asset and the tangency portfolio of the risky assets. If w is the weight on the tangency portfolio so that $1-w$ is the weight on the riskfree asset, the expected return and standard deviation of the combined portfolio are

$$\mu(w) = w \mu_{\tan} + (1-w)r_f, \quad \sigma(w) = |w| \sigma_{\tan}. \quad (7.46)$$

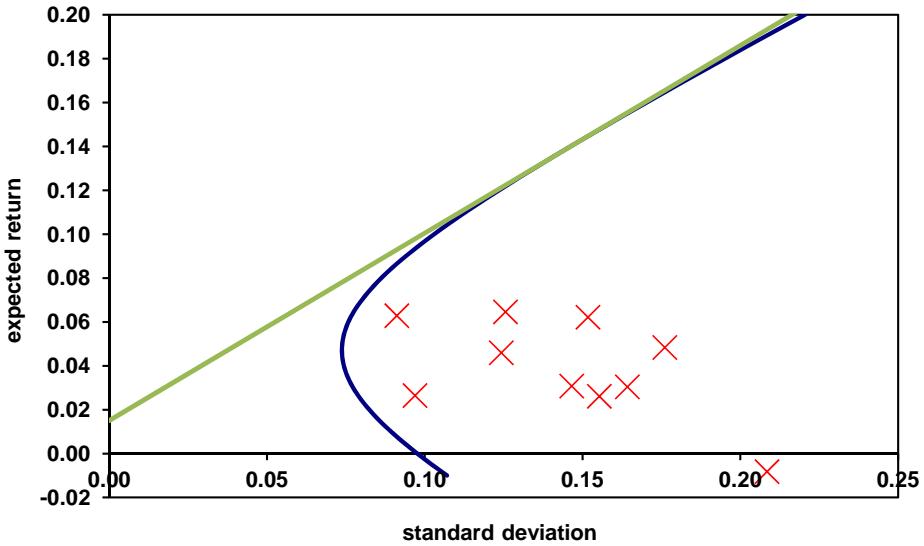


Figure 7.4: The mean-variance frontier of all assets.

The curve shows the mean-variance frontier generated from the 10 individual assets corresponding to the x's. The straight line is the mean-variance frontier of all assets, assuming a riskfree return of 1.5%.

We can therefore generate the mean-variance frontier of all assets by first determining the tangency portfolio of the risky assets—the necessary computation are easily done in Excel—and then considering various combinations of the tangency portfolio and the riskfree asset.

Even with access to a riskfree asset, we might still be interested in the efficient frontier of risky assets. Since the tangency portfolio lies on this frontier, we can generate the entire frontier of risky assets by considering a number of combinations of the tangency portfolio and another known frontier portfolio such as the minimum-variance portfolio.

If all investors agree on the riskfree rate and the expected returns, variances, and covariances of the risky assets, they will all agree on the composition of the tangency portfolio and on the location of the mean-variance frontier. Everybody would then hold exactly the same portfolio of risky assets in some combination with the riskfree asset, depending on their willingness to take risk as we will discuss in Section 7.3 below.

Example 7.3

Recall that Figure 7.1 presented the mean-variance efficient frontier generated by 10 risky assets considered in Example 7.2. Now suppose the investor can also invest in a riskfree asset with a return of $r_f = 0.015 = 1.5\%$. The right-most column of Table 7.1 shows the composition of the tangency portfolio in this case, and Figure 7.4 shows (the upward-sloping part of) the resulting mean-variance frontier of all assets.

In Section 4.1.3 we already found the portfolio of two assets that is maximizing the Sharpe ratio, i.e. the tangency portfolio with $N = 2$ assets. The next example shows mathematically that with $N = 2$ the tangency portfolio from Theorem 7.6 is consistent

with the results from Section 4.1.3.

Example 7.4

With $N = 2$, $\underline{\Sigma}^{-1}$ is given by (7.21) and

$$\begin{aligned}\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \begin{pmatrix} \mu_1 - r_f \\ \mu_2 - r_f \end{pmatrix} \\ &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_2 - r_f) \\ \sigma_1^2(\mu_2 - r_f) - \rho\sigma_1\sigma_2(\mu_1 - r_f) \end{pmatrix}. \end{aligned}\quad (7.47)$$

This implies that

$$\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \left(\sigma_1^2(\mu_2 - r_f) + \sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_1 + \mu_2 - 2r_f) \right),$$

and substituting into Eq. (7.34) we find

$$\boldsymbol{\pi}_{\tan} = \frac{\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})} = \begin{pmatrix} \frac{\sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_2 - r_f)}{\sigma_1^2(\mu_2 - r_f) + \sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_1 + \mu_2 - 2r_f)} \\ \frac{\sigma_1^2(\mu_2 - r_f) - \rho\sigma_1\sigma_2(\mu_1 - r_f)}{\sigma_1^2(\mu_2 - r_f) + \sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_1 + \mu_2 - 2r_f)} \end{pmatrix}. \quad (7.48)$$

If we let $\lambda_1 = \frac{\mu_1 - r_f}{\sigma_1}$ and $\lambda_2 = \frac{\mu_2 - r_f}{\sigma_2}$ denote the Sharpe ratios of the two assets, we can rewrite the tangency portfolio as

$$\boldsymbol{\pi}_{\tan} = \begin{pmatrix} \frac{\lambda_1 - \rho\lambda_2}{\lambda_1 - \rho\lambda_2 + \frac{\sigma_1}{\sigma_2}(\lambda_2 - \rho\lambda_1)} \\ \frac{\lambda_2 - \rho\lambda_1}{\lambda_2 - \rho\lambda_1 + \frac{\sigma_2}{\sigma_1}(\lambda_1 - \rho\lambda_2)} \end{pmatrix}, \quad (7.49)$$

which indeed is consistent with Eq. (4.16). As long as $\lambda_1 > \rho\lambda_2$ and $\lambda_2 > \rho\lambda_1$, both assets have a positive weight. Eq. (7.47) can be written in terms of the Sharpe ratios as

$$\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2\lambda_1\sigma_1 - \rho\sigma_1\sigma_2^2\lambda_2 \\ \sigma_1^2\lambda_2\sigma_2 - \rho\sigma_1^2\sigma_2\lambda_1 \end{pmatrix},$$

and by (7.39) the squared Sharpe ratio is thus

$$\begin{aligned}(\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \lambda_1\sigma_1 \\ \lambda_2\sigma_2 \end{pmatrix} \cdot \begin{pmatrix} \sigma_2^2\lambda_1\sigma_1 - \rho\sigma_1\sigma_2^2\lambda_2 \\ \sigma_1^2\lambda_2\sigma_2 - \rho\sigma_1^2\sigma_2\lambda_1 \end{pmatrix} \\ &= \frac{1}{1-\rho^2} (\lambda_1^2 + \lambda_2^2 - 2\rho\lambda_1\lambda_2).\end{aligned}$$

Hence, with two assets the maximum Sharpe ratio is

$$\frac{\mu_{\tan} - r}{\sigma_{\tan}} = \sqrt{\frac{1}{1-\rho^2} (\lambda_1^2 + \lambda_2^2 - 2\rho\lambda_1\lambda_2)}, \quad (7.50)$$

which is consistent with Eq. (4.17). The maximum Sharpe ratio is always greater than or equal to λ_1 and λ_2 .

With a single risky asset (or asset class), this asset has a weight of 100% in the tangency portfolio, and the Sharpe ratio of the tangency portfolio equals λ_1 . By introducing the second asset, the weight of the first asset in the tangency portfolio changes to the first component in the vector in (7.49). The new weight is smaller than 100% if $\lambda_2 > \rho\lambda_1$ and $\lambda_1 > \rho\lambda_2$ so, under these conditions, the addition of a second asset lowers the investment in the first asset. This seems natural as the investor benefits from diversification. Assuming $\lambda_1, \lambda_2 > 0$, the conditions are definitely true if the correlation ρ is zero or negative.

Example 7.5

Let us consider a numerical example with two risky assets. The first risky asset is a stock index, whereas the second is a long-term government bond. We take the historical estimates from the U.S. shown in Table 6.6 in the preceding chapter as representative of future investment opportunities. All returns are measured per year and we implement the mean-variance analysis with a period length of one year. The historical average real return on the U.S. stock market is $\mu_1 = 8.3\% = 0.083$ with a standard deviation of $\sigma_1 = 19.9\% = 0.199$, whereas the average real return on bonds is $\mu_2 = 2.4\% = 0.024$ with a standard deviation of $\sigma_2 = 10.3\% = 0.103$. The correlation between stock returns and bond returns is $\rho = 0.2$. Because the average real U.S. short-term interest rate is 0.9%, we assume a riskfree rate of $r_f = 0.009$. The Sharpe ratios of the stock index and the bond are therefore $\lambda_1 \approx 0.372$ and $\lambda_2 = 0.146$, respectively.

The tangency portfolio is in this case given by

$$\boldsymbol{\pi}_{\text{tan}} = \begin{pmatrix} 0.713 \\ 0.287 \end{pmatrix}$$

corresponding to 71.3% in the stock index and 28.7% in the long-term bond. The tangency portfolio has a mean return of 6.61%, a standard deviation of 15.1%, and a Sharpe ratio of 0.379 which is a small increase from the Sharpe ratio of the stock index. Adding long-term government bonds to a diversified stock investment does not improve the overall risk-return tradeoff by much given the parameter values used.

In Figure 7.5 the solid blue curve shows the mean-variance efficient portfolios of risky assets, i.e., the combinations of expected returns and volatility that can be obtained by combining the bond and the stock. The solid orange line corresponds to the efficient frontier of all assets, that is, the stock, the bond, and the one-year riskfree asset.

As discussed in Section 6.8, the stock-bond correlations varies over time. The above calculations assume a correlation of 0.2, but suppose instead the correlation is -0.2 . Then the efficient frontier of risky assets changes to the dotted blue curve and the efficient frontier of all assets to the dotted orange line. In particular, the frontier of all assets becomes steeper, which indicates an improved risk-return tradeoff. Indeed, the Sharpe ratio of the new tangency portfolio is 0.434, compared to 0.379 from before and the stock's Sharpe ratio of 0.372. With the negative correlation, the diversification benefits offered by the bond are significantly larger than with the positive correlation.

7.3 The optimal portfolio

Theorem 7.7 concludes that any mean-variance optimizing investor chooses a combination of the riskfree asset and the tangency portfolio of risky assets. But which combination?

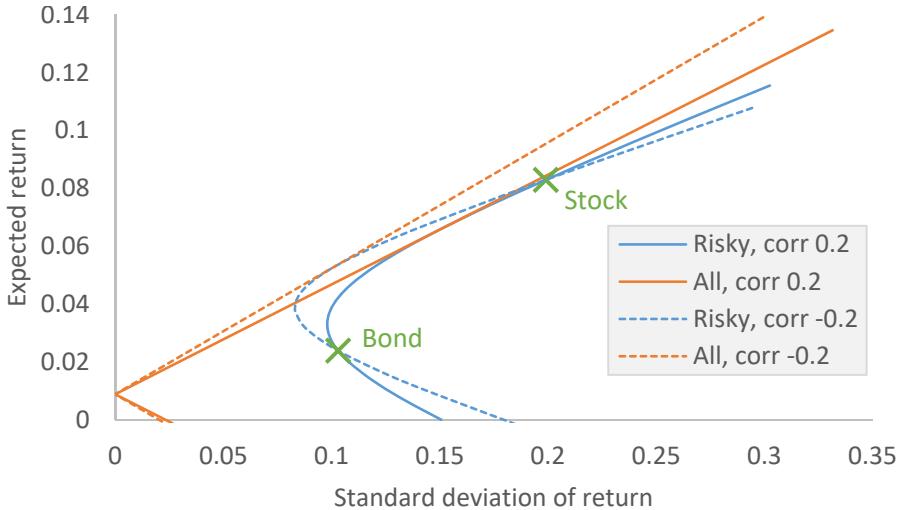


Figure 7.5: The mean-variance frontiers with a stock and a bond.

The figure shows the mean-variance frontier without the riskfree asset (blue curve) and with the riskfree asset (straight orange line). The solid curves are for a stock-bond correlation of 0.2, whereas the dotted curves assume a correlation of -0.2 .

Let w denote the fraction of wealth invested in the tangency portfolio so that the fraction $1 - w$ of wealth is invested in the riskfree asset. Just as in (4.18), the rate of return on the combined portfolio is

$$r(w) = wr_{\tan} + (1 - w)r_f \quad (7.51)$$

with mean and variance

$$\mu(w) = w\mu_{\tan} + (1 - w)r_f = r_f + w(\mu_{\tan} - r_f), \quad (7.52)$$

$$\sigma^2(w) = w^2\sigma_{\tan}^2. \quad (7.53)$$

The problem is to find the optimal w and this depends on the mean-variance trade-off of the investor.

To simplify the discussion, let us focus on the more natural Case 1 in which the relevant mean-variance diagram resembles the left panel of Figure 7.3 and the relevant values of w are non-negative since they correspond to the upward-sloping line.

We can represent the mean-variance preferences of the investor by *indifference curves* in the (σ, μ) -diagram. By definition, the investor is indifferent between all the points on an indifference curve. Mean-variance optimizers have increasing indifference curves because if we start out at given point (σ, μ) and increase the standard deviation slightly, then we need to increase the expected return as well in order to maintain the satisfaction of the investor. Any two indifference curves of the same investor correspond to two different levels of satisfaction and cannot cross each other. The investor prefers the highest possible indifference curve in the (σ, μ) -diagram since that gives the highest expected return for any fixed value of the standard deviation. We expect indifference curves to be convex, and thus becoming steeper as we increase σ , since the investor requires a high increase in μ for a given increase in σ if she already has a high σ . An investor's optimal combination of standard deviation and expected return is found where an indifference curve is tangent to the upward-sloping, mean-variance efficient line.

If we think of standard deviation (or, equivalently, variance) as measuring the risk of

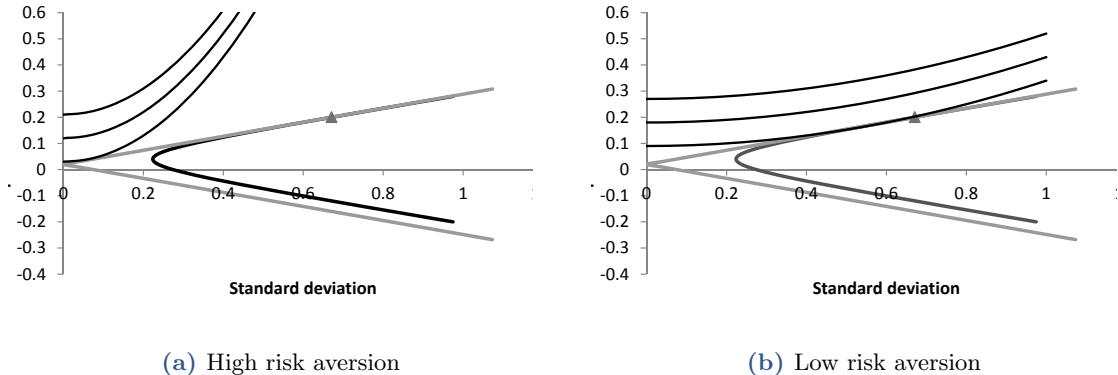


Figure 7.6: Determining the optimal portfolio.

The optimal portfolio for a given investor is determined by the point where the upward-sloping part of the mean-variance frontier is tangent to an indifference curve. The mean-variance frontier is taken from the left panel of Figure 7.3. The indifference curves are constructed from the preferences represented by (7.54) using $a = 5$ in the left panel and $a = 0.5$ in the right panel.

the investment, all mean-variance optimizers are risk averse. A highly risk-averse investor has very steep indifference curves in the (σ, μ) -diagram, as she requires a large increase in μ to compensate for a small increase in σ . Then the indifference curve will be tangent to the efficient frontier somewhere on the left part of the upwards-sloping straight line, corresponding to a low σ . This situation is illustrated in the left panel of Figure 7.6 where we have added three relatively steep indifference curves to the mean-variance picture in the left panel of Figure 7.3. This is implemented by taking a low (but positive) weight w on the tangency portfolio and thus a large weight on the riskfree asset. Conversely, an investor who is only slightly risk averse (i.e., relatively risk tolerant) has flatter, but still increasing, indifference curves. Then the indifference curve will be tangent to the efficient frontier further up the straight line, corresponding to a relatively high w (maybe even higher than 1 which involves borrowing) and a relatively high standard deviation σ . This situation is illustrated in the right panel of Figure 7.6.

To find an explicit expression for the optimal value of w , we need to formalize the mean-variance trade-off of the investor. Suppose the objective of the investor is to maximize the expected rate of return minus a constant times the variance of the rate of return,

$$\max \left(E[r] - \frac{1}{2} \gamma \text{Var}[r] \right), \quad (7.54)$$

where γ is a positive constant. A higher γ corresponds to a larger penalty for variance so it is a measure of the investor's risk aversion. In fact, as we shall see in Section 7.5.3, γ is the so-called relative risk aversion coefficient. With this mean-variance tradeoff, the optimal portfolio is as stated in the following theorem.

Theorem 7.8

With the mean-variance criterion (7.54), the optimal fraction of wealth invested in the

tangency portfolio is

$$w^* = \frac{\mu_{\tan} - r_f}{\gamma \sigma_{\tan}^2} = \frac{\lambda_{\tan}}{\gamma \sigma_{\tan}}, \quad (7.55)$$

where $\lambda_{\tan} = (\mu_{\tan} - r_f)/\sigma_{\tan}$ is the Sharpe ratio of the tangency portfolio. The corresponding vector of portfolio weights in the individual risky assets is

$$\boldsymbol{\pi}^* = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}). \quad (7.56)$$

The optimal fraction to invest in the riskfree asset is $1 - w^* = 1 - \boldsymbol{\pi}^* \cdot \mathbf{1}$.

Proof

With the expressions (7.52) and (7.53) for the mean and the variance, the objective function can be rewritten as

$$f(w) = \mu(w) - \frac{1}{2} \gamma \sigma(w)^2 = r_f + w(\mu_{\tan} - r_f) - \frac{1}{2} \gamma w^2 \sigma_{\tan}^2. \quad (7.57)$$

The derivative is

$$f'(w) = \mu_{\tan} - r_f - \gamma \sigma_{\tan}^2,$$

and solving $f'(w) = 0$, we get the first equality in (7.55). The second equality follows from the definition of the Sharpe ratio. The optimal vector of portfolio weights in the individual risky assets is

$$\boldsymbol{\pi}^* = w^* \boldsymbol{\pi}_{\tan} = \frac{\lambda_{\tan}}{\gamma \sigma_{\tan}} \frac{\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})},$$

which by the use of Eq. (7.40) leads to (7.56).

As we would expect, the fraction w^* of wealth optimally invested in the tangency portfolio is increasing in the expected excess return of the tangency portfolio, decreasing in the variance of the tangency portfolio, and decreasing in the risk aversion γ .

Note that when you apply (7.55), it matters whether you plug in decimal points or percentages on the right-hand side. Suppose that the riskfree rate is 4% and the tangency portfolio has an expected return of 12% and a standard deviation of 20%. Using decimal points you get

$$w^* = \frac{0.12 - 0.04}{\gamma \times (0.20)^2} = \frac{2}{\gamma}.$$

For example for a risk aversion of $\gamma = 4$, you get $w^* = 0.5$. whereas if you use percentages you seem to get

$$w^* = \frac{12 - 4}{\gamma \times (20)^2} = \frac{0.02}{\gamma}.$$

With $\gamma = 4$, you seem to get $w^* = 0.005$, which is off by a factor 100. The explanation is

subtle. If we want to use percentages, we should really include the percentage signs:

$$w^* = \frac{12\% - 4\%}{\gamma \times (20\%)^2} = \frac{0.02(\%)^{-1}}{\gamma}.$$

Now just as $0.02 = 2\%$, we have $0.02(\%)^{-1} = 2$, and then we are back at the correct result of $w^* = 2/\gamma$. To avoid any confusion, use decimal point inputs when applying (7.55).

Suppose we have a single risky asset with expected return μ_1 , standard deviation σ_1 , and Sharpe ratio $\lambda_1 = (\mu_1 - r_f)/\sigma_1$. Then the optimal portfolio consists of the weight

$$\pi_1^* = \frac{\mu_1 - r_f}{\gamma\sigma_1^2} = \frac{\lambda_1}{\gamma\sigma_1}$$

in the risky assets and the weight $1 - \pi_1^*$ in the riskfree asset. What happens to the optimal weight of asset 1 if we add a second risky asset? Asset 2 has expected return μ_2 , standard deviation σ_2 , and Sharpe ratio $\lambda_2 = (\mu_2 - r_f)/\sigma_2$. From (7.47) and (7.56), we see that the optimal weight of asset 1 is then

$$\begin{aligned} \pi_1^* &= \frac{1}{\gamma(1 - \rho^2)\sigma_1^2\sigma_2^2} \left[\sigma_2^2(\mu_1 - r_f) - \rho\sigma_1\sigma_2(\mu_2 - r_f) \right] \\ &= \frac{1}{\gamma(1 - \rho^2)} \left[\frac{\mu_1 - r_f}{\sigma_1^2} - \frac{\rho}{\sigma_1} \frac{\mu_2 - r_f}{\sigma_2} \right] = \frac{1}{\gamma(1 - \rho^2)} \left[\frac{\lambda_1}{\sigma_1} - \frac{\rho}{\sigma_1} \lambda_2 \right] \\ &= \frac{1}{\gamma(1 - \rho^2)} \left[(1 - \rho^2) \frac{\lambda_1}{\sigma_1} + \rho^2 \frac{\lambda_1}{\sigma_1} - \frac{\rho}{\sigma_1} \lambda_2 \right] \\ &= \frac{\lambda_1}{\gamma\sigma_1} + \frac{\rho}{\gamma(1 - \rho^2)\sigma_1} (\rho\lambda_1 - \lambda_2), \end{aligned}$$

so the second term in the last expression is the extra weight put on the first asset due to the inclusion of the second asset. Adding an uncorrelated asset does not change the optimal weight of the first asset. Adding a positively correlated asset lowers the weight of the first asset if and only if $\lambda_2 > \rho\lambda_1$. The investor diversifies his position by spreading it over both risky assets. If both λ_1 and λ_2 are positive, adding a negatively correlated asset increases the optimal weight of the first asset. The investor can increase his exposure to the first asset because the second asset acts a hedge due to the negative correlation. Note that in the two-asset case, the optimal portfolio weights can be written as

$$\pi_1^* = \frac{\lambda_1 - \rho\lambda_2}{\gamma\sigma_1(1 - \rho^2)}, \quad \pi_2^* = \frac{\lambda_2 - \rho\lambda_1}{\gamma\sigma_2(1 - \rho^2)}, \quad (7.58)$$

where the expression for π_1^* follows from the second line in the above calculation and the expression for π_2^* then follows by symmetry.

7.4 Discussion and perspectives

7.4.1 Inputs based on estimation

In order to implement Markowitz' mean-variance analysis we need to fix the investment horizon (the length of the period considered) as well as the set of assets included in the analysis, and we need values for the expected returns and the variance-covariance matrix of the returns of the assets. The expectation μ and the variance-covariance matrix Σ are

often replaced by their sample estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ from a time series of returns, but these estimates can be quite imprecise. This is potentially problematic because the output of the mean-variance optimization turns out to be quite sensitive to the magnitudes of the inputs.

[Chopra and Ziemba \(1993\)](#) show that it is particularly important to obtain precise estimates of the expected returns. Unfortunately, the expected returns are hard to estimate precisely from historical returns, cf. [Merton \(1980\)](#) and our discussion in Section 3.7.2. One idea is to derive so-called implied expected returns from observed market weights of different assets and an equilibrium model such as the CAPM, as we shall discuss in Section 10.1.6. The sample estimate of the variance-covariance matrix can also be imprecise, in particular when the number of assets is large compared to the number of observations of each return in the sample. The most extreme coefficients in the sample variance-covariance matrix tend to extreme because of estimation errors, not because the true coefficient is extreme. The extreme estimates often lead to extreme portfolio weights.

Shrinkage is a common approach to reduce estimation errors. The raw sample estimate of a parameter is combined with other information in a way that moderates extreme sample estimates. In its simplest form, the shrinkage estimate of a parameter is a weighted average of the sample estimate and a target or benchmark value. For example, [Jorion \(1986\)](#) suggests applying the so-called Bayes-Stein shrinkage approach which means that the estimate of the vector of expected returns is

$$\boldsymbol{\mu}_{\text{BS}} = (1 - \varphi)\tilde{\boldsymbol{\mu}} + \varphi\tilde{\boldsymbol{\mu}}_{\min}\mathbf{1}.$$

Here $\tilde{\boldsymbol{\mu}}_{\min} = \mathbf{1} \cdot \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\mu}} / \mathbf{1} \cdot \tilde{\Sigma}^{-1} \mathbf{1}$ is the sample version of the ratio B/C , i.e. the sample estimate of the expected return on the minimum-variance portfolio, see Eq. (7.17) and the expressions for B and C in (7.6). The weighting constant φ is given by

$$\varphi = \frac{N+2}{N+2+T(\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{\min}\mathbf{1}) \cdot \tilde{\Sigma}^{-1} (\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_{\min}\mathbf{1})},$$

where N is the number of assets and T is the number of return observations for each asset. The sample estimate of the expected return on each asset is moved towards the estimated expected return on the minimum-variance portfolio. The Bayes-Stein shrinkage estimate for the variance-covariance matrix is

$$\tilde{\Sigma}_{\text{BS}} = \left(1 + \frac{1}{T+\varphi}\right)\tilde{\Sigma} + \frac{\varphi}{T(T+1+\varphi)} \frac{\mathbf{1}\mathbf{1}^\top}{\mathbf{1} \cdot \tilde{\Sigma}^{-1} \mathbf{1}}.$$

Alternatively, the target or benchmark estimate applied in the shrinkage of the sample estimates can be the parameter estimate when a certain structure is imposed on the return-generating process, e.g. by assuming that returns are generated by a factor model such as the Single-Index model or the Fama-French 3-factor model that we study in Chapter 11, or the assumption that all pairwise return correlations are identical. Given a choice of the shrinkage target, the best weighting of the sample variance-covariance matrix and the shrinkage target still has to be determined.

We refer the interested reader to [Jobson and Korkie \(1980\)](#) and [Michaud \(1989\)](#) for early papers discussing the estimation problem and to [Ledoit and Wolf \(2004, 2017\)](#) for specific examples of shrinkage methods in mean-variance analysis. See also [MacKinlay and Pastor \(2000\)](#), [Garlappi, Uppal, and Wang \(2007\)](#), [Kan and Zhou \(2007\)](#), and [Tu and Zhou \(2011\)](#).

Here is a small example, inspired by Frankfurter, Phillips, and Seagle (1971), illustrating the estimation issue. Consider three risky assets labeled X, Y, and Z whose monthly returns are jointly normally distributed with the moments listed in the upper left panel of Table 7.4. The returns are identically distributed in all months. The annualized expected returns (multiplying by 12, ignoring compounding) are 15%, 14%, and 12%, respectively, whereas the annualized standard deviations (multiplying by $\sqrt{12}$) are 40%, 40%, and 30%. The pairwise return correlations are all 0.6. In addition to the three risky assets, the investor has access to a riskfree asset with a monthly return of 0.125% corresponding to an annualized riskfree rate of 1.5%. These parameters are the *true* parameters of the distribution. They are not known to the investor who has to estimate the parameters from a time series of observations drawn from the distribution.

A number of time series of return observations have been simulated by drawing from the true joint distribution. Each observation is obtained by computing

$$r_X = \mu_X + \sigma_X \varepsilon_1, \quad (7.59)$$

$$r_Y = \mu_Y + \sigma_Y \left(\rho_{XY} \varepsilon_1 + \sqrt{1 - \rho_{XY}^2} \varepsilon_2 \right), \quad (7.60)$$

$$r_Z = \mu_Z + \sigma_Z \left(\rho_{XZ} \varepsilon_1 + \hat{\rho}_{YZ} \varepsilon_2 + \sqrt{1 - \rho_{XZ}^2 - \hat{\rho}_{YZ}^2} \varepsilon_3 \right), \quad (7.61)$$

where μ_X and σ_X are the expectation and standard deviation of the return on X, and μ_Y , σ_Y and μ_Z , σ_Z are the corresponding quantities for Y and Z. The parameters ρ_{XY} , ρ_{XZ} , and ρ_{YZ} are the pairwise return correlations, and $\hat{\rho}_{YZ} = (\rho_{YZ} - \rho_{XY}\rho_{XZ})/\sqrt{1 - \rho_{XY}^2}$. Furthermore, $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are independent draws from a standard normal distribution. Such numbers can be drawn in Excel using the formula `NORMSINV(RAND())`.³ Each time series consists of 60 observations, corresponding to a 5-year period with monthly observations. Based on each time series, the moments are then estimated by standard methods so that the mean is estimated by the average, etc., as explained in Section 3.7.2.

Three sets of estimated parameters based on three different simulated time series have been selected for the following illustrations. The three parameter sets are listed in the upper right panel and the lower panel of Table 7.4. For the true parameter set and each of the three estimated parameter sets, Table 7.5 shows the composition of the minimum-variance portfolio, the tangency portfolio, and the optimal portfolios for investors with a risk aversion coefficient γ equal to 1, 4, or 8. The optimal portfolio is computed from (7.55) using the estimated expected return and variance of the tangency portfolio which again is computed using the estimated moments of the individual assets. The position in the riskfree asset (rows labeled rf) is computed residually so that the portfolio weights sum up to one. For each portfolio, the true expected return and the true standard deviation (using the true moments) are also shown. For the three sets of estimates, the estimated expected return and standard deviation are shown in italics.

Applying the true moments, the minimum-variance portfolio is tilted towards the relatively low-risk asset Z, whereas X and Y have identical weights because of their identical standard deviations and the symmetric correlation structure. The tangency portfolio consists of roughly 50% in Z, 30% in X, and 20% in Y, reflecting the ranking in terms of Sharpe ratios.

Table 7.4 shows that for all three samples the estimated standard deviations and correlations are relatively close to their true values. The expected returns are harder to estimate

³While this is sufficiently precise for our illustrative purpose, an alternative approach that matches the standard normal distribution better should be used when precision is important.

	Expect	Std dev	Correlations			Expect	Std dev	Correlations		
<i>True moments</i>										
X	1.25	11.547	1	0.6	0.6	1.048	12.420	1.000	0.495	0.643
Y	1.167	11.547	0.6	1	0.6	0.158	10.579	0.495	1.000	0.505
Z	1	8.660	0.6	0.6	1	0.701	9.231	0.643	0.505	1.000
<i>Sample 2 estimates</i>										
X	-0.407	12.114	1.000	0.677	0.698	0.486	10.804	1.000	0.584	0.640
Y	-0.013	11.760	0.677	1.000	0.742	1.171	12.081	0.584	1.000	0.711
Z	0.995	10.314	0.698	0.742	1.000	2.212	8.631	0.640	0.711	1.000

Table 7.4: True and estimated moments in simulation experiment.

The table shows four sets of expected returns, standard deviations, and correlation for three assets labelled X, Y, and Z. The upper left corner shows the assumed true parameters, while the other three parameter sets are estimates based on time series of 60 simulated returns drawn from the assumed true distribution. The expectations and standard deviations are shown in percent per month.

	Min	Tang	$\gamma = 1$	$\gamma = 4$	$\gamma = 8$		Min	Tang	$\gamma = 1$	$\gamma = 4$	$\gamma = 8$
<i>True moments</i>											
X	0.122	0.312	0.405	0.101	0.051	0.018	1.112	0.597	0.149	0.075	
Y	0.122	0.192	0.249	0.062	0.031	0.360	-0.968	-0.520	-0.130	-0.065	
Z	0.757	0.496	0.644	0.161	0.080	0.622	0.856	0.460	0.115	0.057	
rf	0.000	0.000	-0.297	0.676	0.838	0.000	0.000	0.463	0.866	0.933	
μ	1.051	1.110	1.403	0.444	0.285	1.064	1.117	0.658	0.258	0.192	
est						0.512	1.612	0.924	0.325	0.225	
σ	8.447	8.713	11.304	2.826	1.413	8.693	13.689	7.352	1.838	0.919	
est						8.490	16.641	8.938	2.234	1.117	
<i>Sample 2 estimates</i>											
X	0.185	-3.210	-1.379	-0.345	-0.172	0.217	-0.666	-1.643	-0.411	-0.205	
Y	0.179	-2.098	-0.901	-0.225	-0.113	-0.046	-0.423	-1.044	-0.261	-0.130	
Z	0.635	6.308	2.710	0.678	0.339	0.828	2.088	5.155	1.289	0.644	
rf	0.000	0.000	0.570	0.893	0.946	0.000	0.000	-1.468	0.383	0.691	
μ	1.076	-0.152	0.006	0.095	0.110	1.047	0.763	1.700	0.519	0.322	
est	0.554	7.608	3.340	0.929	0.527	1.885	3.801	9.198	2.393	1.259	
σ	8.501	44.772	19.238	4.810	2.405	8.574	13.501	33.323	8.331	4.165	
est	9.994	41.730	17.931	4.483	2.241	8.443	12.204	30.121	7.530	3.765	
<i>Sample 3 estimates</i>											

Table 7.5: True vs. estimated mean-variance portfolios.

The table shows the four assets' weights in the minimum-variance portfolio, the tangency portfolio, and the optimal portfolios chosen by investors with risk aversion levels of 1, 4, and 8, respectively. These portfolio weights are shown both in the case where the true moments are used and for three sets of estimates based on samples with 60 simulated observations each. For each portfolio, the true expected return and the true standard deviation are also shown. For the three sets of estimates, the estimated expected return and standard deviation are shown in italics. Expected returns and standard deviations are in percent.

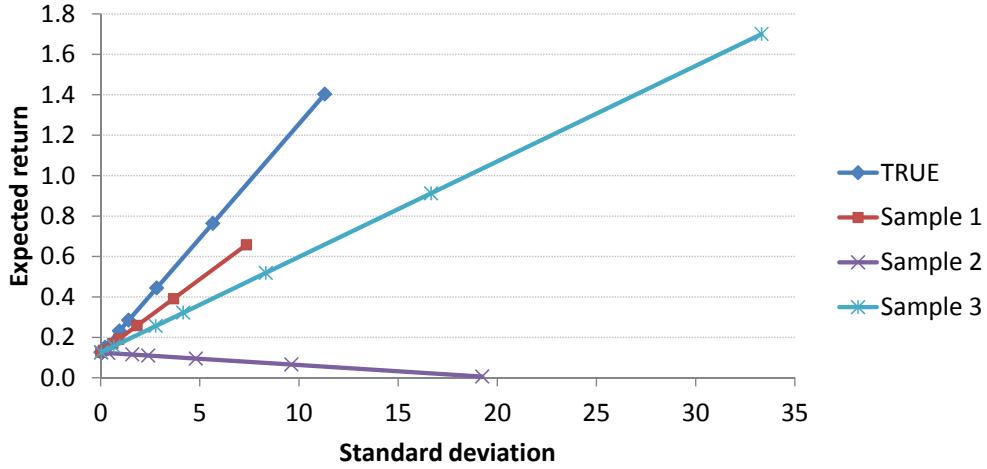


Figure 7.7: Portfolio selections for different parameter sets.

The dark blue line shows the combinations of expected return and standard deviation on the portfolio optimally chosen by investors with the mean-variance trade-off in (7.54) assuming they know the true parameters of the return distribution. The other lines show the combinations such investors would end up with if they based their portfolio decisions on a given estimated parameter set.

precisely, and their values are central to the output from the mean-variance analysis. In sample 1 all expected returns are under-estimated, particularly that of asset Y. As shown in Table 7.5, this leads to a quite extreme tangency portfolio with large long positions in X and Z and a large short position in Y. The true expected returns of the tangency portfolio and the optimal portfolios shown for sample 1 are significantly lower than the expected returns based on the estimates, whereas the true standard deviations are only somewhat smaller. As illustrated by the red line in Figure 7.7, the true risk-return trade-off (Sharpe ratio) on the estimated tangency portfolio—and therefore the portfolios chosen by mean-variance optimizing investors—is not far from that based on the true parameters, but an investor with a given risk aversion coefficient obtains a lower expected return and standard deviation than he thinks.

In sample 2 the estimated expected returns are negative for both assets X and Y, whereas for asset Z the estimate is close to the true value. Asset Z seems to have the highest expected return and the lowest standard deviation and therefore Z looks much more attractive than X and Y. Both the estimated tangency portfolio and the optimal portfolios have a high positive weight on Z and negative weights on X and Y. These portfolios are estimated to have both relatively large expected returns and standard deviations, but the true expected returns are much lower because of the estimation error. In fact, because the true expected return of the estimated tangency portfolio is lower than the riskfree rate, the true expected returns of the optimal portfolios increase for higher levels of the risk aversion coefficient as that involves less short positions in X and Y which have the highest true expected returns. The risk-return tradeoff is thus downward-sloping as illustrated by the purple line in Figure 7.7. In particular, relatively risk-tolerant investors get a risk-return combination very different from what they expected based on their estimates.

Finally, in sample 3 asset Z and the estimated tangency portfolio seem very attractive so the optimal portfolios involve a large positive position in Z together with short positions

in X and Y (and also in the riskfree asset when the risk aversion is sufficiently small). This results in very risky positions with a true expected return far below the estimated expected return. These combinations are on the light-blue line in Figure 7.7.

7.4.2 The number of inputs

The number of inputs to the mean-variance analysis increases dramatically with the number of assets. With N assets, you need estimates of the N expected returns and the N variances of the assets, plus all the pairwise covariances. With N assets, there are $N(N - 1)/2$ covariances. For example, with 20 risky assets you need 20 estimates of expected returns, 20 estimates of return variances, and $20 \times 19/2 = 190$ estimates of covariances, that is a total of 230 inputs. With 100 risky assets, the number grows to 5,150 inputs. And with 500 risky assets, there are 125,750 inputs!

An obvious way to reduce the number of input variables is to reduce the number of assets considered. But of course you do not want to disregard many potentially attractive assets. One solution is to perform the mean-variance analysis for a low number, say 5–15, of *asset classes* instead of a large number of individual assets. Each asset class is represented by some index of assets in that class. For example, the asset classes could be domestic stocks, domestic government bonds, domestic corporate bonds, foreign stocks, foreign government bonds, foreign corporate bonds, and real estate. Each of these classes could be split into a few subclasses. For example, domestic stocks may be split into subclasses according to industry sectors, market capitalization, book-to-market values, or recent performance. Foreign stocks may be split by geographical regions or into developed and emerging markets.

With a low number of asset classes, the number of inputs is limited. Furthermore, it is easier to provide reliable forecasts of expected returns on an asset class level than on the level of individual assets. The returns of asset classes are less volatile over time than the returns of individual assets, which reduces the estimation risk when using historical returns. In addition, some individual assets (at least some stocks) are from time to time severely affected by mergers, technological breakthroughs, patents, and other events that may dramatically change their risk-return profile so that historical returns might not be representative of future returns. Asset classes are less sensitive to such events. Even when considering asset classes, investors do often not fully rely on historical estimates, especially not for their expected returns, but adjust these based, for example, on an assessment of the macroeconomic or industry-specific outlook.

Another way to reduce the number of input variables is to add more structure to the model in the form of additional assumptions. More specifically, if all the common variation in the returns of the assets are driven by a low number of factors, the number of inputs can be reduced substantially. This is the idea of the so-called *factor models*, see Chapter 11.

7.4.3 Portfolio constraints

In the mean-variance analysis presented above, we did not impose any constraints on the individual portfolio weights. For certain input variables, the optimal portfolios can easily involve very large or highly negative positions in some assets. However, some investors face portfolio constraints. For example, some investors are not allowed to—or have the policy not to—take short positions. Others may avoid having an unbalanced portfolio with a single or a few very dominating assets.

If the inputs are based on historical estimates, the tangency portfolio of risky assets will often involve huge positions in some assets that have happened to perform well in the

estimation period, but it may be unlikely that they will perform as well in the future. It may also involve large short positions in other assets that did poorly in the estimation period or just happen to have a very low correlation with the assets that did well. One approach to avoiding such extreme positions is to put upper and lower bounds on each portfolio weight.

The basic idea of the mean-variance analysis still holds if portfolio constraints are added, but analytical derivations as those presented in the preceding sections are very complicated. Some analytical results for the case where the portfolio weights are restricted to being non-negative (short-selling not possible) can be found in Elton, Gruber, and Padberg (1976), Alexander (1993), and Best and Grauer (1991). Alexander, Baptista, and Yan (2007) study the case of value-at-risk type constraints.

In any case, such constrained problems can be solved with various computational software tools. In Excel, the problems are tackled with the so-called **Solver** which in the current version of Excel is available from the Data tab.⁴ When selecting the **Solver**, a window pops up in which you can enter a target cell, an objective (max or min), the cells to vary, and various constraints. A non-negativity constraint can be included simply by ticking a box. You can also choose between different optimization algorithms, but the default suggestion seems to work well for mean-variance problems.

To generate the *constrained mean-variance frontier* we cannot just identify two frontier portfolios and then combine them in various ways as such combinations can easily violate the constraints. We have to solve a number of constrained optimization problems. For a given level of the expected portfolio return, minimize the variance (in the target cell) by varying the portfolio weights (in an array of cells) under the relevant constraints. **Solver** delivers the solution in form of the portfolio weights and the associated variance. You can copy these values to a different place in your Excel sheet and then solve the problem for a new level of the expected return. Doing this for a range of (say 15-20) relevant values of the expected return gives you various points on the mean-variance frontier and the corresponding portfolio weights. Note that when portfolio weights must be non-negative, the largest possible expected return among all portfolios is obtained by a full investment in the asset with the largest expected return. Conversely, the smallest possible expected return among all portfolios is obtained by a full investment in the asset with the smallest expected return. This gives natural bounds on the range of expected returns to consider when generating the frontier.

We can also find the *constrained tangency portfolio* by using the solver. In this case the target cell must contain the Sharpe ratio of the portfolio and the objective is to maximize this value. Similarly, we could determine the *constrained minimum-variance portfolio*. Without portfolio constraints, the efficient frontier of risky assets can be generated by considering a wide range of combinations of the minimum-variance portfolio and the tangency portfolio—or any other two efficient portfolios. When moving along the frontier the portfolio weight of any individual asset is then going to change linearly. This procedure does not work with portfolio constraints because if you move far enough up or down the frontier, the combination of the minimum-variance portfolio and the tangency portfolio will start to violate the constraints.

Figure 7.8 presents an example illustrating the effects on the mean-variance frontier of forbidding short sales of the risky assets. The picture is constructed using data on 11 risky asset classes corresponding to the red squares. The black-dotted hyperbola and the black

⁴Before the first time use, you have to activate the **Solver**. In the current version of Excel do the following: From the File menu, choose Options. In the dialog box, select Add-Ins to the left, and click Go at the bottom of the window. Tick the Solver Add-In box and click OK.

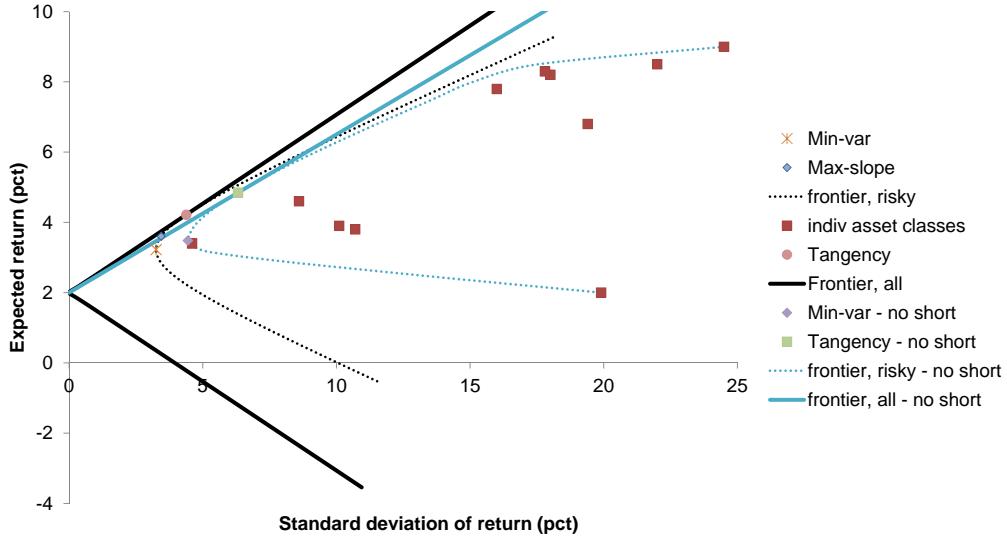


Figure 7.8: Efficient frontiers with and without short sales.

Computations are based on 11 risky asset classes corresponding to the points labeled by the red squares as well as a riskfree rate of 2%.

wedge are the mean-variance frontiers with unconstrained portfolio weights, whereas the curved blue-dotted line and the straight blue line are the frontiers with no short selling of the risky assets. Of course, the constrained frontiers are located to the right of the unconstrained frontiers: when you minimize the variance for a given expected return, you can obtain a lower variance if you minimize over all portfolios than if you minimize only over the portfolios with non-negative weights.

With constraints the efficient frontier of risky assets is not of the nice hyperbolic shape as in the unconstrained case. Even with short-selling constraints the efficient frontier of all assets includes the straight line between the point corresponding to the constrained tangency portfolio and the point corresponding to the riskfree asset. If borrowing is possible (in other words, the riskfree asset can be shorted), we can also obtain the points on the tangent line that lies above to the right of the tangency point, as these points correspond to leveraged investments in the tangency portfolio. In the unconstrained case, the efficient frontier of all assets include the (less interesting) downward-sloping straight line as shown for example in Panel (a) of Figure 7.3. But since this line involves shorting the tangency portfolio, this part is infeasible with short-selling constraints.

Our stylized and analytically tractable mean-variance model in Section 7.2 assumes that the same interest rate applies to borrowing and lending. In reality the borrowing rate of an investor typically exceeds the lending rate. Figure 7.9 illustrates this case. Here the lending rate is $r_l = 0$, and the borrowing rate over the period considered is $r_b = 2\%$. The black hyperbola is the efficient frontier of risky assets (here derived without constraints). The square and the triangle indicate the tangency portfolio computed using the lending rate and the borrowing rate, respectively. The tangency lines are also shown, but the dotted parts of the lines are not feasible. For example, the dotted part of the blue line (the steepest line) corresponds to portfolios involving borrowing at 0%, which is not assumed possible in this example. As always the overall efficient frontier of all assets

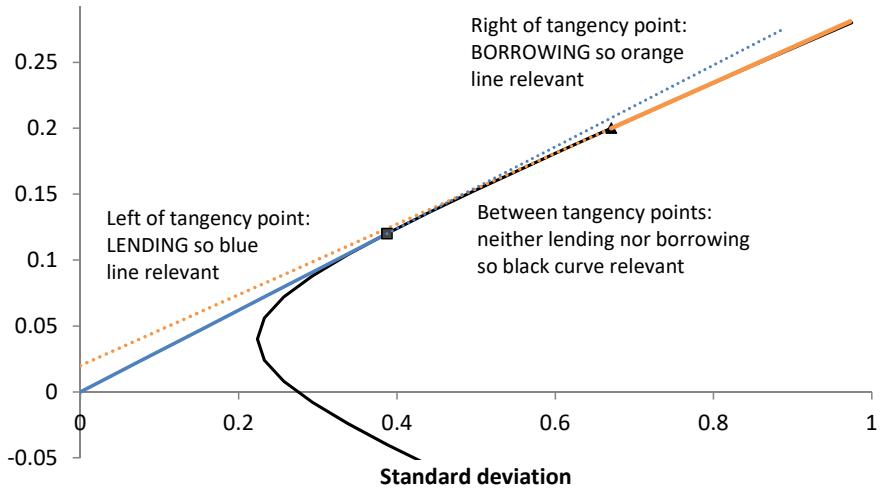


Figure 7.9: Frontier with different borrowing and lending rates.

The blue line is the efficient frontier and the square is the tangency point with a riskfree rate of 0%. The orange line is the efficient frontier and the triangle the tangency point with a riskfree rate of 2%. With a borrowing rate of 2% and a lending rate of 0%, the efficient frontier consists of the solid part of the blue line, the curve between the two tangency points, and the solid part of the orange line.

consists of the points with the lowest standard deviation for some expected return. In this case, the efficient frontier has three parts:

1. the piece of the straight line representing the efficient frontier of all assets with a riskfree rate of r_l that goes from $(0, r_l)$ to the tangency point (the square in the figure),
2. the piece of the efficient frontier of risky assets only that lies between the two tangency points (this is a concave curve although it may appear to be a straight line in the figure),
3. the piece of the straight line representing the efficient frontier of all assets with a riskfree rate of r_b that goes from the tangency point (the triangle in the figure) and up to the right.

In the case where borrowing is not possible at any interest rate, the third part is not present and the efficient frontier only consists of a straight line up to the tangency point associated with r_l and then the curved frontier of risky assets to the right of the tangency point. This implies that a borrowing-constrained investor with low enough risk aversion is not going to invest in the tangency portfolio of risky assets, but instead in some portfolio with both higher standard deviation and higher expected return than the tangency portfolio, and this portfolio is not necessarily well diversified.

7.5 Theoretical foundation

The key assumption of mean-variance analysis is that investors make their investment decision based only on the expectation and the variance (or, equivalently, the standard deviation) of the return on the investment over the period in question. This assumption is satisfied if the returns on risky assets are normally distributed since a normal distribution is fully characterized by its mean and variance, and portfolio returns are then also normally distributed. The normal distribution assumption is not perfect, but not a too bad approximation of reality at least if you focus on stock returns over a month, a quar-

ter, or a year; see the discussion in Section 6.5. In this case, we still have to quantify the mean-variance trade-off of the investor. In particular, we discuss below when the tractable mean-variance criterion (7.54) can be justified, where “tractable” refers to the fact that we derived an explicit solution for the optimal portfolio in this case. First, we need an introduction to the concept of a utility function.

7.5.1 Utility functions

In economics, the preferences of a decision-maker are typically represented by a utility function.⁵ For our one-period investment problem let W_0 denote the wealth of the investor at the beginning of the period and assume that all of this wealth is invested. The wealth at the end of the period is then

$$W = W_0(1 + r), \quad (7.62)$$

where r is the rate of return on the investment over the period. Obviously, the return and thus the end-of-period wealth depend on the portfolio chosen at the beginning of the period, and both the return and the end-of-period wealth are random variables—unless the entire wealth is invested in a riskfree asset. A utility function is then a function $u(W)$ assigning values to each possible level of end-of-period wealth. The objective of the investor is to maximize the expected utility, $E[u(W)]$, over all possible portfolios.

Utility functions are normally assumed to be

1. increasing (so that $u'(W) > 0$, assuming u is differentiable) which means that the investor is greedy—she wants as much wealth as possible; and
2. concave (so that $u''(W) < 0$) which means both (a) marginal utility $u'(W)$ is decreasing in wealth, i.e., the investor appreciates an extra dollar more when poor than when rich, and (b) the investor is risk averse—she rejects any risky gamble where the expected profit is zero or negative.

Figure 7.10 gives an example of such a utility function, namely the square root function $u(W) = \sqrt{W}$, which is certainly increasing and concave.

Consider an investor whose preferences are represented by a square root utility function. Currently she is holding her entire wealth of \$250 in cash. She is offered a gamble or risky investment where she has to invest the \$250. In return she receives either \$100 or \$400, and the two outcomes are equally likely. Note that the expected wealth of the gamble equals the investment, $E[W] = 0.5 \times \$100 + 0.5 \times \$400 = \$250$, so in that sense it is a *fair gamble*. In financial terms, the expected return is zero. The corresponding utility levels are

$$u(\$100) = \sqrt{100} = 10, \quad u(\$400) = \sqrt{400} = 20,$$

so the expected utility is

$$E[u(W)] = 0.5 \times 10 + 0.5 \times 20 = 15.$$

If she rejects the gamble and keeps her \$250 for sure, her utility (which is also the expected utility as there is no uncertainty) is simply

$$u(\$250) = \sqrt{250} \approx 15.81.$$

⁵The concept of a utility function dates back at least to the Swiss mathematician Daniel Bernoulli in 1738 (see English translation in Bernoulli (1954)) and was put on a firm formal setting by von Neumann and Morgenstern (1944). They showed that if the preferences of the decision-maker satisfy a few—apparently reasonable but not undisputed—“behavioral axioms”, then the preferences can indeed be represented by a utility function and the optimal decision maximizes the expected value of the utility function.

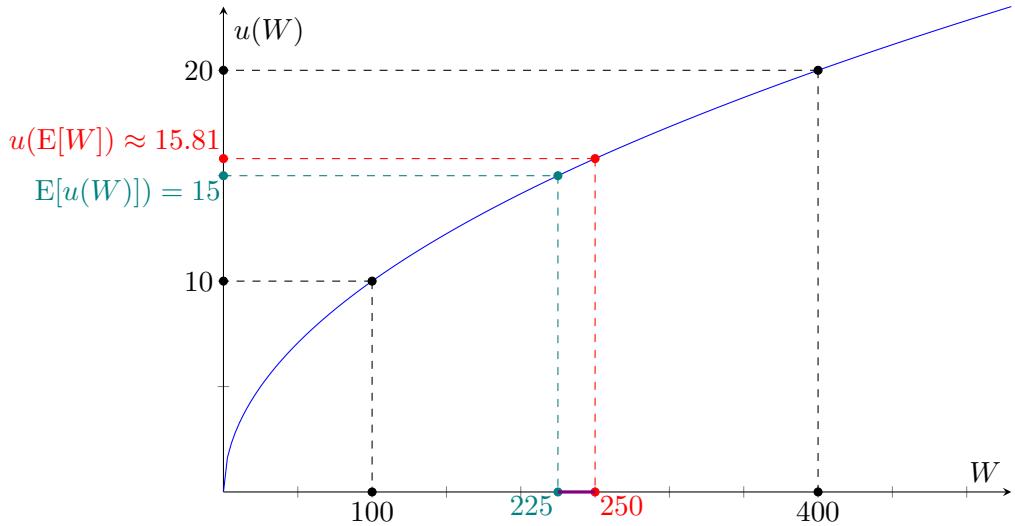


Figure 7.10: A utility function.

The square root function $u(W) = \sqrt{W}$ is a utility function. The dashed lines and the associated numbers are explained in the text.

Evidently, she prefers the safe amount to the fair gamble, which reflects that she is risk averse. In fact, a safe amount of \$225 would give her the same utility as the gamble since $\sqrt{225} = 15 = E[u(W)]$. The \$225 is said to be the *certainty equivalent* of the gamble and is written as $CE = \$225$. She is willing to give up the difference $E[W] - CE = \$250 - \$225 = \$25$ in expected wealth to avoid the risk. This is 10% of her current wealth. Both the \$25 and the 10% are measures of the investor's *degree* of risk aversion. If the investor had been more risk averse, she would have been willing to give up a large amount and thus a larger share of her current wealth to avoid the gamble.

The exact numerical values of the utility function are not informative. We are not claiming that a certain level of wealth can be associated with some unique level of happiness. A utility function ranks alternatives. Saying that $u(\$400) = 20$ and $u(\$100) = 10$ as in the above example does not mean that the investor is twice as happy with a wealth of \$400 as with a wealth of \$100. Given a utility function $u(W)$, any function of the form

$$v(W) = A \times u(W) + B,$$

where $A > 0$ and B is any real number, positive or negative, is also a utility function that leads to exactly the same optimal decision as the original utility function u .

Two frequently used measures of the degree of risk aversion are the Absolute Risk Aversion, ARA, and the Relative Risk Aversion, RRA, defined by

$$\text{ARA}(W) = -\frac{u''(W)}{u'(W)}, \quad \text{RRA}(W) = -\frac{Wu''(W)}{u'(W)} \quad (7.63)$$

and sometimes referred to as the Arrow-Pratt risk aversion measures in honor of Kenneth Arrow and John Pratt whose research in the 1960's made great contributions to our understanding of risk. Both the ARA and RRA are unaffected by the above-mentioned scaling by positive constants or addition or subtraction of constants. The concepts of absolute and relative risk aversion are discussed and illustrated below.

7.5.2 Quadratic utility functions

Suppose an investor has a *quadratic utility function*,

$$u(W) = a + bW - cW^2, \quad (7.64)$$

where a , b , and c are constants. Then the expected utility is

$$\mathbb{E}[u(W)] = a + b\mathbb{E}[W] - c\mathbb{E}[W^2] = a + b\mathbb{E}[W] - c\left(\text{Var}[W] + (\mathbb{E}[W])^2\right),$$

where the last equality is due to Eq. (3.15). As the expected utility only depends on the expectation and variance of the wealth, the mean-variance analysis is appropriate for investors with quadratic utility functions, no matter how the returns are distributed.

Unfortunately, quadratic utility is a poor representation of investor preferences. Since $u'(W) = b - 2cW$, the utility function is only increasing up to a wealth level of $W = b/(2c)$, after which it starts to decrease again. This conflicts with the very reasonable assumption of greediness. It can be shown that quadratic utility has other unrealistic properties.

7.5.3 A utility function supporting the tractable mean-variance criterion

Suppose now that returns and thus end-of-period wealth are normally distributed so that preferences only depend on the mean and the variance. The end-of-period wealth is $W = W_0(1+r)$ where r is the rate of return on the portfolio chosen by the investor. If all assets have normally distributed returns, this is also true for any portfolio. Note that

$$r \sim N(\mu, \sigma^2) \quad \Rightarrow \quad W \sim N\left(W_0(1+\mu), W_0^2\sigma^2\right). \quad (7.65)$$

To find the optimal portfolio among the mean-variance efficient portfolios, a criterion like (7.54) is often assumed. Such a mean-variance criterion has some intuitive appeal, but is it consistent with a reasonable utility function?

Suppose that the investor has a *negative exponential utility function* of wealth,

$$u(W) = -e^{-kW}. \quad (7.66)$$

Since

$$u'(W) = ke^{-kW}, \quad u''(W) = -k^2 e^{-kW}$$

we require $k > 0$ to represent a greedy and risk-averse investor. The risk aversion measures defined in (7.63) become

$$\text{ARA}(W) = k, \quad \text{RRA}(W) = kW,$$

so in particular the absolute risk aversion is independent of the investor's level of wealth. Hence the negative exponential utility function is often called a CARA utility function as it exhibits a constant absolute risk aversion.

It can be shown (see Appendix A) that

$$X \sim N(m, s^2) \quad \Rightarrow \quad \mathbb{E}\left[e^{-kX}\right] = e^{-km + \frac{1}{2}k^2s^2}. \quad (7.67)$$

In our case, this implies that

$$\begin{aligned} \mathbb{E}[u(W)] &= \mathbb{E}\left[-e^{-kW}\right] = -\mathbb{E}\left[e^{-kW}\right] \\ &= -e^{-kW_0(1+\mu)+\frac{1}{2}k^2W_0^2\sigma^2} = -e^{-kW_0\left(1+\mu-\frac{1}{2}kW_0\sigma^2\right)}. \end{aligned} \quad (7.68)$$

Because $-e^{-kW_0(1+x)}$ is increasing in x , the portfolio maximizing $\mathbb{E}[u(W)]$ will also maximize $\mu - \frac{1}{2}kW_0\sigma^2$. This shows that the mean-variance criterion (7.54) is consistent with the investor having a negative exponential utility function. The constant γ equals the product of the investor's absolute risk aversion k and initial wealth W_0 , which is exactly the initial relative risk aversion. An investor with a negative exponential utility function should therefore invest the fraction

$$w^* = \frac{\mu_{\tan} - r_f}{kW_0\sigma_{\tan}^2} \quad (7.69)$$

of wealth in the tangency portfolio and the rest in the riskfree asset.

What does a constant absolute risk aversion mean? Consider an investor with current wealth W and a negative exponential utility function. The investor's certainty equivalent CE for a fair gamble in which she wins x and loses x with equal probabilities is defined by

$$u(\text{CE}) = \frac{1}{2}u(W+x) + \frac{1}{2}u(W-x), \quad (7.70)$$

that is,

$$e^{-k\times\text{CE}} = \frac{1}{2}e^{-k(W+x)} + \frac{1}{2}e^{-k(W-x)}.$$

Solving for CE, we find

$$\begin{aligned} \text{CE} &= -\frac{1}{k} \ln \left(\frac{1}{2}e^{-k(W+x)} + \frac{1}{2}e^{-k(W-x)} \right) = -\frac{1}{k} \ln \left(\frac{1}{2}e^{-kW} [e^{-kx} + e^{kx}] \right) \\ &= -\frac{1}{k} \left(-\ln 2 - kW + \ln [e^{-kx} + e^{kx}] \right) = W + \frac{\ln 2}{k} - \frac{1}{k} \ln [e^{-kx} + e^{kx}]. \end{aligned}$$

Hence, the investor is willing to give up at most

$$W - \text{CE} = \frac{1}{k} \ln [e^{-kx} + e^{kx}] - \frac{\ln 2}{k}$$

to avoid the gamble. Note that this is independent of wealth, which reflects the constant absolute risk aversion. It can be shown that the larger the value of k , the larger the difference $W - \text{CE}$, in line with the interpretation of k as a measure of the degree of absolute risk aversion.

Normally, you would expect a very wealthy individual to be willing to give up less to avoid a fair gamble of, say, 100 dollars than a very poor individual. This suggests that the absolute risk aversion should be a decreasing function of the wealth of the individual. However, with the negative exponential utility function this absolute risk aversion is assumed to be a constant k , which is a major drawback of this utility specification. On the other hand, the negative exponential utility function is mathematically very tractable in combination with normally distributed returns.

7.5.4 Better utility functions

A frequently used utility function is

$$u(W) = \frac{1}{1-\gamma} W^{1-\gamma}, \quad \gamma > 0, \quad \gamma \neq 1. \quad (7.71)$$

Since $u'(W) = W^{-\gamma}$ and $u''(W) = -\gamma W^{-\gamma-1}$, the absolute and relative risk aversion measures in (7.63) are

$$\text{ARA}(W) = \frac{\gamma}{W}, \quad \text{RRA}(W) = \gamma.$$

Hence, utility functions like (7.71) are called CRRA utility as they exhibit constant relative risk aversion, and γ is referred to as the relative risk aversion coefficient. The square-root function we considered earlier corresponds to $\gamma = 0.5$ and thus belongs to this class of utility functions (recall that multiplying by $1/(1-0.5) = 2$ or not is unimportant). Note that the logarithmic utility function,

$$u(W) = \ln W, \quad (7.72)$$

also has a constant relative risk aversion, namely 1.⁶

What does a constant relative risk aversion mean? Let us again consider a fair gamble of $\pm x$ around some current wealth level W . Substituting (7.71) into (7.70), we find that the certainty equivalent is now

$$\text{CE} = \left(\frac{1}{2}\right)^{1/(1-\gamma)} \left((W+x)^{1-\gamma} + (W-x)^{1-\gamma}\right)^{1/(1-\gamma)}, \quad (7.73)$$

and the difference $W - \text{CE}$ is now decreasing in W as we expect it to be. Suppose instead the investor is gambling about some fraction α of her current wealth so $x = \alpha W$. The corresponding certainty equivalent is then

$$\text{CE} = W \left(\frac{1}{2}\right)^{1/(1-\gamma)} \left((1+\alpha)^{1-\gamma} + (1-\alpha)^{1-\gamma}\right)^{1/(1-\gamma)}, \quad (7.74)$$

so the fraction of her current wealth she would give up to avoid the gamble is

$$\frac{W - \text{CE}}{W} = 1 - \left(\frac{1}{2}\right)^{1/(1-\gamma)} \left((1+\alpha)^{1-\gamma} + (1-\alpha)^{1-\gamma}\right)^{1/(1-\gamma)}, \quad (7.75)$$

which is independent of her level of wealth. So when the gamble is relative to her wealth, her relative wealth sacrifice is independent of her current wealth. This is the meaning of a constant relative risk aversion. Of course, the fraction $(W - \text{CE})/W$ is increasing in the

⁶Here is the mathematical background. Except for a constant, the utility function

$$u(W) = \frac{W^{1-\gamma} - 1}{1-\gamma}$$

is identical to the function specified in (7.71). The two utility functions are therefore equivalent in the sense that they generate the same ranking of alternatives and, hence, the same optimal choices. The advantage in using the latter definition is that this function has a well-defined limit as $\gamma \rightarrow 1$. From l'Hôpital's rule we have that

$$\lim_{\gamma \rightarrow 1} \frac{W^{1-\gamma} - 1}{1-\gamma} = \lim_{\gamma \rightarrow 1} \frac{-W^{1-\gamma} \ln W}{-1} = \ln W$$

so in that sense $\ln W$ corresponds to $\gamma = 1$.

Gamble of $x = \pm 0.5$								
			CARA $k = 2$			CRRA $\gamma = 2$		
W	$W + x$	$W - x$	CE	will give up		CE	will give up	
1	1.5	0.5	0.7831	0.2169	21.69%	0.75	0.2500	25.00%
10	10.5	9.5	9.7831	0.2169	2.17%	9.975	0.0250	0.25%
100	100.5	99.5	99.7831	0.2169	0.22%	99.75	0.0025	0.0025%
Gamble of $x = \pm 0.05W$								
			CARA $k = 2$			CRRA $\gamma = 2$		
W	$W + x$	$W - x$	CE	will give up		CE	will give up	
1	1.05	0.95	0.9975	0.0025	0.25%	0.9975	0.0025	0.25%
10	10.5	9.5	9.7831	0.2169	2.17%	9.975	0.025	0.25%
100	105	95	95.3466	4.6534	4.65%	99.75	0.25	0.25%

Table 7.6: Comparing CARA and CRRA.

The table shows how much an investor with CARA utility or CRRA utility is willing to sacrifice in order to avoid a fair gamble of either a fixed amount (top panel) or a fixed fraction (lower panel) of her current wealth.

$\gamma = \text{RRA}$	$\alpha = 1\%$	$\alpha = 10\%$	$\alpha = 50\%$
0.5	0.00%	0.25%	6.70%
1	0.01%	0.50%	13.40%
2	0.01%	1.00%	25.00%
5	0.02%	2.43%	40.72%
10	0.05%	4.42%	46.00%
20	0.10%	6.76%	48.14%
50	0.24%	8.72%	49.29%
100	0.43%	9.37%	49.65%

Table 7.7: CRRA utility.

The fraction of wealth a CRRA investor is willing to sacrifice to avoid a fair gamble of a fraction α of her current wealth.

coefficient γ and in the “size” α of the gamble.

Table 7.6 compares a CARA utility function with $k = 2$ and a CRRA utility function with $\gamma = 2$. The upper panel considers a gamble of a fixed amount of 0.5. With CARA utility, the investor sacrifices 0.2169 to avoid the gamble whether she has a wealth of 1, 10, or 100, whereas the amount sacrificed by an investor with CRRA utility decreases in wealth as we would expect. The lower panel considers a gamble of 5% of current wealth. Here the CRRA investor would sacrifice the same fraction of wealth to avoid the gamble independently of her wealth level. In contrast, the wealth share sacrificed by the CARA investor increases in wealth.

Focusing on CRRA utility, Table 7.7 shows the relative wealth sacrifice for various values of γ and α . For example, an individual with $\gamma = 5$ is willing to sacrifice 2.43% of the safe wealth in order to avoid a fair gamble of 10% of that wealth. Of course, even extremely risk-averse individuals will not sacrifice more than they can lose but in some cases it is pretty close. Looking at these numbers, it is hard to believe in γ -values outside, say, the interval [1,10].

Figure 7.11 illustrates the CRRA utility function for different value of γ . Some of

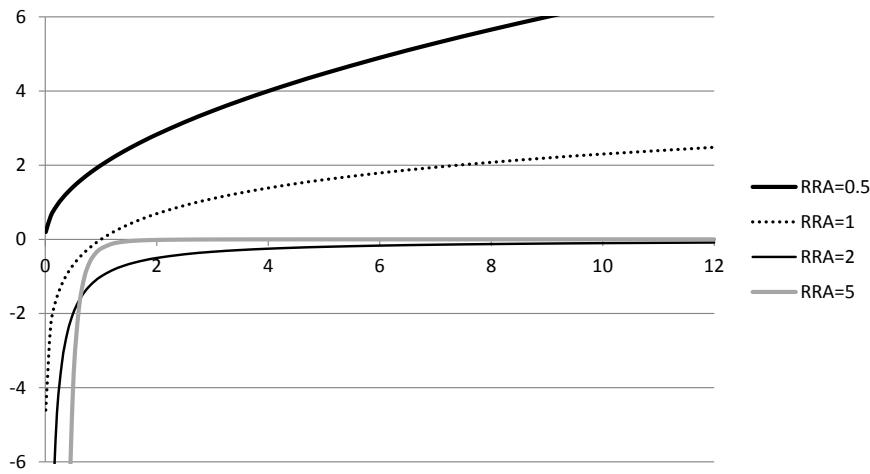


Figure 7.11: Illustration of CRRA utility functions.

The figure shows graphs of the CRRA utility functions for four different values of the relative risk aversion (RRA) parameter γ . For $\gamma = 1$, the function is $u(w) = \ln W$. For $\gamma \neq 1$, the function is $u(w) = W^{1-\gamma}/(1-\gamma)$.

the functions are negative and it may seem strange to assign a negative utility to some wealth. But again the precise numerical values of the utility function carry no meaning. Utility functions are only used for ranking different alternatives and finding the best of all alternatives. If you do not like a utility function taking negative values, feel free to add a huge positive constant.

A constant relative risk aversion seems much more reasonable than a constant absolute risk aversion, and various empirical studies of consumer and investor behavior also report some support for the CRRA function with γ -values in the range 1-5, cf., e.g., Meyer and Meyer (2005), Guiso and Sodini (2013), and the references therein. While not perfect, the CRRA function seems as a reasonable starting point for an analysis of optimal decisions.

So why are we not applying CRRA utility in the mean-variance setting? The CRRA utility function is only defined for positive wealth or, in other words, the utility of a negative wealth is $-\infty$. An investor with this utility function would never consider a portfolio that could lead to a negative wealth in the future. But the normality assumption of the mean-variance analysis implies that wealth can be negative, which is thus incompatible with CRRA utility. Or, more formally, with normally distributed returns, the only investment decision which does not lead to an expected utility of $-\infty$ is a full investment in the riskfree asset. If we instead assumed that log-returns were normally distributed, the value of any unlevered, long-only portfolio would always stay positive, so this could potentially be combined with CRRA utility. However, the problem with this assumption is that if the assets in the portfolio have normally distributed log-returns, then the log-return on the portfolio is not normally distributed, cf. Theorem 3.4. We return to the combination of CRRA utility and normally distributed log-returns in the next chapter.

7.5.5 Further critique of the mean-variance analysis

There are several other critical issues with the mean-variance approach to investment decisions:

1. Investors typically get utility from consumption at many points in time and not simply

- the wealth level at one particular date.
2. Even in the case where the investor only obtains utility from wealth at one date, she has the opportunity to change her portfolio over time, which she would normally do as new information arises (e.g., when stock prices and interest rates change) or simply because time passes. Investors live in a dynamic world and take decisions dynamically. Of course, the existence of transaction costs is a reason for not changing the portfolio too frequently, but if we are really worried about transaction costs we should explicitly model that imperfection; the analysis of such models is quite difficult, however.
 3. Consumption and investment decisions are generally not to be separated from each other. After all, investments are meant to generate future consumption.
 4. The financial investment decisions of an individual should be seen in connection with other factors that are essential for the welfare of the individual, such as the magnitude and uncertainty of labor income and her housing decisions (owning or renting; size and quality of house/apartment; mortgages).

Chapter 8 describes numerous models of multi-period investments taking the above aspects into account. As we shall see, the optimal investment decision is still quite closely related to that found in the current chapter so, in spite of all the critique, the mean-variance analysis remains useful.

7.6 Exercises

Exercise 7.1. Suppose you can invest in two assets. Asset 1 has an expected return of 6% and a standard deviation of 0.3. Asset 2 has an expected return of 9% and a standard deviation of 0.4.

- (a) Using Excel or another appropriate software tool, compute the combinations of standard deviation and expected returns that can be obtained by combining the two assets for the following different values of the correlation coefficient ρ : $-0.5, 0, 0.5$. Illustrate your findings graphically.
- (b) For each of the three correlations, what is the portfolio generating the minimum variance and the corresponding standard deviation and expected return?

In the following assume the two assets have a correlation of $\rho = 0$. Furthermore, assume you can also invest in a riskfree asset.

- (c) Determine the tangency portfolio of the two risky assets if the riskfree rate is 1%. What is the expected return and the standard deviation of the tangency portfolio? Illustrate the mean-variance frontier graphically.
- (d) Answer the same questions as above for a riskfree rate of 3%. Compare with the results for a riskfree rate of 1%.

Exercise 7.2. In a market with many assets, but no riskfree asset, the following portfolios, among others, exist:

Portfolio	Expected return	Standard deviation
P1	9%	21%
P2	5%	7%
P3	15%	36%
P4	12%	15%

- (a) Can we determine whether one or more of the portfolios lie on the efficient frontier? If so, which portfolios are efficient?
- (b) Can we determine whether one or more of the portfolios do *not* lie on the efficient frontier? If so, which portfolios are inefficient?

Exercise 7.3. Show Equation (7.8).

Exercise 7.4. Show that the minimum-variance portfolio has the property that (7.20) is satisfied.
Hint: Consider a portfolio consisting of a fraction w in the risky asset/portfolio with return r and a fraction $(1 - w)$ in the minimum-variance portfolio. Compute the variance of the return on this portfolio and realize that the variance has to be minimized for $w = 0$.

Exercise 7.5. Let r_{mv} denote the return on any mean-variance efficient portfolio of risky assets and let r denote another, not necessarily efficient, portfolio of risky assets with $E[r] = E[r_{\text{mv}}]$.

- (a) Show that $\text{Cov}[r_{\text{mv}}, r] = \text{Var}[r_{\text{mv}}]$.
- (b) Show that $\text{Corr}[r_{\text{mv}}, r] > 0$.

Exercise 7.6. Give a mathematical proof of Theorem 7.5. *Hint:* First show that the covariance between the return on the efficient portfolio with mean μ_1 and the return on the efficient portfolio with mean μ_2 is equal to $(C\mu_1\mu_2 - B[\mu_1 + \mu_2] + A)/D$.

Exercise 7.7. You have the following information about the return distribution of three stocks:

Stock	Exp. return (%)	Standard dev. (%)	Correlations		
			A	B	C
A	10	14	1	0.2	0.4
B	12	20	0.2	1	0.7
C	18	30	0.4	0.7	1

Also, the riskfree rate of return is 5%. Note that the questions below do not require the use of a computer, only your calculator, but feel free to do it in Excel as well.

- (a) Determine the variance-covariance matrix, $\underline{\Sigma}$.
- (b) Compute the expected return and standard deviation on a portfolio consisting of an equal weight in all three stocks.

The inverse of the variance-covariance matrix, $\underline{\Sigma}^{-1}$, is found to be (check this in Excel!):

$$\underline{\Sigma}^{-1} = \begin{pmatrix} 61.7 & 6.8 & -14.7 \\ 6.8 & 49.8 & -24.5 \\ -14.7 & -24.5 & 25.3 \end{pmatrix}$$

- (c) Find the (global) minimum-variance portfolio consisting of the three stocks.
- (d) Find the tangency portfolio of the three stocks.
- (e) Assume that you wish to hold a portfolio with an expected return of 18% and the lowest possible variance by combining the riskfree asset and A, B and C. Determine the weights in A, B, and C.
- (f) Assume that you wish to hold a portfolio with an expected return of 18% and the lowest possible variance, but that you do not have access to investing in a riskfree asset. Determine the weights in A, B and C.
- (g) Sketch the solutions to the preceding questions (b)-(f) in a (standard deviation, expected return)-diagram. You don't need to compute portfolio standard deviations for this question, merely indicate relative values.
- (h) Assume that the returns of all stocks are normally distributed. Compute the probability that the return is less than zero for (1) the minimum variance portfolio in question (c), and (2) the portfolio in question (b).
- (i) Assume that an investor at time 0 has invested \$1 mill. in the minimum-variance portfolio found in question (c). Compute the investor's maximum loss on a yearly basis with, respectively, 95% and 99% probability under the normal distribution assumption.

Exercise 7.8. You have the following information about the returns on three risky assets over the next year:

Asset	Expected return	Standard deviation	Correlations		
			X	Y	Z
X	20%	40%	1	0.4	0
Y	24%	50%	0.4	1	0.1
Z	4%	12%	0	0.1	1

- (a) Compute the variance-covariance matrix Σ and its inverse Σ^{-1} .
- (b) Determine the (global) minimum-variance portfolio of the three assets. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the minimum-variance portfolio.

In addition to the three risky assets, you can also invest in a riskfree asset with a return of $r_f = 3\%$.

- (c) Compute the Sharpe ratios of the three assets.
- (d) Determine the tangency portfolio of the three risky assets. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the tangency portfolio.
- (e) Construct a (standard deviation, expected return)-diagram in which you plot the efficient frontier of risky assets and the efficient frontier of all assets. Mark the points corresponding to the minimum-variance portfolio, the tangency portfolio, and the riskfree asset.

Consider an investor maximizing $E[r] - \frac{1}{2}\gamma \text{Var}[r]$, where γ is the risk aversion parameter. (Here it is important to represent returns by decimal points, so that an expected return of, say, 20% is written as 0.2, and a return variance of, say, 900(%)² is written as 0.09.)

- (f) What is the optimal portfolio if $\gamma = 1$? What is then the expected return and the standard deviation of the investor's portfolio?
- (g) Answer the previous question again assuming $\gamma = 10$ and compare your results.

Exercise 7.9. This is a continuation of Exercise 4.10, so you will have to do that first. First, perform a mean-variance analysis without imposing any constraints on portfolio weights.

- (a) Compute the inverse of the variance-covariance matrix.
- (b) Compute the minimum-variance portfolio and the maximum-slope portfolio. Check that the portfolio weights seem reasonable.
- (c) Construct a diagram in which you draw the efficient frontier of the six ETFs. Indicate where each of the six ETFs is placed in the diagram.
- (d) Compute the tangency portfolio. Check that the portfolio weights seem reasonable.
- (e) Draw the (unconstrained) efficient frontier of all assets in the diagram constructed above.

Next, perform a mean-variance analysis imposing short-selling constraints on the six ETFs, that is the portfolio weights of the ETFs have to be non-negative. In this case, you need to use the Solver in Excel.

- (f) Compute at least ten points on the constrained efficient frontier of risky assets and sketch the frontier in the diagram constructed earlier. Compare with the unconstrained frontier of risky assets.
- (g) Compute the constrained minimum-variance portfolio of the six ETFs. Check that the portfolio weights seem reasonable and compare with the unconstrained minimum-variance portfolio found in (b).
- (h) Compute the constrained tangency portfolio (hint: find the portfolio with maximal Sharpe ratio). Check that the portfolio weights seem reasonable and compare with the unconstrained tangency portfolio found in (d).
- (i) Draw the constrained efficient frontier of all assets in the diagram constructed earlier. Compare with the unconstrained frontier of all assets.

Suppose that you will at most accept a monthly standard deviation of 5%.

- (j) What is your best choice of portfolio in each of the following cases:

1. You invest only in the six ETFs and are not subject to any portfolio constraints.
2. You invest both in the riskfree asset and the six ETFs and you are not subject to any portfolio constraints.
3. You invest only in the six ETFs and are not allowed to have short positions.
4. You invest both in the riskfree asset and the six ETFs and you are not allowed to have short positions in the ETFs.

Compare your answers for the four cases.

Exercise 7.10. You consider investing over the next month in a riskfree asset and the following five exchange-traded funds (ETFs):

Name	Description
EurStx	Vanguard FTSE Europe: tracks index of European stocks
USsmall	Vanguard Small-Cap Blend: tracks index of small U.S. stocks
U\$large	Vanguard Large-Cap Blend: tracks index of large U.S. stocks
LTbonds	Vanguard Long-Term Bond: tracks index of long-term U.S. government bonds
Comm	iShares GSCI Commodity-Indexed Trust: tracks commodity return index

Based on observed monthly returns over the past five years, the following estimates of the expectations, standard deviations, and correlations have been derived:

	Correlations						
	Exp ret	Std Dev	EurStx	USsmall	U\$large	LTbonds	Comm
EurStx	0.0131	0.068	1.00	0.84	0.90	-0.04	0.63
USsmall	0.0185	0.061	0.84	1.00	0.95	-0.21	0.55
U\$large	0.0144	0.046	0.90	0.95	1.00	-0.15	0.63
LTbonds	0.0071	0.032	-0.04	-0.21	-0.15	1.00	-0.25
Comm	0.0014	0.061	0.63	0.55	0.63	-0.25	1.00

Here returns are not written in percentages so, for example, the expected monthly return on EurStx is 0.0131 or 1.31%. Likewise for standard deviations so, for example, the standard deviation of EurStx is 0.068 or 6.8%. In addition to the five risky ETFs, you can also invest in a riskfree asset with a monthly return of $r_f = 0.0001 = 0.01\%$. In questions (a)-(c), assume that there are no restrictions on the portfolios you can choose.

- (a) Determine the (global) minimum-variance portfolio of the five ETFs. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the minimum-variance portfolio.
- (b) Determine the tangency portfolio of the five ETFs. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the tangency portfolio.
- (c) Consider the five combinations of the minimum-variance portfolio and the tangency portfolio in which the weight of the tangency portfolio is 0.1, 0.5, 2, 3, and 4, respectively. For each combination determine the expected return and standard deviation. Use these five combinations, the tangency portfolio, and the minimum-variance portfolio to construct a diagram showing the efficient frontier of risky assets. Also draw the efficient frontier of all assets in the diagram. Mark the points corresponding to the riskfree asset, each of the five ETFs, the minimum-variance portfolio, and the tangency portfolio.
- (d) Now assume that you are not allowed to short sell neither the riskfree asset (i.e., you cannot borrow) nor any of the five ETFs. Determine the constrained tangency portfolio of the five ETFs and its expected return and standard deviation. In the diagram constructed in question (c), draw a sketch of the efficient frontier of all assets when short sales are prohibited.
- (e) Suppose you are willing to accept a standard deviation of your monthly return equal to 0.06 or 6%. In the following consider portfolios involving, at most, the five ETFs and the riskfree asset. Among all portfolios having a standard deviation of 6%, which portfolio has

the highest expected return and how large is this expected return? Among all the portfolios not involving short positions and having a standard deviation of 6%, which portfolio has the highest expected return and how large is this expected return? Comment on the differences between the answers to these two questions.

Exercise 7.11. You consider investing over the next month in a riskfree asset and five exchange-traded funds (ETFs) replicating various U.S. industry sectors. The five ETFs are labelled TEC (for technology stocks), FIN (for stocks of financial companies), UTL (for stocks in utilities), IND (for stocks in industrial companies), and CON (for stocks in manufacturers of consumer products). Based on observed monthly returns over the past five years, the following estimates of expectations, standard deviations, and correlations have been derived:

	Correlations						
	Exp ret	Std Dev	TEC	FIN	UTL	IND	CON
TEC	0.0157	0.0467	1	0.75	0.39	0.82	0.68
FIN	0.0131	0.0781	0.75	1	0.35	0.89	0.70
UTL	0.0083	0.0358	0.39	0.35	1	0.43	0.66
IND	0.0163	0.0618	0.82	0.89	0.43	1	0.74
CON	0.0120	0.0331	0.68	0.70	0.66	0.74	1

Here returns are not written in percentages so, for example, the expected monthly return on TEC is 0.0157 or 1.57%. Likewise for standard deviations so, for example, the standard deviation of TEC is 0.0467 or 4.67%. In addition to the five risky ETFs, you can also invest in a riskfree asset with a monthly return of $r_f = 0.0002 = 0.02\%$. In questions (a)-(c), assume that there are no restrictions on the portfolios you can choose.

- (a) Determine the (global) minimum-variance portfolio of the five ETFs. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the minimum-variance portfolio.
- (b) Determine the tangency portfolio of the five ETFs. Do the portfolio weights seem reasonable given the inputs? Compute the expected return and the standard deviation of the return on the tangency portfolio.
- (c) Consider the six combinations of the minimum-variance portfolio and the tangency portfolio in which the weight of the tangency portfolio is -0.5, -0.1, 0.1, 0.5, 2, and 4, respectively. For each combination determine the expected return and standard deviation. Use these six combinations (add more if you like), the tangency portfolio, and the minimum-variance portfolio to construct a diagram showing the efficient frontier of risky assets. Also draw the efficient frontier of all assets in the diagram. Mark the points corresponding to the riskfree asset, each of the five ETFs, the minimum-variance portfolio, and the tangency portfolio.

Suppose your objective is to maximize

$$E[r] - \frac{1}{2}\gamma \text{Var}[r],$$

where $E[r]$ and $\text{Var}[r]$ are the expected return and return variance per month represented as decimal points (for example $E[r_{TEC}] = 0.0157$, $\text{Var}[r_{TEC}] = (0.0467)^2 = 0.00218089$). The constant γ is your (relative) risk aversion.

- (d) Assume you can invest without any portfolio constraints in both the five ETFs and the riskfree asset. What is your optimal portfolio if $\gamma = 10$ and if $\gamma = 20$? Mark the points corresponding to the two portfolios in your diagram.
- (e) Now assume you can invest in both the five ETFs and the riskfree asset, but that you are not allowed to have any negative portfolio weights. What is then your optimal portfolio if $\gamma = 10$ and if $\gamma = 20$? Mark the points corresponding to these two portfolios in your diagram. Compare with the optimal unconstrained portfolios found in Question (e).

Exercise 7.12. You consider investing in four stocks over the next year. Their rates of return over the next year are assumed to be jointly normally distributed with the following key parameter values:

Asset	Mean	Std dev	Correlations			
			1.0	0.1	0.4	0.5
Stock1	0.08	0.42	1.0	0.1	0.4	0.5
Stock2	0.07	0.51	0.1	1.0	0.0	0.1
Stock3	0.12	0.40	0.4	0.0	1.0	0.8
Stock4	0.12	0.38	0.5	0.1	0.8	1.0

- (a) Compute the variance-covariance matrix and its inverse.
- (b) Compute the minimum-variance portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters presented in the above table.
- (c) Compute the tangency portfolio assuming a riskfree rate of $r_f = 1\% = 0.01$. Discuss whether the portfolio weights seem reasonable given the input parameters presented in the above table.
- (d) Construct a diagram in which you draw the efficient frontier of the four risky assets. Indicate where each of the stocks is located in the diagram.

The riskfree rate of 1% used above is the lending rate, whereas the borrowing rate is 6%.

- (e) Compute the tangency portfolio assuming a riskfree rate of $r_f = 0.06$.
- (f) Draw the efficient frontier of all assets in the diagram constructed above. Explain the shape of the efficient frontier.
- (g) What are the portfolio weights for the efficient portfolio of all assets having an expected return of (i) 10%, (ii) 12%, and (iii) 14%, respectively?
- (h) If you are not allowed to short any of the stocks, what is then the efficient portfolio with an expected return of 14%?
- (i) What is the 5% value-at-risk for the unconstrained efficient portfolio with an expected return of 10%? (If r denotes the return on the portfolio, the 5% value-at-risk is defined as the value of x for which $\text{Prob}(r \leq x) = 0.05$.)
- (j) Of all portfolios with an expected return of 10%, which portfolio has the lowest absolute value of the 5% value-at-risk? Is this portfolio on the efficient frontier? Explain why or why not.

Exercise 7.13. You have some wealth $W_0 > 0$ that you want to invest over the next year. You have decided to invest all the wealth either in a riskfree asset or in the stock market index (via an exchange-traded fund). The riskfree asset has a log-return of r_f , so by investing in this asset your wealth at the end of the year will be $W = W_0 e^{r_f}$ for sure. The stock market index has a log-return of \tilde{r} which is assumed to be normally distributed with mean μ and variance σ^2 .

Since you are risk-averse, you will choose to invest in the stock market index only if μ is sufficiently larger than r_f . But how much larger than r_f does μ have to be? To answer this question, assume that you want to maximize the expected utility of your wealth at the end of the year, and that you have the CRRA utility function

$$u(W) = \frac{1}{1-\gamma} W^{1-\gamma},$$

where γ is the relative risk aversion and we assume that $\gamma > 1$.

- (a) Show that if you invest all of your wealth in the stock market index, your expected utility is given by

$$\mathbb{E}[u(W)] = \frac{1}{1-\gamma} W_0^{1-\gamma} e^{(1-\gamma)\mu + \frac{1}{2}(1-\gamma)^2 \sigma^2}.$$

Hint: Feel free to use the fact that when $X \sim N(m, s^2)$ and k is a constant, then $\mathbb{E}[e^{kX}] = e^{km + \frac{1}{2}k^2s^2}$ (this follows from Appendix A).

- (b) Show that you prefer investing in the stock market index if and only if

$$\mu > r_f + \frac{1}{2}(\gamma - 1)\sigma^2.$$

Carefully explain how you arrive at this condition.

- (c) Assume that $r_f = 0.01$ and $\sigma = 0.2$. What is the minimum value of μ for which the full stock investment is preferred to the riskfree investment if (i) $\gamma = 2$, (ii) $\gamma = 5$, and (iii) $\gamma = 8$? Briefly comment on your findings.

Exercise 7.14. In Markowitz' mean-variance portfolio choice model, the unconstrained efficient frontier of risky assets is generated by the minimum-variance portfolio and the maximum-slope portfolio. Let μ_{\min} and σ_{\min}^2 denote the expectation and the variance of the rate of return on the minimum-variance portfolio. Similarly, let μ_{slope} and σ_{slope}^2 denote the expectation and the variance of the rate of return on the maximum-slope portfolio.

- (a) Suppose you invest a fraction w_{slope} of your wealth in the maximum-slope portfolio and a fraction $1 - w_{\text{slope}}$ in the minimum-variance portfolio. Show that you can write the expectation and the variance of the rate of return on this combined portfolio as

$$\begin{aligned} E[r] &= \mu_{\min} + w_{\text{slope}} (\mu_{\text{slope}} - \mu_{\min}), \\ \text{Var}[r] &= \sigma_{\min}^2 + w_{\text{slope}}^2 (\sigma_{\text{slope}}^2 - \sigma_{\min}^2). \end{aligned}$$

- (b) Suppose your objective is to maximize $E[r] - \frac{\gamma}{2} \text{Var}[r]$ over all portfolios of risky assets (you do not have access to a riskfree asset). Here $\gamma > 0$ is your risk aversion parameter. Show that your optimal portfolio consists of investing the fraction

$$w_{\text{slope}}^* = \frac{\mu_{\text{slope}} - \mu_{\min}}{\gamma (\sigma_{\text{slope}}^2 - \sigma_{\min}^2)}$$

of wealth in the maximum-slope portfolio and the fraction $1 - w_{\text{slope}}^*$ of wealth in the minimum-variance portfolio.

- (c) Let r^* denote the rate of return on the optimal portfolio found in the previous question. Show that the relation

$$E[r^*] - \mu_{\min} = \gamma (\text{Var}[r^*] - \sigma_{\min}^2)$$

holds.

Exercise 7.15. You have \$10,000 that you want to invest over the next year. You evaluate various investment strategies by the expected utility of the wealth at the end of the year, and you apply the CRRA utility function

$$u(W) = \frac{1}{1-\gamma} W^{1-\gamma},$$

where $\gamma > 1$ is the relative risk aversion and where W is your end-of-year wealth in thousands of dollars. In other words, if you end up with a wealth of \$12,000, use $W = 12$ in the utility function.

You consider investing all the wealth in the stock market index (via an exchange-traded fund), and you believe that the log-return \tilde{r} of the stock market index is normally distributed with mean $\mu = 0.05$ and standard deviation $\sigma = 0.20$.

- (a) Show that if you invest all of your wealth in the stock market index, your expected utility is given by

$$E[u(W)] = \frac{1}{1-\gamma} W_0^{1-\gamma} e^{(1-\gamma)\mu + \frac{1}{2}(1-\gamma)^2\sigma^2}.$$

Hint: Feel free to use the fact that when $X \sim N(m, s^2)$ and k is a constant, then $E[e^{kX}] = e^{km + \frac{1}{2}k^2s^2}$ (this follows from Appendix A).

- (b) Suppose the log riskfree rate (also known as the continuously compounded riskfree rate) over the next year is 0.01 or 1%. If your relative risk aversion is $\gamma = 2$, would you prefer investing all of your wealth in the stock market index or in the riskfree asset? What is the answer if $\gamma = 5$? What is the value of γ for which you are indifferent between investing in the stock market and investing in the riskfree asset?
- (c) Your cousin Stella works for a hedge fund and she recommends that you invest all of your wealth in that fund. The fund is rather risky with a return standard deviation of $\sigma_F = 0.40$, whereas you do not yet know its expected rate of return μ_F . If your relative risk aversion is $\gamma = 2$, how big does μ_F have to be in order for you to prefer investing all your wealth in Stella's fund rather than in the stock index or the riskfree asset? What is the answer if $\gamma = 5$?

CHAPTER 8

Multi-period portfolio choice

The previous chapter showed how to find the optimal portfolio of an investor who invests for a single period and only cares about the mean and variance of the return over that period. In short, any risk-averse investor should invest in some mix of the riskfree asset and the tangency portfolio of risky assets. For investors with constant absolute risk aversion (i.e., a negative exponential utility function) we know exactly which mix is optimal.

But investors generally stay in the market for multiple periods and want to rebalance their portfolio every now and then. Many individual investors invest in financial assets primarily to save for their retirement or for some large expenses down the road such as the college tuition for their children or the down payment on a house or an apartment, and thus also have—or at least should have—a long-term perspective. The optimal portfolio for an investor may vary over time because (i) market conditions change, such as the expected returns, variances, and correlations of financial assets, and (ii) relevant investor characteristics change, such as wealth, labor income level, health, or simply the investor’s age. We are therefore really trying to find an optimal *dynamic portfolio* or *investment strategy* that tells the investor which portfolio to have at different points in time depending on the market conditions and the investor’s personal characteristics.

Determining an optimal dynamic portfolio is a computationally challenging task. We are not going into detailed derivations but present a number of models and their conclusions. More details can be found in the original papers referred to throughout this chapter, the book by [Campbell and Viceira \(2002\)](#), and the lecture notes by [Munk \(2017\)](#). As we shall see, a lot of the insights and results from the one-period setting carry over to the multi-period setting, but there are also some important differences between the optimal investments in a multi-period setting and the optimal portfolio in the one-period setting.

First, Section 8.1 sets up a basic model of long-term investments and explains what we can conclude about the optimal investment strategies within that model. In particular, we find a simple expression for the optimal investment strategy of an investor with a constant relative risk aversion. Somewhat surprisingly, the optimal investment strategy in this multi-period model is similar to the optimal portfolio in the one-period mean-variance framework. Section 8.2 discusses what we can say about the optimal investments for alternative investor preferences. Section 8.3 extends the basic model to the case where investment opportunities (expected returns and standard deviations of risky assets, as well as the riskfree rate of return) vary over time.

8.1 Merton's basic model of long-term investments

Samuelson (1969) and Merton (1969, 1971) founded the modern theory of long-term or multi-period investments. They successfully solved for the optimal investment strategy under a number of assumptions that, while not perfectly realistic, are reasonable enough to serve as a good starting point. Samuelson used a discrete-time model where the investor can rebalance the portfolio only at some prespecified dates, e.g. every month or every quarter. Merton used a continuous-time model where the investor can rebalance the portfolio at any point in time. We will follow Merton's approach. Below we describe the assumptions and conclusions of Merton's model.

8.1.1 Assumptions

The key assumptions are the following:

1. The investor can get a riskfree log-return r_f per year, and r_f is constant over time.
2. The log-returns on all risky assets are normally distributed, and the expectations, variances, and correlations of returns over any fixed time horizon stay constant.
3. The investor can trade in the above assets at zero transaction costs at any time, and the assets are perfectly divisible.
4. The investor has a fixed investment horizon ending, say, at time T . In this section we assume that the investor only cares about the wealth she ends up with at time T . Hence, she invests with the objective of maximizing her expected utility of terminal wealth, $E[u(W_T)]$, where u is an increasing and concave function corresponding to the investor being greedy and risk averse, see Section 7.5.
5. The investor has to decide on which investment strategy to follow. An investment strategy consists of a portfolio $\boldsymbol{\pi}_t$ at any point in time t from now until the terminal time T . Here each $\boldsymbol{\pi}_t$ is a vector of the fractions of wealth invested in the risky assets at time t with the remaining wealth being invested in the riskfree asset. Note that $\boldsymbol{\pi}_t$ might be state dependent, i.e. depend on which values some key variables have at time t , e.g. the wealth of the investor.
6. The investor starts out with a certain financial wealth and receives no income from non-financial sources (e.g. labor income). The investor has no non-financial assets such as real estate or other durable goods. The investor can only follow investment strategies that for sure lead to a non-negative terminal wealth; she cannot end up indebted. Since the investor does not receive any labor income, her financial wealth has to stay non-negative throughout life. She can borrow at the riskfree rate r_f in order to invest more in risky assets, but should the value of her risky investments drop to the debt she carries, she would have to sell all risky assets and pay back the debt.

Assumption 2 means that over any time interval, say from time t to time $t + \Delta t$, the gross return on any risky asset i is of the form

$$R_{i,t,t+\Delta t} = \exp \left\{ \left(\mu_i - \frac{1}{2} \sigma_i^2 \right) \Delta t + \sigma_i \varepsilon_{i,t} \sqrt{\Delta t} \right\}, \quad (8.1)$$

where μ_i and $\sigma_i > 0$ are constants, and where $\varepsilon_{i,t}$ is a random variable that follows a standard normal distribution (mean zero, variance one). The log-return $r_{i,t,t+\Delta t}^{\log} = \ln R_{i,t,t+\Delta t}$ is then

$$r_{i,t,t+\Delta t}^{\log} = \left(\mu_i - \frac{1}{2} \sigma_i^2 \right) \Delta t + \sigma_i \varepsilon_{i,t} \sqrt{\Delta t}, \quad (8.2)$$

which implies that the log-return is normally distributed

$$r_{i,t,t+\Delta t}^{\log} \sim N\left(\left[\mu_i - \frac{1}{2}\sigma_i^2\right]\Delta t, \sigma_i^2\Delta t\right). \quad (8.3)$$

In the one-period mean-variance model we assumed that the rate of return is normally distributed which has the inappropriate implication of assigning a positive probability to a rate of return less than -100% . Here, the assumption of a normally distributed log-return is more appropriate as log-returns are unbounded.

From Theorem 3.1, the expectation of the rate of return $r_{i,t,t+\Delta t} = R_{i,t,t+\Delta t} - 1$ is

$$\mathbb{E}[r_{i,t,t+\Delta t}] = \exp\left\{\left[\mu_i - \frac{1}{2}\sigma_i^2\right]\Delta t + \frac{1}{2}\sigma_i^2\Delta t\right\} - 1 = \exp\{\mu_i\Delta t\} - 1 \approx \mu_i\Delta t,$$

where the approximation is based on $e^x \approx 1 + x$ for x close to zero and is thus better for Δt small. From (8.3) it is clear that

$$\mathbb{E}[r_{i,t,t+\Delta t}^{\log}] = \left(\mu_i - \frac{1}{2}\sigma_i^2\right)\Delta t, \quad (8.4)$$

$$\text{Var}[r_{i,t,t+\Delta t}^{\log}] = \sigma_i^2\Delta t. \quad (8.5)$$

In particular, σ_i^2 is the annualized variance and σ_i the annualized standard deviation of the log-return, and σ_i is often called the *volatility* of the asset price.

The returns of different risky assets can be correlated. Suppose $\text{Corr}[\varepsilon_{i,t}, \varepsilon_{j,t}] = \rho_{ij}$. Since $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$ both have a standard deviation of 1, it follows that $\text{Cov}[\varepsilon_{i,t}, \varepsilon_{j,t}] = \rho_{ij}$. More importantly, ρ_{ij} is also the correlation between the log-returns of assets i and j over any period. To see this, first compute the covariance where we can ignore the first term of the log-return as this is not random:

$$\begin{aligned} \text{Cov}[r_{i,t,t+\Delta t}^{\log}, r_{j,t,t+\Delta t}^{\log}] &= \text{Cov}[\sigma_i\varepsilon_{i,t}\sqrt{\Delta t}, \sigma_j\varepsilon_{j,t}\sqrt{\Delta t}] \\ &= \sigma_i\sigma_j\Delta t \text{Cov}[\varepsilon_{i,t}, \varepsilon_{j,t}] = \rho_{ij}\sigma_i\sigma_j\Delta t. \end{aligned}$$

Hence the correlation is

$$\text{Corr}[r_{i,t,t+\Delta t}^{\log}, r_{j,t,t+\Delta t}^{\log}] = \frac{\text{Cov}[r_{i,t,t+\Delta t}^{\log}, r_{j,t,t+\Delta t}^{\log}]}{\text{Std}[r_{i,t,t+\Delta t}^{\log}]\text{Std}[r_{j,t,t+\Delta t}^{\log}]} = \frac{\rho_{ij}\sigma_i\sigma_j\Delta t}{\sqrt{\sigma_i^2\Delta t}\sqrt{\sigma_j^2\Delta t}} = \rho_{ij}. \quad (8.6)$$

Let N denote the number of risky assets we can invest in, as in the preceding chapter. Let $\boldsymbol{\mu}$ denote the vector of expected return parameters and $\underline{\Sigma}$ the annualized variance-covariance matrix,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix}, \quad \underline{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1N}\sigma_1\sigma_N \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho_{2N}\sigma_2\sigma_N \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N}\sigma_1\sigma_N & \rho_{2N}\sigma_2\sigma_N & \dots & \sigma_N^2 \end{pmatrix}.$$

We can estimate the values in $\boldsymbol{\mu}$ and $\underline{\Sigma}$ from past returns on the assets. Given a time-series of log-returns on the assets with a specific observation frequency Δt (e.g. $\Delta t = 1/12$ for monthly observations), we can estimate σ_i^2 and ρ_{ij} from the sample variances and

correlations of the log-returns by using (8.5) and (8.6). Based on (8.4), we can estimate μ_i from the sample arithmetic average of the log-returns and the σ_i -estimate. Recall from (3.94) that the arithmetic average of the log-returns is closely related to the geometric average of the rates of return.

As stated in Theorem 3.4 we know that if individual assets' log-returns are normally distributed, then the log-return on a buy-and-hold portfolio is not normally distributed. However, it turns out that if the portfolio is continuously rebalanced to maintain constant portfolio weights, then the log-return on the portfolio is in fact normally distributed. The proof of this result relies on rather advanced mathematics, but the interested reader can find the details, e.g., in Munk (2013). We state the result in the following theorem.

Theorem 8.1

Suppose the assumptions of Merton's basic model of long-term investments are satisfied. Consider a constant-weight investment strategy of the riskfree asset and the N risky assets. Let π be the vector of portfolio weights of the risky assets so that $1 - \pi \cdot \mathbf{1}$ is the weight of the riskfree asset. Then the log-return $r_{\pi,t,t+\Delta t}^{\log}$ on this constant-weight portfolio over the interval $[t, t + \Delta t]$ is normally distributed:

$$r_{\pi,t,t+\Delta t}^{\log} \sim N \left(\left[r_f + \pi \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{1}{2} \pi \cdot \underline{\Sigma} \pi \right] \Delta t, \pi \cdot \underline{\Sigma} \pi \Delta t \right). \quad (8.7)$$

In fact, even when the portfolio weights are allowed to vary deterministically over time (e.g., with the age of the investor), the log-return on a continuously rebalanced portfolio is normally distributed.

Note that keeping your portfolio weights constant requires trading. For example, if you do not rebalance a two-stock portfolio and stock 1 increases more in price than stock 2, then the portfolio weight of stock 1 goes up and the weight of stock 2 declines. To maintain a constant portfolio weight you have to sell units of the more successful stocks in your portfolio and purchase additional units of the less successful stocks, a "sell winners, buy losers" strategy. Of course, in reality you cannot rebalance the portfolio continuously, and you would not want to do so as trading involves transaction costs, no matter how small they are nowadays. If you only rebalance at a regular frequency (say daily, weekly, or monthly) or when the current portfolio weights have deviated substantially from the targeted constant weights, your portfolio return is typically close to that suggested by the above continuous-time limit. See also the discussion in Section 8.1.5.

8.1.2 Two-fund separation for a general utility function

We can form a tangency portfolio of the risky assets exactly as in the one-period model of the preceding chapter:

$$\pi_{\tan} = \frac{\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})}. \quad (8.8)$$

Although we now operate in a multi-period setting, we have assumed that r_f , $\boldsymbol{\mu}$, and $\underline{\Sigma}$ are all constant over time, which implies that the tangency portfolio also stays the same.

Based on rather complicated mathematics, Merton (1969, 1971) showed that the **two-fund separation** result of the one-period model carries over to the multi-period setting:

Theorem 8.2

Given the assumptions of Merton's basic model of long-term investments, the optimal portfolio of any investor at any point in time is a combination of the riskfree asset and the tangency portfolio of the risky assets as defined in (8.8). The optimal combination of wealth invested in the riskfree asset and in the tangency portfolio can vary over time and differ across investors depending on their risk aversion.

A consequence of two-fund separation is that all investors should hold risky assets in the same proportion, i.e., π_i/π_j is the same for all investors for any pair (i, j) of risky assets.

We can illustrate the investment problem in a (standard deviation, mean)-diagram as we did in the one-period setting. Now we form such a picture for any point in time, and along the axis we have the standard deviation and expectation of the instantaneous rate of return. Each risky asset corresponds to a point in the diagram and, with μ_i and σ_i being constant, the point does not move around. We can form the efficient frontier of risky assets as in the one-period model. Since we have also assumed constant correlations, this frontier stays the same over time. Finally, because of the assumption of a constant riskfree rate, the wedge representing the efficient frontier of all assets remains constant over time. Any investor picks a portfolio corresponding to a point on the upward-sloping part of the wedge and such a point refers to some combination of the riskfree asset and the tangency portfolio of risky assets. So as long as the investment opportunities (as characterized by r_f and the μ_i 's, σ_i 's, and ρ_{ij} 's) are not time-varying, the extension from a one-period setting to a multi-period setting does not really change the results.

8.1.3 Optimal investment strategy for CRRA utility

The two-fund separation theorem does not say *which* combination of the riskfree asset and the tangency portfolio a given investor should hold at any given point in time. But, in fact, Merton (1969) also found a specific answer to that question for the quite reasonable CRRA utility function $u(W) = \frac{1}{1-\gamma}W^{1-\gamma}$ that was introduced in Section 7.5.4. Recall that γ is the constant relative risk aversion coefficient. Since we assume that the individual only cares about her wealth W_T at some given future point in time T , her objective is to maximize $E[\frac{1}{1-\gamma}W_T^{1-\gamma}] = \frac{1}{1-\gamma}E[W_T^{1-\gamma}]$. The next theorem characterizes the optimal investment strategy which turns out to be a strategy holding portfolio weights constant over time. In the proof of the theorem, we confirm that this is the best portfolio weight vector among all investment strategies with constant portfolio weights. We skip the much more complicated proof that the stated strategy is also the best among all possible investment strategies including all strategies with non-constant weights.

Theorem 8.3

Under the assumptions of Merton's basic model of long-term investments, the optimal investment strategy for an investor with CRRA utility of terminal wealth and a relative risk aversion of γ is to hold, at any time, the portfolio

$$\boldsymbol{\pi} = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) \quad (8.9)$$

of risky assets as well as the fraction $1 - \boldsymbol{\pi} \cdot \mathbf{1}$ of wealth in the riskfree asset. Equivalently, the optimal strategy is to invest the fraction $[\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})]/\gamma$ of wealth in the tangency portfolio of risky assets and the fraction $1 - [\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})]/\gamma$ of wealth in the riskfree asset.

Proof

Suppose that the individual only considers investment strategies for which the portfolio weight vector is constant over the entire time period $[0, T]$. The log-return on such a strategy is $r_{\boldsymbol{\pi}, 0, T}^{\log} = \ln R_{\boldsymbol{\pi}, 0, T}$, where $R_{\boldsymbol{\pi}, 0, T}$ is the gross return. We know from Theorem 8.1 that

$$r_{\boldsymbol{\pi}, 0, T}^{\log} = \ln R_{\boldsymbol{\pi}, 0, T} \sim N \left(\left[r_f + \boldsymbol{\pi} \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{1}{2} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \right] T, \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} T \right). \quad (8.10)$$

If we invest W_0 at time 0 in the constant-weight portfolio $\boldsymbol{\pi}$ until time T , we end up with a wealth of

$$W_T = W_0 R_{\boldsymbol{\pi}, 0, T}. \quad (8.11)$$

For a given choice of $\boldsymbol{\pi}$, the expected CRRA utility of terminal wealth is thus

$$E[u(W_T)] = \frac{1}{1-\gamma} E[W_T^{1-\gamma}] = \frac{1}{1-\gamma} W_0^{1-\gamma} E[(R_{\boldsymbol{\pi}, 0, T})^{1-\gamma}] \quad (8.12)$$

Since the gross return is lognormally distributed, it follows from Theorem A.2 in Appendix A that

$$\begin{aligned} E[(R_{\boldsymbol{\pi}, 0, T})^{1-\gamma}] &= \exp \left\{ (1-\gamma) \left(r_f + \boldsymbol{\pi} \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{1}{2} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \right) T + \frac{1}{2} (1-\gamma)^2 \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} T \right\} \\ &= \exp \left\{ (1-\gamma) \left(r_f + \boldsymbol{\pi} \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{\gamma}{2} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \right) T \right\}. \end{aligned}$$

Hence, the expected utility is

$$E[u(W_T)] = \frac{1}{1-\gamma} W_0^{1-\gamma} \exp \left\{ (1-\gamma) \left(r_f + \boldsymbol{\pi} \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{\gamma}{2} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \right) T \right\}. \quad (8.13)$$

The portfolio solving the problem

$$\max_{\boldsymbol{\pi}} f(\boldsymbol{\pi}), \quad \text{where } f(\boldsymbol{\pi}) = \left(r_f + \boldsymbol{\pi} \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{\gamma}{2} \boldsymbol{\pi} \cdot \underline{\Sigma} \boldsymbol{\pi} \right), \quad (8.14)$$

will also maximize $E[u(W_T)]$ since the function $x \mapsto \frac{1}{1-\gamma} W_0^{1-\gamma} \exp\{(1-\gamma)Tx\}$ is increasing. The derivative of $f(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ is the vector

$$f'(\boldsymbol{\pi}) = \boldsymbol{\mu} - r_f \mathbf{1} - \gamma \underline{\Sigma} \boldsymbol{\pi}$$

and by equating this with zero and solving for $\boldsymbol{\pi}$ we find (8.9). The second-order condition for a maximum can be verified, so this *is* the best of all constant-weight portfolio strategies for an investor with CRRA utility.

While focusing on constant-weight portfolios seems restrictive, Merton (1969) showed that the optimal constant-weight investment strategy we just derived is in fact the optimal among all investment strategies. As the proof involves advanced mathematics, we skip it here. The courageous reader is referred to [Munk \(2017, Ch. 6\)](#).

The optimal investment strategy has a number of notable and reasonable properties:

1. The higher the risk aversion coefficient γ , the lower the investment in the risky assets and the higher the investment in the riskfree asset.
2. The optimal portfolio weights are independent of the wealth of the investor, which is an implication of the investor's constant relative risk aversion. Of course, the higher the wealth of the investor, the higher the amount of money invested in each asset.
3. The optimal investment strategy is independent of the horizon of the investor. The assumption of constant investment opportunities—i.e. constant riskfree rate, means, variances, and correlations—is crucial for obtaining a horizon-independent optimal portfolio.
4. The fraction of wealth invested in each asset is to be kept constant over time. As explained above, this requires continuous rebalancing of the portfolio since the asset prices vary all the time. To keep weights constant you have to sell winners and buy losers.
5. Suppose that there is just a single risky asset, which could represent the tangency portfolio or maybe a stock market index fund. Let μ and σ^2 denote the expectation and variance parameters for this asset. Then (8.9) implies that the fraction of wealth optimally invested in the risky asset is

$$\pi = \frac{\mu - r_f}{\gamma \sigma^2}, \quad (8.15)$$

which is increasing in the expected return and decreasing in the variance, as we would expect. The portfolio weight is often written in terms of the Sharpe ratio $\lambda = (\mu - r_f)/\sigma$ of the risky asset:

$$\pi = \frac{\lambda}{\gamma \sigma}. \quad (8.16)$$

8.1.4 Discussion of the optimal investment strategy for CRRA utility

As explained in Section 6.5, stock investments have typically outperformed bond investments for long investment horizons, whereas over shorter horizons it is seen more often that bond investments outperform stocks. Referring to these empirical facts, many investment consultants recommend that long-term investors should place a large part of their wealth in stocks and then gradually shift from stocks to bonds as they get older and their investment horizon shrinks. However, this recommendation conflicts with the optimal portfolio strategy we have derived above. According to our analysis, the optimal portfolio weights of CRRA investors are independent of the investment horizon. Is this because our model of the financial asset prices is inconsistent with the basic empirical facts? No!

To see this let us consider the simple case with a single risky asset representing the stock index. According to Eq. (8.3), the T -year log-return on the risky asset is then normally

distributed:

$$r_{0,T}^{\log} \sim N\left((\mu - \frac{1}{2}\sigma^2)T, \sigma^2 T\right). \quad (8.17)$$

For a general constant x , we thus have from Eqs. (3.23) and (3.26) that

$$\text{Prob}\left(r_{0,T}^{\log} < x\right) = N\left(\frac{x - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}\right), \quad \text{Prob}\left(r_{0,T}^{\log} > x\right) = N\left(\frac{(\mu - \frac{1}{2}\sigma^2)T - x}{\sigma\sqrt{T}}\right).$$

In the model, r_f denotes the log of the riskfree rate per year, so over a T -year period the log of the riskfree return is Tr_f . Therefore, the probability that over a T -year period an investment in the risky asset outperforms a riskfree investment is

$$\text{Prob}\left(r_{0,T}^{\log} > Tr_f\right) = N\left(\frac{(\mu - \frac{1}{2}\sigma^2)T - Tr_f}{\sigma\sqrt{T}}\right) = N\left(\frac{(\mu - r_f - \frac{1}{2}\sigma^2)\sqrt{T}}{\sigma}\right). \quad (8.18)$$

Figure 8.1 illustrates the relation between this outperformance probability and the investment horizon T . The curves differ with respect to the presumed expected rate of return on the stock, i.e., μ , whereas the riskfree rate is 2% and the volatility of the stock is 20% for all curves. As discussed in Section 6.5, U.S. stocks have had an average excess rate of return of 8–9% per year over the past century or so. A μ -value of 12% corresponds to an expected excess rate of return of 8% per year since $0.12 - 0.02 - (0.20)^2/2 = 0.08$. However, it should be emphasized that historical estimates of expected rates of return, volatilities, and correlations are not necessarily good predictors of the future values of these quantities. In particular, there are reasons to believe that the average return on the U.S. stock market over the past century is higher than what the stock market is currently offering in terms of expected returns. Probably the curve labeled $\mu = 8\%$ is more representative of the current investment opportunities. In any case, it is tempting to conclude from the graph that long-term investors should invest more in stocks than short-term investors. Why does the optimal portfolio derived previously not reflect this property?

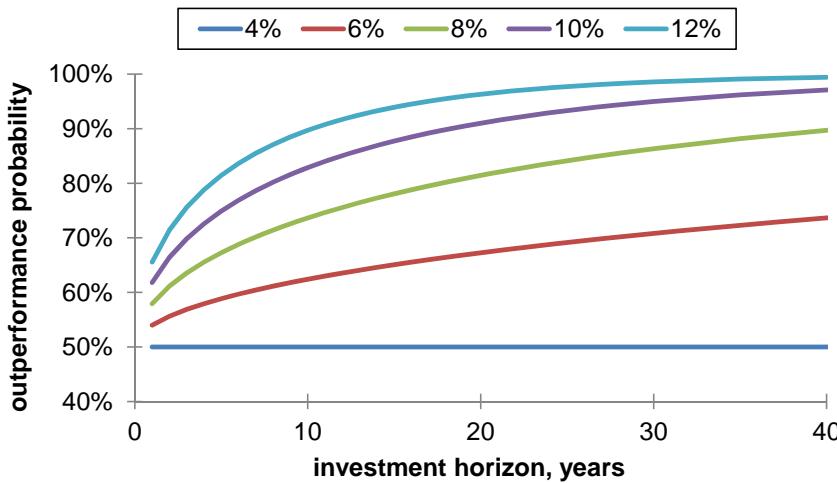
It is important to realize that the optimal decision cannot be based just on the probabilities of gains and losses. After all, most individuals would reject a gamble with a 99% probability of winning 1 dollar and a 1% probability of losing a million dollars. The magnitudes of gains and losses are also important for the optimal investment decision. Let us look at the probability that the T -year rate of return $r_{0,T}$ on the stock is K percentage points lower than the riskfree rate of return $e^{r_f T} - 1$:

$$\begin{aligned} \text{Prob}\left(r_{0,T} < e^{r_f T} - 1 - K\right) &= \text{Prob}\left(r_{0,T}^{\log} < \ln(e^{r_f T} - K)\right) \\ &= N\left(\frac{\ln(e^{r_f T} - K) - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}\right). \end{aligned} \quad (8.19)$$

Table 8.1 shows such probabilities for various combinations of the return shortfall constant K and the horizon assuming $r_f = 2\%$, $\mu = 8\%$, and $\sigma = 20\%$ (the numbers in the row labeled 0% are equal to 100% minus the outperformance probabilities shown in Figure 8.1.) Over a 10-year period the rate of return on a riskfree investment at a rate of 2% per year is

$$\left(e^{0.02 \times 10} - 1\right) \times 100\% \approx 22.1\%.$$

The table shows that with a 16% probability a stock investment over a 10-year period gives a rate of return which is lower than $22.1\% - 25\% = -2.9\%$, and there is a 1.6%

**Figure 8.1: Outperformance probabilities.**

The figure shows the probability that a stock investment outperforms a riskfree investment over different investment horizons. For all curves the riskfree interest rate is 2%, and the volatility of the stock is 20%. Each of the curves correspond to the value of the parameter μ shown in the legend.

Excess return on bond	1 year	10 years	40 years
0%	42.1%	26.4%	10.3%
25%	5.4%	16.0%	8.7%
50%	0.0%	7.1%	7.1%
75%	0.0%	1.6%	5.6%
100%	0.0%	0.0%	4.1%

Table 8.1: Underperformance probabilities.

The table shows the probability that a stock investment over a period of 1, 10, and 40 years provides a percentage return which is at least 0, 25, 50, 75, or 100 percentage points lower than the riskfree return. The numbers are computed using the parameter values $\mu = 8\%$, $r_f = 2\%$, and $\sigma = 20\%$.

probability that the stock return is lower than $22.1\% - 75\% = -52.9\%$. Over a 40-year period the riskfree return is 122.6%. There is a 4.1% probability that a stock investment yields a return at least 100 percentage points lower, i.e., lower than 22.6%. Over longer periods the probability that stocks underperform bonds is lower, but the probability of extremely bad stock returns is larger than over short periods. The expected excess return on the stock increases with the length of the investment horizon, but so does the variance of the return. Any risk-averse investor has to consider this trade-off. For a CRRA investor in our simple financial model, the two effects offset each other exactly so that the optimal portfolio is independent of the investment horizon!

8.1.5 Loss due to suboptimal investments

How important is it to invest in precisely the optimal portfolio? Obviously, investing in any other portfolio leads to a lower expected utility. As explained earlier, the exact numerical value of a utility function says very little and so does the difference between

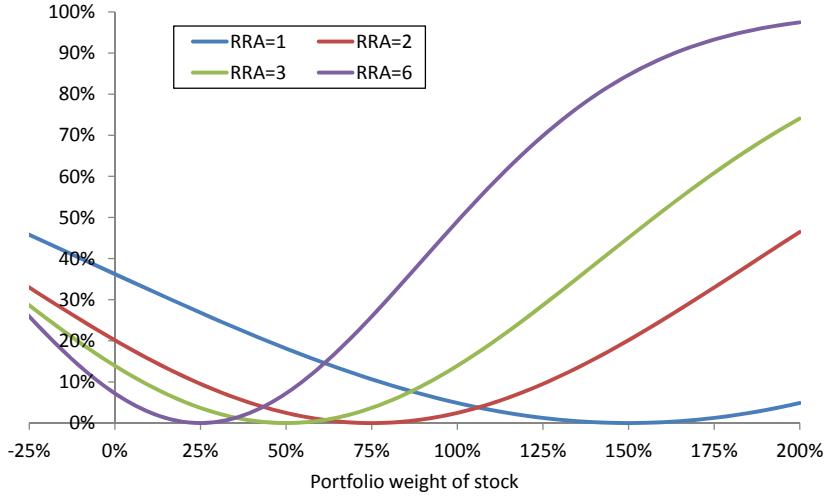


Figure 8.2: Welfare losses for different levels of risk aversion.

The figure shows the wealth-equivalent utility loss $\ell(\pi)$ from applying a suboptimal constant portfolio weight π instead of the optimal portfolio weight. The loss is depicted as a function of π with different curves for different levels of the relative risk aversion γ . The investment horizon is $T = 10$ years and the stock has a Sharpe ratio of 0.3 and a volatility of $\sigma = 0.2$.

the optimal utility and the utility derived from following a different investment strategy. But the utility difference can be transformed into a wealth difference which is directly interpretable. Let us again consider the case with a single risky asset representing the stock market index. Then it can be shown that a CRRA investor holding some fraction π of wealth in the stock index over the horizon $[0, T]$ is willing to give up a fraction

$$\ell(\pi) = 1 - \exp \left\{ -\frac{\gamma\sigma^2}{2} (\pi^* - \pi)^2 T \right\} \approx \frac{\gamma\sigma^2}{2} (\pi^* - \pi)^2 T \quad (8.20)$$

of her wealth in order to always invest the optimal fraction $\pi^* = (\mu - r_f)/(\gamma\sigma^2)$; a proof can be found in Munk (2017, Sec. 6.7). The number $\ell(\pi)$ is called the wealth-equivalent loss from following the suboptimal strategy π . The approximation is based on $e^x \approx 1 + x$ for x near 0, so it is only useful for short horizons.

Figure 8.2 illustrates the wealth-equivalent loss $\ell(\pi)$ as a function of the portfolio weight π for four different levels of the relative risk aversion γ . The investment horizon is fixed to 10 years, the Sharpe ratio of the stock is assumed to be $(\mu - r_f)/\sigma = 0.3$, and the volatility of the stock is assumed to be $\sigma = 0.2$ so that its excess expected return is $\mu - r_f = 0.06 = 6\%$. We see that the losses are relatively flat around the optimal portfolio weight. Large deviations from the optimal portfolio weight are necessary to obtain substantial losses. Highly risk-averse individuals are more sensitive to deviations from the optimal portfolio weight. Figure 8.3 depicts the wealth-equivalent loss $\ell(\pi)$ as a function of π for different investment horizons. Clearly, the individual suffers a bigger loss from following a suboptimal strategy over longer periods.

A related question is how important the frequency of portfolio rebalancing is. In theory, it is optimal to rebalance continuously in time to make sure that the portfolio weights are always equal to (8.9) in the case of multiple risky assets or (8.15) in the case of a single risky asset. But continuous rebalancing is not practically possible and, even with tiny

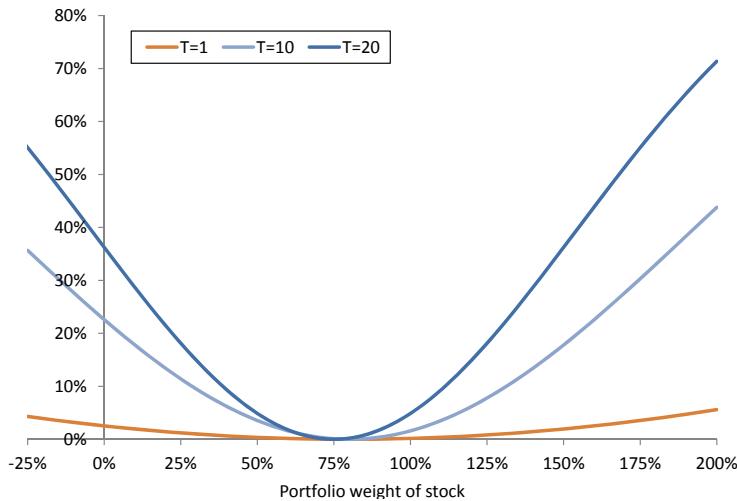


Figure 8.3: Welfare losses for different investment horizons.

The figure shows the wealth-equivalent utility loss $\ell(\pi)$ from applying a suboptimal constant portfolio weight π instead of the optimal portfolio weight. The loss is depicted as a function of π with different curves for different investment horizons T . The relative risk aversion is $\gamma = 2$ and the stock has a Sharpe ratio of 0.3 and a volatility of $\sigma = 0.2$.

trading costs per transaction, continuous rebalancing would be infinitely expensive. It is therefore interesting to see how bad it is to rebalance in a non-continuous way.

A very simple strategy is to predetermine a finite number of trading dates. At each trading date the portfolio is rebalanced so that the portfolio weights coincide with the solution for the continuous-time case. In between trading dates, the portfolio weights deviate somewhat from the truly optimal weights. Ignoring the trading costs, such a strategy is suboptimal and the investor incurs a utility loss, which can be transformed into a wealth-equivalent loss as above. In this case, however, it is not possible to express the wealth-equivalent loss by an explicit formula. It can be estimated by so-called Monte Carlo simulation as explained in [Munk \(2017, Sec. 6.8\)](#).

As an example, assume $r_f = 0.02$ and a single risky asset with $\sigma = 0.2$ and $\mu = 0.08$. Consider an investor with a relative risk aversion of $\gamma = 2$ and an investment horizon of $T = 10$ years. The optimal strategy is to have $\pi = 0.75 = 75\%$ of the wealth invested in the stock at any point in time. With only quarterly rebalancing of the portfolio, [Munk \(2017\)](#) reports a wealth-equivalent loss of only 0.26%. This indicates that it is not important to rebalance the portfolio very frequently. More frequent rebalancing reduces the wealth-equivalent loss, but also lead to higher transaction costs. Between two adjacent rebalancing dates the portfolio weight of the stock can deviate somewhat from the optimal weight, but the deviation is typically rather small, and we have already seen above that expected utility is relatively insensitive to small deviations from the optimal strategy.

[Rogers \(2001\)](#) provides a more formal analysis of the impact of infrequent portfolio rebalancing. [Branger, Breuer, and Schlag \(2010\)](#) perform a detailed Monte Carlo simulation study, also for some models with non-constant investment opportunities. Their study confirms that for investment problems involving only stocks and bonds, relatively infrequent rebalancing induces only small wealth-equivalent losses.

It is complicated to derive the truly optimal investment strategy in a model including trading costs. The simplest type of trading costs to handle is proportional costs meaning that the trading costs are proportional to the value of the stocks being traded. The

theoretical work by Magill and Constantinides (1976), Constantinides (1979, 1986), and Davis and Norman (1990) considered a model with a riskfree asset and a single risky asset. The riskfree asset is traded without costs, whereas the risky asset is traded with proportional costs. They conclude that as long as the fraction π of wealth held in the risky asset is in a certain interval $[\pi_1, \pi_2]$, it does not pay off to trade the risky asset. The higher the transaction costs, the wider the no-trade interval. When the price of the asset moves enough that π is about to break one of the boundaries of that interval, it is optimal to trade just sufficiently to keep π in the interval. Rather than fixing the time period between portfolio adjustments, it is thus better to rebalance whenever the price of the risky asset has moved by a certain percentage. However, determining exactly how big a percentage it has to move involves quite complicated calculations. Rebalancing at a regular time frequency is simpler to implement and often does quite well as indicated by the small loss for quarterly rebalancing reported above. For a different approach to portfolio choice with transaction costs, see Garleanu and Pedersen (2013).

8.1.6 Merton versus Markowitz

The assumptions of constant relative risk aversion, normally distributed log-returns, and portfolio rebalancing that underlie Merton's dynamic model are more realistic than the assumptions of Markowitz' static mean-variance model with constant absolute risk aversion, normally distributed rates of return, and no portfolio rebalancing. Nevertheless, the models seem to give identical results! The optimal portfolio in Eq. (8.9) for the dynamic model is identical to the optimal portfolio in Eq. (7.56) for the static model. However, r_f , μ , and Σ have slightly different interpretations in the two models. In the dynamic Merton model r_f is the log riskfree rate (or continuously compounded riskfree rate) per year, μ contains logs of expected gross returns per year, and Σ contains variances and covariances of assets' log-returns per year. In the static Markowitz model r_f is the riskfree rate of return over the length of the period considered, μ is the vector of expected rates of return over the period, and Σ contains variances and covariances of assets' rates of return over the period. This difference in interpretations seems to imply that you should apply different values of r_f , μ , and Σ in the two models and therefore you will get different optimal portfolios, although in practice the differences are likely to be small.

The similarity of the results of the two models suggests that the static mean-variance model can generate useful portfolio recommendations also for a world in which investors have long-term objectives and regularly rebalance their portfolios, provided that the assumptions listed in Section 8.1.1 are valid or at least close to reality. For example, suppose you believe in these assumptions but face portfolio constraints that make the optimal unconstrained portfolio in Eq. (8.9) infeasible. Unfortunately, deriving the optimal investment strategy in the dynamic model is extremely complicated when portfolio constraints are involved. Alternatively, you can derive the optimal constrained portfolio in the mean-variance setting following the procedure explained in Section 7.4.3 and then make sure that your portfolio is regularly rebalanced to match those portfolio weights. This strategy is likely to be very close to the unknown optimal dynamic strategy. In this case, when applying the mean-variance model, you should ideally use values of r_f , μ , and Σ that reflect their interpretation in the dynamic model.

In Section 9.1 we are going to use a similar procedure to find optimal investment strategies for investors that receive a regular income, e.g. labor income. Also in this case, it is computationally difficult to find optimal investment strategies in the more realistic dynamic model, but we can do it in an extended version of the mean-variance model.

8.2 Alternative preferences

8.2.1 CRRA utility of consumption

In the preceding section we assumed that the investor maximizes utility of her wealth at some future point in time. From the perspective of an individual investor (or a household), it is really the consumption of goods throughout life that determines welfare rather than wealth at a single point in time. The utility of consumption at a given point in time, say time t , can again be modeled by a utility function $u(c_t)$, where c_t could represent the number of consumption goods consumed at time t or, given the diversity of consumption goods, maybe the money spent on consumption at time t . It makes sense that u is increasing since people generally want to consume as much as possible, although for some specific goods there might be an upper limit. It also makes sense that u is a concave function since this means that the marginal utility $u'(c_t)$ is decreasing.

Merton (1969) considered a continuous-time model in which the individual can consume and rebalance her portfolio continuously in time, and the individual at some initial time 0 wants to maximize expected life-time utility of consumption, $E \left[\int_0^T e^{-\theta t} u(c_t) dt \right]$. Here, the Greek letter θ (“theta”) is the individual’s subjective time preference rate or utility discount rate. A high θ means that future consumption is weighted less, corresponding to a more impatient consumer. The individual only invests in order to generate higher future consumption.

Merton showed that the results from the basic model are still true. For a general utility function, two-fund separation still applies: any investor combines the riskfree asset with the tangency portfolio of risky assets. And for the CRRA utility function, the optimal portfolio at any time is still given by Eq. (8.9). Of course, with the same initial wealth W_0 , an investor consuming throughout life has a lower wealth at any time $t > 0$ than an investor not consuming, so the amounts invested in each asset and the number of shares held of each asset differ between the model with consumption and the model without. But the optimal fraction of wealth invested in each asset is the same in the two models.

Merton further derived the optimal consumption strategy, that is how much to consume at each point in time, and thus also how much to invest (the residual wealth after any consumption). The optimal consumption over a tiny interval $[t, t+dt]$ equals $c_t \times dt$, where

$$c_t = \frac{A}{1 - e^{-A[T-t]}} W_t. \quad (8.21)$$

Here W_t is the wealth of the individual at time t and A is the constant

$$A = \frac{\theta + r_f(\gamma - 1)}{\gamma} + \frac{1}{2} \frac{\gamma - 1}{\gamma^2} \lambda^2 \quad (8.22)$$

with

$$\lambda^2 = (\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})$$

being the square of the Sharpe ratio of the tangency portfolio, cf. (7.39). We assume $A \neq 0$. In particular, under the reasonable assumptions that $\gamma > 1$ and that θ and r_f are both non-negative, then $A > 0$. It is optimal for the individual to consume a time-varying fraction of wealth. The consumption-wealth ratio c_t/W_t is increasing over time t for a fixed end date T , increasing in the time preference rate θ , and—assuming $\gamma > 1$ —increasing in the riskfree rate r_f and the squared Sharpe ratio of the tangency portfolio. All these properties are in line with intuition. For example, higher returns in form of a riskfree return or risk premiums on the risky assets imply that the individual can afford a higher

level of consumption throughout life.

The initial expectation of future optimal consumption is

$$E[c_t] = c_0 \exp \left\{ \frac{1}{\gamma} \left(r_f - \theta + \frac{\gamma+1}{2\gamma} \lambda^2 \right) t \right\} \quad (8.23)$$

(see Munk 2017, Ch. 6). In the model, consumption is thus expected to either increase with age, decrease with age, or to be age-independent depending on whether the constant $r_f - \theta + \frac{\gamma+1}{2\gamma} \lambda^2$ is positive, negative, or zero. Economists frequently assume a time preference rate close to the riskfree rate and since γ and λ^2 are positive, this constant is then positive so that consumption should increase, on average, over life. However, various experimental studies (see Andersen, Harrison, Lau, and Rutström (2014) and the references therein) find that some individuals have a significantly higher time preference rate, which would lead to a decreasing expected consumption pattern over life.

Empirical studies show a hump-shaped consumption pattern over the life cycle (Browning and Crossley 2001, Gourinchas and Parker 2002) so that consumption typically increases up to an age of around 45-50 years and then drops until retirement and probably even also in retirement. The relatively simple consumption-investment model considered above cannot generate such a pattern. We return to this issue below.

8.2.2 Subsistence consumption

The life-time consumption-investment decision problem can be solved for utility functions that are slightly different from the CRRA utility function and maybe more realistic. One example is the so-called subsistence power utility function

$$u(c) = \frac{1}{1-\gamma} (c - \bar{c})^{1-\gamma}, \quad \text{for } c > \bar{c}, \quad (8.24)$$

where \bar{c} represents a subsistence level of consumption that the individual must have to survive. Any consumption above the subsistence level increases the utility of the individual. In this case the relative risk aversion is

$$\text{RRA}(c) = -\frac{cu''(c)}{u'(c)} = \frac{\gamma c}{c - \bar{c}} = \frac{\gamma}{1 - (\bar{c}/c)},$$

which is now decreasing in the level of consumption. If you consume little so that c is close to (but larger than) \bar{c} , the relative risk aversion is high as the individual will not risk a further decline in her consumption. But if consumption is high, the relative risk aversion is lower. In the limit as consumption goes to infinity, the ratio \bar{c}/c goes to zero, and we can see that the relative risk aversion approaches γ . Some empirical studies (Cicchetti and Dubin 1994, Ogaki and Zhang 2001) find support of such a specification of preferences.

If we assume continuous consumption, the present value of the subsistence consumption from some time t and to the terminal time T is

$$f(t) = \int_t^T e^{-r_f(s-t)} \bar{c} ds = \bar{c} \int_t^T e^{-r_f(s-t)} ds = \frac{\bar{c}}{r_f} \left(1 - e^{-r_f(T-t)} \right).$$

At time t the individual has to invest the amount $f(t)$ in the riskfree asset to be sure to cover the future subsistence consumption. Obviously, the buffer $f(t)$ is decreasing over time as less and less money has to be reserved when the remaining horizon shortens. The remaining “disposable” wealth $W_t - f(t)$ can be used on additional consumption and on

investments. The utility of additional consumption is as in the case of CRRA utility discussed above so she optimally spends a fraction of the disposable wealth on additional consumption. The optimal total consumption is therefore

$$c_t = \bar{c} + \frac{A}{1 - e^{-A[T-t]}} (W_t - f(t)), \quad (8.25)$$

where A is still given by (8.22).

Likewise, the disposable wealth is optimally invested as the wealth in the CRRA case. This means that individual invests the fraction $\frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})$ of disposable wealth in the risky assets. As a fraction of total wealth W_t , the optimal portfolio weights are thus

$$\pi_t = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) \frac{W_t - f(t)}{W_t} = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) \left(1 - \frac{f(t)}{W_t} \right). \quad (8.26)$$

Note that with subsistence power utility the optimal portfolio weights are no longer independent of time nor independent of the wealth of the investor.

If we fix wealth W_t , the portfolio weights are increasing in time. The individual moves from the riskfree asset to risky assets over the life cycle. This contrasts the typical practitioner advice of “more stocks when you are young,” which is often motivated by the observation that stocks tend to outperform bonds over long horizons.

If we fix time t , the portfolio weights are increasing in wealth. An investor with a wealth of only $f(t)$ has to invest it all in the riskfree asset so the portfolio weights of the risky assets have to be zero. As wealth increases, the individual optimally invests a larger and larger fraction of wealth in the risky assets. This behavior is also typically seen empirically: more wealthy investors tend to have higher fractions of wealth invested in the stock market, cf., e.g., Guiso and Sodini (2013).

8.2.3 Consumption habits

So far we have assumed that the utility of consumption at one point in time is independent of the consumption at other points in time. An interesting alternative is the idea of *habit formation*: an individual having consumed a lot in the past demands higher consumption in the future. In other words, the utility of a certain level c_t of consumption right now is decreasing in the past consumption rates. Preferences with consumption habits have been studied by economists at least since Ryder and Heal (1973), and some empirical studies of consumer behavior find evidence of habit formation, cf., e.g., Carrasco, Labeaga, and Lopez-Salido (2005), Browning and Collado (2007), and Ravina (2019).

From a mathematical point of view, habit formation is very similar to the idea of subsistence consumption except that the minimum possible consumption at any time t is no longer a constant \bar{c} , but a time-varying level \bar{c}_t which is a measure of past consumption. The necessary buffer $f(t)$ is now more complicated to compute, but the form of the optimal consumption and investment strategy remains the same as above, as has been shown formally by Detemple and Zapatero (1992) and Munk (2008).

Concerning the consumption pattern over the life cycle, Kraft, Munk, Seifried, and Wagner (2017) show that by combining habit formation with a high time preference rate the optimal consumption may indeed have the hump shape reported by empirical studies. In the absence of habit formation, an impatient individual would prefer a decreasing consumption path over life. However, because of habit formation, a high initial consumption would lead to high required consumption in the future. To cover the future required consumption, wealth is set aside, but the necessary amount decreases with age which allows

consumption to increase in the early part of life. At some age, the impatience outweighs the habit concerns so that consumption starts to decrease.

8.3 Time-varying investment opportunities

The basic model of the previous section assumes constant investment opportunities, i.e., constant interest rates, expected rates of return, volatilities, and correlations. However, it is well-documented that these quantities vary over time in a stochastic manner. This situation is referred to as time-varying or stochastic investment opportunities.

The main effect of allowing investment opportunities to vary over time is easy to explain. Risk-averse investors with time-additive utility prefer a relatively stable consumption across possible states of the world. Therefore, a risk-averse individual chooses a portfolio with high positive returns in states with relatively bad future investment opportunities (or bad future labor income for that sake, cf. Section 9.1), and the individual is willing to give up some returns or some consumption in good times to achieve that protection in bad times. This mechanism is known as **intertemporal hedging**. Because of intertemporal hedging, the optimal investment strategy is different from the case with constant investment opportunities and therefore different from the optimal portfolio in the one-period mean-variance model.

Intertemporal hedging was first identified in rather abstract settings by Merton (1971, 1973a), and specific models with time-varying investment opportunities were not solved until many years later. As such models are quite complex, we do not go into computational details but present only some basic economic arguments and conclusions below.

Which aspects of investment opportunities do long-term investors care about? Think about the usual mean-variance diagram with the standard deviation of returns along the horizontal axis and the expected returns along the vertical axis, where returns are computed over some short period of time. A short-term investor is really only interested in the location of the *capital market line*, i.e., the straight line going through the point corresponding to the riskfree return over the period and the point corresponding to the tangency portfolio of risky assets. The line is characterized by the intercept and the slope, i.e., the riskfree rate and the Sharpe ratio of the tangency portfolio. If all investors agree on the inputs to the mean-variance analysis, the tangency portfolio has to be the market portfolio of the risky assets, say the stock market index if we focus on stocks. Long-term investors care about how the capital market line moves around over time so they care about variations in the short-term riskfree rate and in the Sharpe ratio of the tangency or market portfolio (see Nielsen and Vassalou (2006) for a formal proof hereof). Variations in the expected returns and standard deviations of the individual risky assets and in their pairwise correlations only matter to the extent that they cause fluctuations in the Sharpe ratio of the tangency or market portfolio.

8.3.1 Time-varying equity risk premium

As discussed in Chapter 6, the expected excess returns on individual stocks and stock indices vary over time. Merton's basic model of long-term investments assumes that returns in one period are statistically independent of returns in other periods, but some empirical studies have concluded that stock returns tend to exhibit short-term momentum and long-term reversal. Of course, the future returns are uncertain and there is no guarantee that the momentum or reversal happens at any particular occasion. What the studies suggest is that, more often than not, positive returns over a few weeks or months are followed by positive returns in the subsequent weeks and months, but followed by

negative returns a few years into the future. If returns indeed follow these patterns, what are the implications for optimal investments?

Momentum magnifies both the upside potential and the downside risk of stock investments in the short run. It also suggests that timing is important. With a short investment horizon, say up to one year, your optimal stock investment should be increasing in the stock returns in the recent past. Reversals reduce the risk of stock investments in the longer run. Should stocks fall in the near future, they tend to recover within a few years. The reversal effect is also referred to as mean reversion in stock prices or in stock returns.

A few papers have formulated extensions of Merton's basic investment model to capture the reversal effect, cf. [Kim and Omberg \(1996\)](#) and [Wachter \(2002\)](#). They consider a single risky asset representing the stock market index and assume that the expected return μ_t at any time t is correlated with realized past returns. A negative correlation implies return reversal since a period of positive returns are then followed by lower-than-average expected returns over the following period. Since the riskfree rate and the volatility of the stock index are still assumed constant, the variations in the expected return of the stock index translate directly into variations of the Sharpe ratio $\lambda_t = (\mu_t - r_f)/\sigma$. Of course, in order to apply such a model, you need to be able to precisely estimate the market Sharpe ratio at any given point in time, which is a questionable premise.

With empirically realistic parameter values the reversal feature leads only to a modest reduction in the risk of a long-run stock investment as indicated by Figure 8.4. Here the riskfree rate is assumed to be 1% and the stock index volatility 16%. The Sharpe ratio is assumed to be 0.25 at the moment and fluctuate around that level (corresponding to an expected return of 5%) with an annualized standard deviation of $\sigma_\lambda = 0.065$. Stock returns and changes in the expected stock return (or Sharpe ratio) over the next period are assumed to have a correlation of $\rho = -0.8$. With the specific modeling assumptions, the log-return on the stock is still normally distributed over any investment horizon. The difference between the return distribution with reversal or mean reversion and the return distribution without is small over a 5-year horizon (left panel), but over a 30-year horizon (right panel) it is clear that the reversal feature implies a return distribution which has slimmer tails (and thus lower risk) and more probability mass near the median. The reversal feature increases the probability of a full stock investment outperforming a full riskfree investment over any horizon, but again differences are modest except for long horizons.

In such a model of time-varying expected returns, [Kim and Omberg \(1996\)](#) find that the optimal investment strategy of an investor with CRRA utility of terminal wealth is

$$\pi_t = \frac{\lambda_t}{\gamma\sigma} + \frac{\gamma-1}{\gamma}(-\rho)\frac{\sigma_\lambda}{\sigma}(A_1(T-t) + A_2(T-t)\lambda_t), \quad (8.27)$$

where A_1 and A_2 are rather complicated functions. In the realistic case where $\gamma > 1$, both these functions are positive and increasing in the remaining time horizon $T - t$, and $A_1(0) = A_2(0) = 0$. If we compare (8.27) with Eq. (8.16) that states the optimal stock index weight in the case of constant investment opportunities, two differences are apparent. First, the constant weight becomes time varying since the Sharpe ratio λ_t now fluctuates over time. When the stock index has a high Sharpe ratio, you should invest a larger fraction of your wealth in the stock than when the index has a low Sharpe ratio. Secondly, the additional term in (8.27) reflects intertemporal hedging. This term shows how a longer-term investor optimally deviates from the portfolio optimal for a very short time horizon, which is the term $\lambda_t/(\gamma\sigma)$ familiar from the one-period mean-variance model.

As the remaining time horizon $T - t$ shrinks to zero, the intertemporal hedge term van-

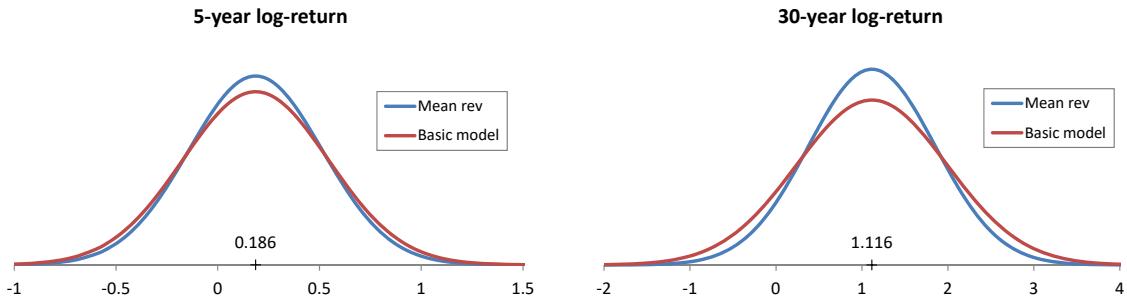


Figure 8.4: Return distribution with mean reversion.

The figure illustrate the effects of mean reversion on the distribution of the log-return on the stock index over T years. The graphs show the distribution of log-return in Merton's basic model without mean reversion (red solid curve) and the extended model with mean reversion (blue dotted curve). The number above the horizontal axis is the expected log-return.

ishes since investors with very short horizons do not care about the risk of bad investment opportunities in the future. The intertemporal hedging term also vanishes for investors with a relative risk aversion of $\gamma = 1$, which corresponds to the logarithmic utility function. But for investors with a higher risk aversion and a horizon longer than a single period, the intertemporal hedging term is present. With the negative correlation consistent with return reversals, we have $(-\rho)$ positive, and the standard deviations σ_λ and σ are also positive, of course. So for $\gamma > 1$, the intertemporal hedge term in (8.27) is positive. Hence the optimal weight of the stock index is higher than it would be if the Sharpe ratio would be constant at its current level. Moreover, the magnitude of the intertemporal hedging term and thus the total stock weight are increasing in the remaining time horizon.

Why is that so? One argument is that, due to the return reversal, the risk of a stock investment no longer grows proportionally with the investment horizon, but at a slower rate, as reflected by the probability distributions and outperformance probabilities presented above. Another argument is that stocks in the new model have a built-in hedge against bad times, i.e., periods with low expected returns. Should we end up in a period with low expected return and thus low Sharpe ratio, it often follows a period with high realized returns and thus with a relatively high wealth.

Figure 8.5 illustrates the optimal portfolio weight of the stock as a function of the investment horizon, both in Merton's basic model—where the weight is independent of the horizon—and in the model with return reversals, and for a relative risk aversion of 2 and 5. The figure is produced assuming reasonable parameter values and assuming that the current Sharpe ratio equals its long-run level. With the given numbers, for any investment horizon the optimal stock weight in the basic model would be 78.1% for a risk aversion of 2 and 31.3% for a risk aversion of 5. In the extended model (blue curves) the optimal stock weight is only slightly higher. Even for very long investment horizons of 40-50 years the stock weight should only be increased up to around 87% for a risk aversion of 2 and 37.4% for a risk aversion of 5. The return reversal has only a modest effect on the optimal portfolio. This should not come as a big surprise given the modest effect return reversals were shown to have on the return probability distribution. The mean reversion model is consistent with the common advice that young investors should have a higher portfolio weight on the stock market than old investors, but the model cannot justify large differences in portfolio weights across age groups.

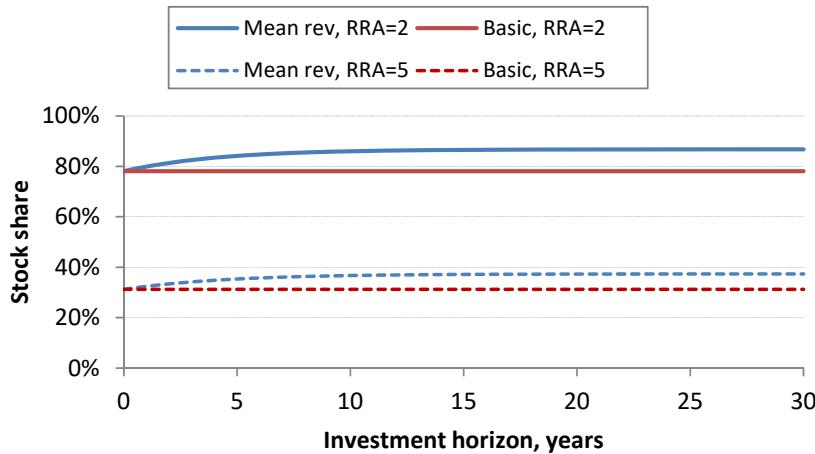


Figure 8.5: Mean reversion and stock weight.

The figure shows the optimal portfolio weight of the stock index as a function of the investment horizon. The red curves are from Merton's basic model without mean reversion, and the blue curves from the extended model with mean reversion. The solid curves are for a relative risk aversion of $\gamma = 2$, and the dashed curves for a relative risk aversion of $\gamma = 5$.

While the intertemporal hedging term has a modest size, it is important to use the current value of the Sharpe ratio in the mean-variance component of the optimal portfolio. Just applying (8.16) with a long-term average value of the Sharpe ratio can lead to severe mistakes if the current value of the Sharpe ratio is far away from its long-term average. Eq. (8.27) shows that the stock share π_t is linearly increasing in the Sharpe ratio λ_t . As the Sharpe ratio fluctuates, so does the optimal portfolio composition. By simulating possible paths of the Sharpe ratio, we get possible paths of the optimal stock share in the portfolio. The five paths shown in Figure 8.6 illustrate that the optimal stock share may fluctuate considerably over time.

The model assumes that the investor can directly observe the current Sharpe ratio of the stock market index. In practice, the investor has to estimate the Sharpe ratio. As discussed in Section 3.7.2 it is difficult to precisely estimate the expected return on the index—and thus also its Sharpe ratio—even if we assume it is constant over time. It seems even more difficult to estimate an expected return or a Sharpe ratio fluctuating over time. In the light of this uncertainty about the key model input, investors should probably be even more skeptical about letting the stock-bond allocation vary substantially over time because of hard-to-detect swings in Sharpe ratios.

Wachter (2002) considers a very similar model, but assumes that the individual has CRRA utility of consumption. This leads to intertemporal hedging demands that are even lower than shown above.

Koijen, Rodriguez, and Sbuelz (2009) extend the model described above to include the short-term momentum feature in addition to the (longer-term) return reversal feature. The model set-up and the solution for the optimal investment strategy are then more complex. The authors report that the intertemporal hedging term is now negative for fairly short investment horizons and positive for investment horizons longer than roughly 5 years, where the mean reversion feature begins to dominate.

In practice, Sharpe ratios or expected returns are not directly observable, so it might make more sense to set up a model in which these quantities depend on some directly

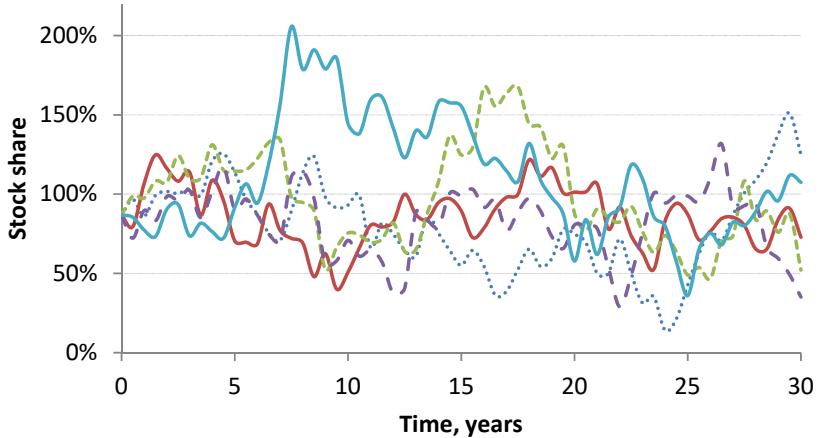


Figure 8.6: Mean reversion and stock weight.

The figure shows five simulated paths of how the optimal portfolio stock share varies over time. The investor initially has a 30-year investment horizon and has a relative risk aversion of $\gamma = 2$.

observable variable. As explained in Section 6.5, stock returns appear to be predictable by a number of variables such as the price-dividend or price-earnings ratio. Indeed, several academic papers have solved for optimal investments with a single stock (i.e., the index) when the expected stock return varies over time with the price-dividend ratio, see, e.g., Campbell and Viceira (1999) and Barberis (2000). The models are very similar to those described above and lead to similar conclusions. The higher the price-dividend ratio, the lower the expected return over the next period. An increase in the price-dividend ratio in one period thus leads to a decrease in the expected return over the next period. Holding the dividend fixed, this is exactly the return reversal idea again. The models rely on an estimated relation between expected returns and the predictor variable, but also this relation is difficult to pin down precisely based on historical observations, which again may cause skepticism in the models' conclusions about how the optimal portfolio should depend on the predictor.

In sum, many market practitioners and academic researchers seem to believe in some form of counter-cyclical variations in expected returns and Sharpe ratios on the stock market. In bad times, expected returns and Sharpe ratios over, say, the next year tend to be high. In good times, expected returns and Sharpe ratios tend to be low. Investors with a constant relative risk aversion can exploit this through market timing by having a higher stock weight in bad times than in good times. However, since it is extremely difficult to precisely quantify how expected returns and Sharpe ratios vary with the state of the economy, it is also difficult to pin down how much the stock weight should be varied over the business cycle. The available models show that variations in expected returns and Sharpe ratios do not generate a substantial horizon-dependence of the optimal portfolio.

8.3.2 Time-varying stock market volatility

The volatility of the stock market index and the volatilities of individual stocks vary over time. For example, Bloom (2009) shows that the volatility of the S&P 500 index over the period from 1960 to 2010 has fluctuated between about 10% in some calm periods to about 50% during the 2008 credit crunch and in the days following the Black Monday

stock market crash in 1987. So at a given point in time, the current stock market volatility can be far from the long-term average volatility of 15-20%. How does that influence the optimal investment strategy?

If the expected excess stock return $\mu_t - r_f$ would simply be proportional to the current volatility, the Sharpe ratio of the stock $\lambda = (\mu_t - r_f)/\sigma_t$ would still be constant. If we continue to assume a constant riskfree rate, the capital market line would then never move, and long-term investors would not engage in intertemporal hedging. The optimal portfolio weight of the stock index would then be

$$\pi_t = \frac{1}{\gamma} \frac{\lambda}{\sigma_t},$$

cf. Eq. (8.16), and thus still vary over time. If the stock market volatility goes up, while the Sharpe ratio stays the same, you should reduce the fraction of wealth invested in the stock market to maintain the same overall riskiness of your portfolio. But there would be no difference between how a short-horizon investor and a long-horizon investor would react to changes in the volatility, as long as they have the same risk aversion coefficient.

If variations in stock market volatility induce variations in the Sharpe ratio of the stock market, long-horizon investors are engaging in intertemporal hedging. Precisely how depends on the assumed link between the volatility and the Sharpe ratio, as well as the dynamics of the volatility. Which asset would investors then use for the hedging activities? Suppose the Sharpe ratio of the stock index is proportional to the volatility. Empirical estimates of the correlation between the stock and its instantaneous variance are negative. Volatility tends to go up when stock prices go down and vice versa. Consequently, volatility risk can be hedged by investing more in the stock index. A low volatility represents a situation of bad investment opportunities since the Sharpe ratio is then also low. Due to the negative correlation, stocks have a built-in hedge: should investment opportunities deteriorate (falling variance), the stock will typically increase substantially in price. For details, see the specific models of [Liu \(1999, 2007\)](#) and [Kraft \(2005\)](#).

Another possibility is to hedge volatility risk by investing in options on the stock market index. As we shall see in Chapter 15, the price of an option is highly dependent on the volatility of the underlying asset, so if the volatility of the stock market index changes, so does the price of an option on the index. The main model of option pricing is the Black-Scholes-Merton model, which we will introduce in Section 15.6, but that model assumes a constant volatility. More advanced models of option pricing that allow for stochastic volatility are necessary to use if we want to figure how to invest in options in order to hedge volatility risk. For an example of a specific stochastic volatility model of optimal investments in stocks and options, the courageous reader is referred to [Liu and Pan \(2003\)](#). For reasonable parameter values, the hedge component of the optimal portfolio seems to be small as in the case with a time-varying equity premium. Hence, there is not much difference between the optimal investments of the long-term power-utility agent and the agent optimizing the mean-variance tradeoff period by period.

8.3.3 Time-varying interest rates

Interest rates vary over time in a manner which is not perfectly predictable. Figure 5.4 shows how the three-month, one-year, and ten-year interest rates in the United States developed over the period 1954–2021. Interest rates of all maturities vary over time.

The intercept of the capital market line in the usual standard deviation-mean diagram is the riskfree rate over the same period for which the standard deviations and means in

the diagram are computed. For an investor rebalancing her portfolio every month, the yield on a one-month government bond would be a reasonable value to use for the riskfree rate. A month from now, the yield on a one-month government bond might be different from the yield on a one-month government bond right now. No matter the rebalancing frequency, the corresponding riskfree rate does fluctuate over time.

Long-term investors may want to hedge against variations in the intercept of the capital market line, i.e., the short-term interest rate. Other things equal, a decline in the short-term interest rate represents a deterioration of investment opportunities. Hence, long-term risk-averse individuals would consider tilting the portfolio towards assets that provide high returns should the short-term interest rate fall. The downside is that such a tilt lowers returns should the short-term interest rate increase, but individuals are willing to accept this in order to get the protection against a severe decline in wealth or consumption in the bad states of the world. Bonds have this property: if interest rates fall, bond prices increase leading to capital gains. Consequently, investors can hedge against interest rate changes by investing more in long-term bonds than they would have done just considering the expectation and standard deviation of the bonds' returns over the next short period.

Several academic papers have extended Merton's basic investment model of Section 8.1 to the case of time-varying interest rates. These extended models have to make specific assumptions about how interest rates can vary over time and how changes in interest rates are correlated with changes in stock prices.

The first specific model of this type was considered by Sørensen (1999). As in the basic model, the investor is assumed to have CRRA utility of wealth at some future date T with a relative risk aversion of γ . Sørensen assumes that the investor can invest in a stock index, a zero-coupon government bond maturing at time T , and a money market account (sometimes referred to as cash) that at any point in time gives a return equal to the then prevailing short-term interest rate. The short-term interest rate is assumed to fluctuate around a certain long-run level so that it always tend to move closer to the long-run level (a feature known as mean reversion), but shocks to the economy might pull the short-term interest rate further away from the long-run level.

Sørensen derives the optimal investment strategy under these assumptions. Let λ_S and λ_B denote the Sharpe ratios of the stock index and the bond, and let ρ denote the correlation between changes in the stock index and the bond price. Then the fraction of wealth optimally invested in the stock index is

$$\pi_t^S = \frac{\lambda_S - \rho\lambda_B}{\gamma\sigma_S(1 - \rho^2)}, \quad (8.28)$$

which is still independent of time, wealth, and the levels of the stock index and the short-term interest rate. The fraction of wealth optimally invested in the long-term bond is

$$\pi_t^B = \frac{\lambda_B - \rho\lambda_S}{\gamma\sigma_B(t)(1 - \rho^2)} + 1 - \frac{1}{\gamma}, \quad (8.29)$$

where $\sigma_B(t)$ is the volatility of the bond price at time t . The remaining fraction of wealth, $1 - \pi_t^S - \pi_t^B$, is invested in the money market account.

If we compare with (7.58), we see that the optimal stock weight above is exactly as it would be for a short-horizon investor, i.e., the stock is not used for intertemporal hedging against interest rate fluctuations. The first term of the optimal bond weight is the mean-variance component, i.e., the optimal bond weight of a short-term investor. The second term of the optimal bond weight is the intertemporal hedging term. The bond is used

γ	tangency	bond	stock	cash	exp. return	volatility
0.5	5.03	1.44	3.59	-4.03	0.296	0.758
1	2.51	0.72	1.79	-1.51	0.153	0.379
2	1.26	0.36	0.90	-0.26	0.081	0.189
2.515	1.00	0.29	0.71	0.00	0.066	0.151
3	0.84	0.24	0.60	0.16	0.057	0.126
5	0.50	0.14	0.36	0.50	0.038	0.076
20	0.13	0.04	0.09	0.87	0.016	0.019
50	0.05	0.01	0.04	0.95	0.012	0.008

Table 8.2: Ignorant's stock-bond-cash portfolio.

The table shows the stock-bond-cash portfolio weights for CRRA investors ignoring interest rate fluctuations. Each row assumes a given relative risk aversion parameter γ as shown in the first column. The other columns show the optimal weight in the tangency portfolio, the corresponding weights in the bond, the stock, and in cash, as well as the expected return and volatility of the optimal portfolio.

for hedging interest rate variations because bond prices are perfectly negatively correlated with interest rate variations (at least under the assumptions of the model). In contrast, the stock index is only partially correlated with interest rate movements and thus not as effective a hedging instrument as the bond.

We see that the log utility investor ($\gamma = 1$) does not hedge, but sticks to the “myopic” portfolio, that is the portfolio that is optimal in the very short run. The hedge position of a less risk-averse investor ($\gamma < 1$) is negative, while a more risk-averse investor ($\gamma > 1$) takes a long position in the bond in order to hedge interest rate risk. An infinitely risk-averse investor ($\gamma \rightarrow \infty$) will invest her entire wealth in the zero-coupon bond maturing at T , which makes sense again because this bond is the truly riskfree asset for the investor. By investing money in this bond today, she knows exactly what she gets back at time T .

The long-term bond does not have to be the zero-coupon bond maturing at the end of the investor’s horizon. Any other government bond would do. Under the assumptions of the model, all bonds have the same Sharpe ratio and the same correlation with the stock index. Different bonds have different volatilities, however. Since the volatility measures the standard deviation of returns over the next short period, long-maturity bonds have larger volatilities than short-term bonds. If we invest in a different bond than the zero-coupon bond maturing at T , we have to adjust the expression for the optimal bond weight. Of course, $\sigma_B(t)$ should now be the volatility of the bond we invest in. Moreover, the intertemporal hedge term $(1 - [1/\gamma])$ has to be scaled by the ratio of the volatility of the zero-coupon bond maturing at time T to the volatility of the bond we invest in. As the fraction of wealth invested in the money market account is residually determined, it will also change when we change the maturity of the bond we invest in.

Example 8.1

In Example 7.5 we computed the tangency portfolio of the stock index and a long-term bond based on historically reasonable inputs. Focus on the case with a stock-bond correlation of 0.2. The tangency portfolio consists of 71.3% in the stock index and 28.7% in the bond. Let us now assume that the input values are appropriate over very short time intervals. In Figure 7.5 the solid blue curve shows the mean-variance efficient portfolios of

risky assets, i.e., the combinations of expected returns and volatility that can be obtained by combining the bond and the stock. The solid orange line corresponds to the optimal portfolios for investors assuming constant investment opportunities with an interest rate equal to the long-term average.

We know from Theorem 8.3 that CRRA investors ignoring interest rate risk choose a portfolio of risky assets given by $\boldsymbol{\pi} = \frac{1}{\gamma} [\mathbf{1} \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r\mathbf{1})] \boldsymbol{\pi}^{\text{tan}}$, where γ is the relative risk aversion and r is the current short-term interest rate. The portfolio is independent of the investment horizon. Table 8.2 shows the portfolio allocation for various γ -values. The numbers in the column “tangency” represent the fraction of wealth invested in the tangency portfolio, which is then split into the bond and the stock in the following two columns. The cash position (i.e., the amount deposited at the short-term interest rate) is determined residually so that weights sum to one. The last two columns show the instantaneous expected rate of return and volatility of the portfolio.

Now let us look at investors who realize that interest rates vary over time and consequently alter their investment strategy (except for log-utility investors). We assume they invest in a 10-year zero-coupon government bond regardless of their investment horizon. The short-term interest rate is assumed to fluctuate around a long-run level of 1%. The parameters of the interest rate dynamics are chosen so that the volatility of a 10-year zero-coupon bond according to the model equals the historical estimate of 10.3%. The current short rate is assumed to equal the long-term level of 0.9%.

Table 8.3 shows the optimal portfolios for CRRA investors with different combinations of risk aversion and investment horizon. The column ‘hedge’ shows the hedge demand for the 10-year zero-coupon bond which the investors are allowed to trade in. While the weight on the tangency portfolio and thus the stock is independent of the investment horizon, this is not true for the weight on the hedge portfolio and hence not true for the total weight on the bond and on cash. The ratio of the bond weight to the stock weight is shown in the column ‘bond/stock’. For investors with risk aversion above one, the optimal bond weight and the bond-stock ratio is clearly increasing in the investment horizon when all investors invest in the same bond. Longer-term investors choose portfolios with higher volatility, i.e. they take on more short-term risk, but the main point is that long-term investors do not choose their portfolio according to the short-term risk/return trade-off. It is important to emphasize that the portfolio weights on the bond and thus the bond/stock ratio will depend on the maturity (and payment schedule) of the bond, the investor is trading in. In particular, a recommendation of a particular bond weight or bond/stock ratio should always be accompanied by a specification of what bond the recommendation applies to.

Next we compare the current mean/variance tradeoff chosen by different investors. As discussed above, CRRA investors that either have a zero (or very, very short) horizon or do not take interest rate risk into account pick a portfolio that corresponds to a point on the straight line in Figure 8.7. This is the instantaneous mean-variance efficient frontier. Similarly, each of the other two curves corresponds to the combinations chosen by CRRA investors with a given non-zero horizon (one-year or 30-year horizon) who take interest rate risk into account. Since these curves lie to the right of the instantaneous mean-variance frontier, all these investors could obtain a higher instantaneous expected rate of return for the same volatility by choosing a different portfolio. But the long-term investors are willing to sacrifice some expected return in the short term in order to hedge changes in interest rates and place themselves in a better position if interest rates should decline.

γ	tangency	hedge	bond	stock	$\frac{\text{bond}}{\text{stock}}$	cash	exp. return	volatility
$T = 1$; bond volatility 2.6%								
0.5	5.029	-0.255	1.186	3.588	0.331	-3.774	0.289	0.748
1	2.515	0.000	0.721	1.794	0.402	-1.515	0.151	0.379
2	1.257	0.127	0.488	0.897	0.544	-0.385	0.082	0.195
5	0.503	0.204	0.348	0.359	0.970	0.293	0.040	0.086
10	0.251	0.229	0.302	0.179	1.681	0.519	0.027	0.052
20	0.126	0.242	0.278	0.090	3.102	0.632	0.020	0.037
$T = 10$; bond volatility 10.3%								
0.5	5.029	-1.000	0.441	3.588	0.123	-3.029	0.281	0.724
1	2.515	0.000	0.721	1.794	0.402	-1.515	0.151	0.379
2	1.257	0.500	0.860	0.897	0.959	-0.757	0.086	0.215
5	0.503	0.800	0.944	0.359	2.631	-0.303	0.047	0.132
10	0.251	0.900	0.972	0.179	5.418	-0.151	0.034	0.113
20	0.126	0.950	0.986	0.090	10.992	-0.076	0.028	0.107
$T = 30$; bond volatility 11.0%								
0.5	5.029	-1.070	0.371	3.588	0.103	-2.959	0.280	0.723
1	2.515	0.000	0.721	1.794	0.402	-1.515	0.151	0.379
2	1.257	0.535	0.895	0.897	0.998	-0.792	0.087	0.217
5	0.503	0.856	1.000	0.359	2.788	-0.359	0.048	0.137
10	0.251	0.963	1.035	0.179	5.770	-0.215	0.035	0.119
20	0.126	1.017	1.053	0.090	11.735	-0.142	0.029	0.113

Table 8.3: Optimal portfolios acknowledging stochastic interest rates.

The table shows the optimal portfolios of long-term investors with CRRA utility of terminal wealth for different combinations of the investment horizon T and the relative risk aversion γ . The bond volatility written next to the investment horizon refers to the zero-coupon bond maturing at the end of that horizon.

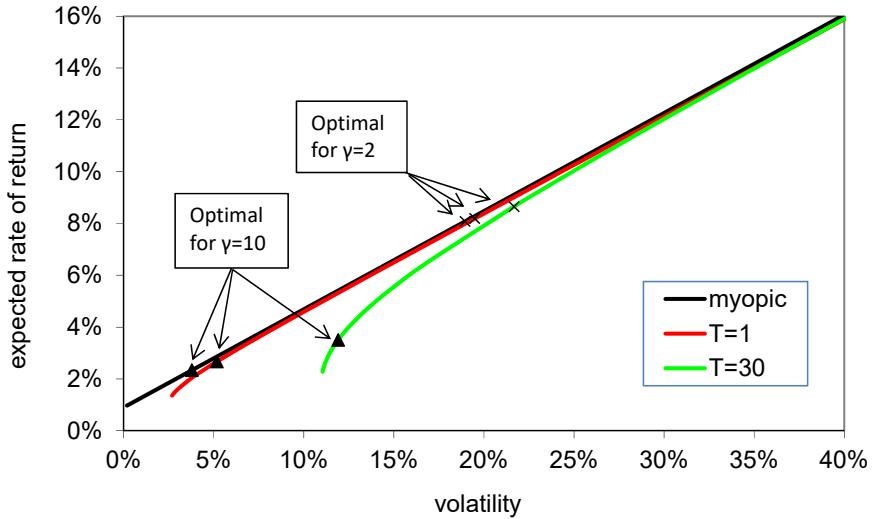


Figure 8.7: Optimal frontiers with stochastic interest rates.

Each curve contains the combinations of current expected rate of return and volatility for CRRA investors with a given investment horizon T . The black line represents the myopic case and is identical to the mean-variance frontier. The red curve is for $T = 1$ and the green for $T = 30$. The three points marked with an ‘x’ indicate the optimal choice when the relative risk aversion is $\gamma = 2$. The three points marked with a triangle indicate the optimal choice for $\gamma = 10$.

Brennan and Xia (2000), Deelstra, Grasselli, and Koehl (2000), Campbell and Viceira (2001), Munk and Sørensen (2004), Sangvinatsos and Wachter (2005), and Liu (2007), among others.

8.4 Conclusion

This chapter has extended the one-period portfolio choice analysis of Chapter 7 to a multi-period setting which is more realistic for most investors. The first part of the chapter showed that, under a specific set of assumptions, the optimal portfolio rule is more or less the same in the multi-period framework as in the one-period case. Among the underlying assumptions supporting this result is the assumption that the riskfree rate as well as the expected returns, volatilities, and correlations of risky assets stay constant over time which conflicts with empirical evidence. We saw that a relaxation of this assumption opens up for intertemporal hedging considerations and twists the optimal portfolio rules. For example, long-term investors may prefer to have a larger share of long-term bonds and stocks in their portfolio than short-term investors.

The subsequent Chapter 9 focuses on the investment decisions of individual or household investors. Besides being multi-period investors, they have special features that can significantly affect their optimal portfolio. An individual possess an intangible asset in terms of human capital, i.e., the present value of the individual’s future income. In addition, many individuals own a large tangible asset, namely the apartment or house they live in. As we shall see, the magnitude and risk characteristics of these assets matter for their optimal financial decisions.

Many investors—in particular large institutional investors—follow investment strategies that in some sense are designed to “beat the market”, i.e., provide large returns relative to the risk taken. Of course, to assess whether a strategy beats the market, we need to

quantify what the market would expect the return on the strategy to be. This is the subject of the two subsequent chapters. The basic model for determining the “fair” return for a given level of risk is the so-called Capital Asset Pricing Model (CAPM) which is the subject of Chapter 10. Numerous empirical studies claim to detect strategies that perform better than they should according to the CAPM, and many investors are thus tempted to follow such strategies. Examples are value/growth tilts, income investing strategies, market timing, momentum or trend-following strategies, reversal/contrarian strategies, “opposites attract” strategy, and the $1/N$ diversification strategy. Some of them are explained and discussed in the following chapters.

8.5 Exercises

Exercise 8.1. You are considering how to invest your retirement savings and have decided to use Merton’s basic model for long-term investment with a constant relative risk aversion of $\gamma = 4$. You estimate the (continuously compounded) riskfree rate to be $r_f = 0.01$ per year.

First, you consider investing only in the riskfree asset and the domestic stock market index. Your estimates of the parameters for the domestic stock market index are $\mu_1 = 0.09$ (expected return per year) and $\sigma_1 = 0.20$ (volatility or standard deviation per year).

- (a) What is your optimal combination of the riskfree asset and the domestic stock market index according to Merton’s model?
- (b) Suppose you invest in this optimal portfolio right now and do not trade in the financial market over the next month. Suppose that the return on the domestic stock market index over this month exceeds the riskfree return over the month. If you want to follow the optimal investment strategy, what should you then do:
 - (i) Purchase additional domestic stocks
 - (ii) Sell some of the domestic stocks you own
 - (iii) Do not trade at all

Explain your answer.

Next, you consider also investing in an exchange-traded fund (ETF) of stocks in emerging markets. The relevant parameters of this ETF are $\mu_2 = 0.13$ (expected return per year) and $\sigma_2 = 0.40$ (volatility or standard deviation per year).

- (c) Assume that the correlation between the domestic stock index and the emerging markets ETF is $\rho = 0.5$. What is your optimal combination of the riskfree asset, the domestic stock market index, and the emerging markets ETF according to Merton’s model? In particular, is the portfolio weight of the domestic stock index smaller or larger than in question (a)?
- (d) How does your answer to question (b) change if the correlation is $\rho = 0.75$ or $\rho = 0.85$? Comment on your results.

Inspired by the above questions you want to analyze more formally how the addition of a second risky asset affects the optimal portfolio weight of the first risky asset.

- (e) Show that the addition of the second risky asset leads to a larger optimal weight of the first risky asset if and only if

$$\frac{\mu_1 - r_f}{\sigma_1} > \frac{\mu_2 - r_f}{\rho\sigma_2}.$$

Exercise 8.2. After careful introspection, you have realized that you have CRRA utility of terminal wealth with a relative risk aversion of $\gamma = 4$. You have decided to invest only in the stock market index (via an ETF) and a riskfree asset following Merton’s basic model for long-term investments. The riskfree log-return is $r_f = 2\%$. The volatility of the stock market index is $\sigma = 20\%$. Your estimate of the log expected stock return is $\mu = 10\%$.

- (a) What is your optimal portfolio if your investment horizon is $T = 10$ years? Or $T = 30$ years?

- (b) Suppose you have mis-estimated the expected stock return and that the true value of μ is really 8%. What is then the truly optimal portfolio for you? In terms of the percentage wealth loss defined in (8.20), how much are you losing by using your wrong μ -estimate if $T = 10$ and if $T = 30$?
- (c) Answer the preceding question again assuming that the true μ is 12%.

CHAPTER 9

Household portfolio choice

The topic of this chapter is the financial decisions of individual persons or households. An individual should realize that he or she may live for many years and is thus to be considered as a multi-period investor. Therefore, the ideas and results introduced in the previous chapter are also relevant for household investors. This chapter outlines some special characteristics of household investors and discusses how these characteristics should affect the financial decisions of such investors.

Most household investors receive an income stream from labor market activities, welfare transfers from the government, etc. The human capital of an individual is the present value of the income the individual is going to receive in her remaining lifetime. Most young individuals have a human capital that significantly exceeds their current financial wealth. Intuitively, the human capital can have a large effect on how the individual should invest her financial wealth. In the special case where the human capital can be considered riskfree, it is like having a large portfolio of riskfree bonds that offer regular payouts. With such a large indirect position in riskfree assets, it makes sense to invest the financial wealth primarily in risky assets, e.g., the stock market. On the other hand, if the income and thus human capital would be highly correlated with the stock market, the individual should probably invest the financial wealth in riskfree assets. Section 9.1 provides details and presents models for calculating the human capital of an individual and for determining the implications on the financial portfolio.

Many household investors own the home they live in, and the home is often the most valuable tangible asset of the household. Therefore, the optimal financial portfolio of a homeowner can depend on the risk characteristics of the home value, e.g., the expected changes and the volatility of house and apartment prices, the correlations of these prices with the stock market and with human capital, etc. Section 9.2 presents some historical facts about home prices and discusses the implications of owner-occupied housing for financial portfolio decisions.

The decision on how to allocate the financial wealth to different financial assets is not the only important financial decision of an individual person. The individual enters each period with some financial wealth from last period, hopefully with a nice return, and with some additional income received last period. The individual then has to decide how much of this disposable wealth she would like to consume in the coming period with the remaining disposable wealth being saved and invested in financial assets. Consumption

this period increases the immediate utility of the individual, whereas savings can generate additional consumption and utility in future periods. In particular, individuals like to smooth consumption over their life cycle: they tend to save and accumulate wealth when earning income in the labor market so that they can decumulate wealth in retirement to finance consumption exceeding the typically relatively low pension from the government. Section 9.3 presents some simplified calculations on how much of your income to save for retirement. Finally, Section 9.4 concludes.

9.1 Labor income

The analysis of portfolio decisions in the two preceding chapters has ignored a key determinant of individuals' decisions over the life cycle: labor income. When starting to make consumption and investment choices, the financial wealth an individual has is typically small compared to the human capital of the individual, i.e., the present value of future income.

Here is a back-of-the-envelope computation. Suppose the individual expects to earn \$50,000 each year (after income taxes) over a 40-year period until retirement. Assuming a discount rate of 5%, the present value of the life-time income is

$$\sum_{t=1}^{40} \$50,000 \times (1.05)^{-t} = \$50,000 \times \frac{1 - (1.05)^{-40}}{0.05} \approx \$857,954.$$

Here we have even disregarded the increase in income individuals often experience at least up to an age around 50 years, after which income tends to flatten out or even decrease slightly until retirement (see below). It is not obvious what the appropriate discount rate is, and the present value is quite sensitive towards it. With a discount rate of 2%, the present value is \$1,367,774. With a discount rate of 10%, it is \$488,953. In any case, this is far more than the financial wealth held by a typical young individual.

The magnitude and risk characteristics of human capital are therefore likely to have a significant impact on the optimal consumption and investment decisions of the individual. Before we go into formal modeling, let us illustrate the effects by a numerical example.

9.1.1 A motivating example

The following example is inspired by Jagannathan and Kocherlakota (1996). Assume that investment opportunities are constant and that a single risky financial asset (representing the stock market index) is traded. With constant interest rates the riskfree asset is equivalent to any government bond. Consider an investor with a financial wealth of \$50,000 and a risk aversion of $\gamma = 2$. Assume that the riskfree rate is $r = 4\%$, the expected return on stocks is $\mu = 10\%$, and the volatility of the stock is $\sigma = 20\%$. The Sharpe ratio is thus $\lambda = (\mu - r)/\sigma = 0.3$. We know from Section 8.1.3 that, in the absence of labor income, it is optimal for the investor to have 75% of her wealth invested in stocks and 25% in the riskfree asset, i.e., the bond. When the investor receives labor income it seems fair to conjecture that she will invest her financial wealth such that the riskiness of her total position corresponds to investing 75% of her total wealth in stocks.

First assume that the investor has a labor income stream with a present value of \$50,000 and, hence, a total wealth of \$100,000. It is then optimal to have a total position of \$75,000 in stocks and \$25,000 in the riskfree asset. How the financial wealth is to be allocated depends on the riskiness of her income. The left part of Table 9.1 considers three cases:

- (a) If the labor income is completely riskfree, it is equivalent to a position of \$0 in

	Short horizon		Long horizon	
	Stock invest	Bond invest	Stock invest	Bond invest
<i>Riskfree income</i>				
Human capital	0 (0%)	50,000 (100%)	0 (0%)	450,000 (100%)
Financial inv.	75,000 (150%)	-25,000 (-50%)	375,000 (750%)	-325,000 (-650%)
Total position	75,000 (75%)	25,000 (25%)	375,000 (75%)	125,000 (25%)
<i>Modestly risky income</i>				
Human capital	25,000 (50%)	25,000 (50%)	225,000 (50%)	225,000 (50%)
Financial inv.	50,000 (100%)	0 (0%)	150,000 (300%)	-100,000 (-200%)
Total position	75,000 (75%)	25,000 (25%)	375,000 (75%)	125,000 (25%)
<i>Very risky income</i>				
Human capital	50,000 (100%)	0 (0%)	450,000 (100%)	0 (0%)
Financial inv.	25,000 (50%)	25,000 (50%)	-75,000 (-150%)	125,000 (250%)
Total position	75,000 (75%)	25,000 (25%)	375,000 (75%)	125,000 (25%)

Table 9.1: Human capital and investments.

The table shows the optimal investment strategy for all combinations of two levels of human capital and three types of labor income risk. The financial wealth is \$50,000. In the left (right) part of the table, the human capital is \$50,000 (\$450,000) corresponding to a relatively short (long) investment horizon.

stocks and \$50,000 in the riskfree asset. To obtain the desired overall riskiness, she has to allocate her financial wealth of \$50,000 by investing \$75,000 in stocks and -\$25,000 in the riskfree asset. This corresponds to a stock investment of 150% of the financial wealth, financed in part by borrowing 50% of the financial wealth. The certain labor income corresponds to the returns of a riskfree investment. Hence the financial wealth (and more) has to be invested in stocks to achieve the desired balance between risky and riskfree returns.

- (b) If the labor income is quite risky and corresponds to an equal combination of stocks and bonds, the entire financial wealth (100%) is to be invested in stocks.
- (c) If the labor income is extremely risky and corresponds to a 100% investment in stocks, the financial wealth is to be split equally between stocks and bonds.

Clearly, the optimal allocation of wealth depends on the risk profile of labor income.

Next, let us consider an investor with the same risk aversion, but a longer horizon and, consequently, a higher capitalized labor income, namely \$450,000. The right part of Table 9.1 shows the allocation of the financial wealth needed to obtain the desired 75-25 split between risky and riskfree returns. Comparing with the left part of the table, we see that the younger investor has a significantly higher fraction of financial wealth invested in stocks than the older investor, except for the case where the income is completely stock-like. The optimal stock weight in the portfolio is clearly depending on the investment horizon.

According to empirical studies, the correlation between labor income and the stock market index is very small for most individuals.¹ In that case, labor income resembles

¹Davis and Willen (2000) find that – depending on the individual's sex, age, and educational level – the correlation between aggregate stock market returns and labor income shocks is between -0.25 and 0.3, while the correlation between industry-specific stock returns and labor income shocks is between -0.4 and 0.1. Campbell and Viceira (2002) report that the correlation between aggregate stock market returns and labor income shocks is between 0.328 and 0.516. Heaton and Lucas (2000) find that the labor income of

a riskfree investment more than a stock investment, and the fraction of financial wealth invested in stocks should increase with the length of the investment horizon—in line with typical investment advice. However, for some investors the labor income may be highly correlated with the stock market, or at least some individual stocks, and in that case the weight of stocks in the financial portfolio should decrease with the length of the horizon.

9.1.2 The size of human capital

How large is human capital at different stages of life? Generally, we can think of the human capital at a given point in time as the sum of the discounted expected after-tax income over the remaining life. Assume that income is paid out at the end of each year, let I_s denote the after-tax income in year s , and let r_s denote the appropriate annualized discount rate for income in year s . Then the human capital at the end of year t —excluding the income I_t just received—is

$$L_t = \sum_{u=1}^T E_t[I_{t+u}] (1 + r_{t+u})^{-u}, \quad (9.1)$$

where the income stream is assumed to continue for T more years. In line with the back-of-the-envelope calculation in the beginning of this section, suppose first that you expect that after-tax income remains constant and equal to the most recent income of I_t and that the discount rate is a constant r . Then the human capital is simply the present value of an annuity, i.e.,

$$L_t = I_t \frac{1 - (1 + r)^{-T}}{r}. \quad (9.2)$$

We can easily incorporate a fixed growth rate of g in expected annual income. In this case, the human capital is

$$L_t = \begin{cases} T \times I_t, & \text{if } g = r, \\ I_t \frac{1+g}{r-g} \left[1 - \left(\frac{1+g}{1+r} \right)^T \right], & \text{if } g \neq r. \end{cases} \quad (9.3)$$

The calculation is similar to that for the price-dividend ratio in Section 6.1. We could even distinguish between periods of life with different income growth rates.

What do we know about typical life-cycle patterns in labor income? Data can be found, for example, in the Survey of Consumer Finances (SCF) in the United States. In U.S. data, labor income is generally found to be hump shaped over working life: rapidly increasing in early years, then flattening out with a peak at an age of 45–55, and then declining somewhat until retirement. Moreover, the life-cycle pattern over the working phase is well approximated by the exponential of a polynomial of order 3 or slightly higher. For example, these facts have been shown for various U.S. data sets by Attanasio and Weber (1995), Cocco, Gomes, and Maenhout (2005), and Guvenen, Karahan, Ozkan, and Song (2021). The income in retirement often stems from various pension saving schemes and government benefits and is more or less constant over time. The red curve in the left panel of Figure 9.1 shows the typical income pattern over the life cycle. Here the annual income at the age of 25 is fixed at 15 (thousands of U.S. dollars, after tax), which seems reasonable given the median before-tax family income of 35.1 for the age group “less than 35” in the 2010 SCF, cf. Bricker, Kennickell, Moore, and Sabelhaus (2012, Table 1). After

entrepreneurs typically is more highly correlated with the overall stock market (0.14) than with the labor income of ordinary wage earners (-0.07).

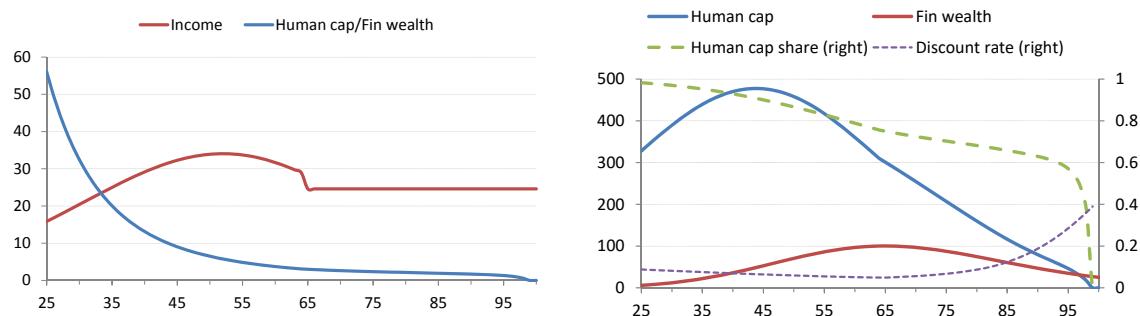


Figure 9.1: Income and wealth over the life cycle.

The left panel shows how expected income in thousand USD (red curve) and the ratio of expected human capital to financial wealth (blue) vary with age measured in years. In the right panel the expected human capital (blue) and the financial wealth (red) as functions of age are measured on the left axis in thousands of USD, whereas the expected human capital's share of total wealth (dashed) and the income discount rate (dotted) as functions of age are measured on the right axis. See [Munk \(2020\)](#) for details.

the assumed retirement age of 65, the income is equal to a constant state pension such as Social Security in the U.S. For details on the construction of the figure, see [Munk \(2020\)](#).

What is the appropriate discount rate for future income? Surely, it depends on the riskiness of the labor income, including the risk that the individual passes away and therefore does not receive the future income. Labor income tends to be more volatile for young workers than old workers, which suggests that the discount rate should decline with age. On the other hand, mortality risk is increasing in age, which works in the other direction. The dotted line in the right panel of Figure 9.1 (use right axis) shows an estimated life-cycle profile of the discount rate, which is first slightly decreasing in age, but increases relatively rapidly in retirement due to the substantial increase in mortality risk (again, see [Munk \(2020\)](#) for details). The blue curve in the right panel depicts the resulting human capital, which starts out at around 330,000 USD, grows for some years because of the high early income growth rates, and then starts to decline because of the fewer remaining years of income—and eventually also due to the decrease in expected income and the increase in the discount rate.

For many individuals the financial wealth also has a hump-shaped pattern over life. Young individuals save to build up a buffer against bad times or to finance sizeable future expenditures (down payment on house or apartment, college tuition for kids) or for their retirement. Financial wealth typically increases until (or even some years after) retirement, after which individuals finance consumption exceeding the pension income by reducing their savings. Some individuals leave a bequest. This typical pattern is also well approximated by the exponential of a third-order polynomial. Based on numbers from the SCF, the red curve in the right panel illustrates the financial wealth over the life cycle of a median U.S. individual.

Continuing the quantitative exercise, the dashed curve in the right panel shows that the human capital's share of total wealth starts at around 98%, remains above 90% until age 46, above 80% until age 59, above 70% until age 76, and above 60% until age 94. The blue curve in the left panel shows the ratio of human capital to financial wealth, which starts at around 55 and declines smoothly over life.

There is a large variation in income and wealth paths across individuals as, e.g., indicated by the huge difference between means and medians of income and net worth at different

ages in the SCF data (Bricker et al. 2012). Of course, if we scale either income or financial wealth up or down and fix the other, the human capital's share of total wealth changes. However, across individuals, income and wealth often move together since higher-earning individuals tend to build more wealth. Hence, we expect less cross-sectional variation in the human capital's share of total wealth than seen in income or wealth.

9.1.3 The extended mean-variance model

In Merton's basic multi-period investment model of Section 8.1, the optimal portfolio is the same as in the much simpler mean-variance model with a slight reinterpretation of the inputs. If we want to explore the effects of labor income on optimal investments, it therefore makes sense to start by extending the mean-variance framework to the case of labor income. The following presentation is based on Munk (2020), who provides more details and discussions.²

Let F denote the financial wealth and L the human capital ("L" for labor income) of the agent so that total wealth is the sum $W = F + L$. The agent makes a buy-and-hold investment decision for a period of a given length. The current date is labeled as time t and the end of the period is labeled as time $t + 1$. The agent has the mean-variance objective

$$\max \left\{ E_t \left[\frac{W_{t+1}}{W_t} \right] - \frac{\gamma}{2} \text{Var}_t \left[\frac{W_{t+1}}{W_t} \right] \right\}, \quad (9.4)$$

where $E_t[\cdot]$ and $\text{Var}_t[\cdot]$ denote the expectation and variance conditional on the information available at time t . The parameter $\gamma > 0$ is the relative risk aversion. This is similar to (7.54), except that the agent now cares about the return on *total wealth*, not just on financial wealth.

Suppose that the agent can invest in a riskfree asset with rate of return r_f over the period and in a number of risky assets with rates of return given by the vector \mathbf{r}_{t+1} . The expected rates of return are represented by $\boldsymbol{\mu}$ and the variance-covariance matrix by $\underline{\Sigma}$. Let $\boldsymbol{\pi}_t$ denote the vector of fractions of financial wealth invested in the risky assets. The financial wealth not invested in the risky assets, $F_t(1 - \boldsymbol{\pi}_t \cdot \mathbf{1})$, is invested in the riskfree asset. We assume that $\boldsymbol{\mu} \neq r_f \mathbf{1}$ and that $\underline{\Sigma}$ is a positive definite matrix.

The return on total wealth depends both on the return on the financial investments and the return on the human capital. We let r_{t+1}^L denote the rate of return on human capital over the period. The return on human capital comes from both the labor income received during the period (like a dividend) and the possible reassessment of the present value of future labor income. The value at time $t + 1$ of the human capital including the labor income received over the period is $L_t(1 + r_{t+1}^L)$. The variance of the return on total wealth is, of course, depending on the variance on the financial returns and the variance on the human capital return, as well as the covariance between these returns. We let $\text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L]$ denote the vector of covariances between the returns on the individual risky assets and the return on human capital. We have the following key result:

²Mayers (1972) derives an equation for the optimal financial portfolio of a mean-variance investor with a nonmarketable asset such as human capital. While very similar to the equation we derive below, his equation does not directly show the importance of the relative size of human capital to financial wealth, and he does not consider the implications for life-cycle portfolio decisions.

Theorem 9.1

In the extended mean-variance model described above, the time t portfolio maximizing (9.4) is also maximizing

$$f(\boldsymbol{\pi}_t) = \boldsymbol{\pi}_t \cdot (\boldsymbol{\mu} - r_f \mathbf{1}) - \frac{\gamma}{2} \left((1 - h_t) \boldsymbol{\pi}_t \cdot \underline{\Sigma} \boldsymbol{\pi}_t + 2h_t \boldsymbol{\pi}_t \cdot \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L] \right), \quad (9.5)$$

where $h_t = L_t/(F_t + L_t)$ is the human capital's share of total wealth.

Without any constraints on the portfolio $\boldsymbol{\pi}_t$ that can be chosen, the solution is

$$\boldsymbol{\pi}_t = \frac{1}{\gamma} (1 + \ell_t) \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) - \ell_t \underline{\Sigma}^{-1} \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L], \quad (9.6)$$

where $\ell_t = L_t/F_t = h_t/(1 - h_t)$ is the human-to-financial wealth ratio.

Proof

The rate of return on the financial investment is

$$\frac{F_{t+1}}{F_t} - 1 = (1 - \boldsymbol{\pi}_t \cdot \mathbf{1}) r_f + \boldsymbol{\pi}_t \cdot \mathbf{r}_{t+1} = r_f + \boldsymbol{\pi}_t \cdot (\mathbf{r}_{t+1} - r_f \mathbf{1})$$

so that

$$F_{t+1} = F_t (1 + r_f + \boldsymbol{\pi}_t \cdot (\mathbf{r}_{t+1} - r_f \mathbf{1})).$$

The end-of-period total wealth is therefore

$$W_{t+1} = F_t (1 + r_f + \boldsymbol{\pi}_t \cdot (\mathbf{r}_{t+1} - r_f \mathbf{1})) + L_t (1 + r_{t+1}^L),$$

where r_{t+1}^L is the rate of return on human capital with expectation μ_L and standard deviation σ_L . Dividing by $W_t = F_t + L_t$, we obtain

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \frac{F_t}{F_t + L_t} (1 + r_f + \boldsymbol{\pi}_t \cdot (\mathbf{r}_{t+1} - r_f \mathbf{1})) + \frac{L_t}{F_t + L_t} (1 + r_{t+1}^L) \\ &= (1 - h_t) (1 + r_f + \boldsymbol{\pi}_t \cdot (\mathbf{r}_{t+1} - r_f \mathbf{1})) + h_t (1 + r_{t+1}^L). \end{aligned}$$

Now, we can calculate the expectation and variance as

$$\begin{aligned} \mathbb{E}_t \left[\frac{W_{t+1}}{W_t} \right] &= (1 - h_t) (1 + r_f + \boldsymbol{\pi}_t \cdot (\boldsymbol{\mu} - r_f \mathbf{1})) + h_t (1 + \mu_L), \\ \text{Var}_t \left[\frac{W_{t+1}}{W_t} \right] &= (1 - h_t)^2 \boldsymbol{\pi}_t \cdot \underline{\Sigma} \boldsymbol{\pi}_t + h_t^2 \sigma_L^2 + 2h_t (1 - h_t) \boldsymbol{\pi}_t \cdot \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L]. \end{aligned}$$

After substitution of these expressions into the objective (9.4), we can ignore any terms not involving $\boldsymbol{\pi}_t$ and divide by the non-negative number $1 - h_t$ to arrive at the objective (9.5).

Note that

$$\begin{aligned} f'(\boldsymbol{\pi}_t) &= \boldsymbol{\mu} - r_f \mathbf{1} - \frac{\gamma}{2} \left((1 - h_t) 2 \underline{\Sigma} \boldsymbol{\pi}_t + 2h_t \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L] \right) \\ &= \boldsymbol{\mu} - r_f \mathbf{1} - \gamma \left((1 - h_t) \underline{\Sigma} \boldsymbol{\pi}_t + h_t \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L] \right), \end{aligned}$$

where we have used the differentiation rules (4.40) and (4.41). The expression (9.6) for the optimal unconstrained portfolio now follows from solving $f'(\boldsymbol{\pi}_t) = \mathbf{0}$ for $\boldsymbol{\pi}_t$.

We can interpret the objective (9.5) as a certainty equivalent of the excess return on the portfolio since it is the excess expected return less a penalty for the return uncertainty. The penalty is a weighted average of the portfolio's return variance $\boldsymbol{\pi}_t \cdot \underline{\Sigma} \boldsymbol{\pi}_t$ and its covariance $\boldsymbol{\pi}_t \cdot \text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L]$ with the return on the human capital. For young individuals, the human capital share h_t is typically close to 1, so the covariance-term dominates the penalty, whereas the portfolio's own variance is relatively unimportant. For old individuals, the human capital is smaller, and then the portfolio variance becomes more important and the covariance with human capital less important.

Looking at the optimal portfolio expression (9.6), we see that, for $\ell_t = 0$, we are back to the conclusion (7.56) from the standard mean-variance model. The ratio ℓ of human capital to financial wealth is obviously crucial for the optimal portfolio. This ratio is typically very large for young individuals and small for older individuals, and the variations in this ratio over life is arguably the most important generator of age-dependence in portfolio decisions. By applying the simple setting above for different values of ℓ , we have effectively introduced a life-cycle perspective on portfolio choice. We assume henceforth that the covariance $\text{Cov}_t[\mathbf{r}_{t+1}, r_{t+1}^L]$ is constant over time.

Let us now specialize to the case with a single risky asset representing the stock market index. Then (9.6) implies that the unconstrained optimal fraction of financial wealth invested in the stock index at time t is

$$\pi_{St} = \frac{\mu_S - r_f}{\gamma\sigma_S^2} (1 + \ell_t) - \ell_t \frac{\rho_{SL}\sigma_L}{\sigma_S} = \frac{\mu_S - r_f}{\gamma\sigma_S^2} + \ell_t \left(\frac{\mu_S - r_f}{\gamma\sigma_S^2} - \frac{\rho_{SL}\sigma_L}{\sigma_S} \right), \quad (9.7)$$

where σ_L is the standard deviation of the return on the human capital and ρ_{SL} is the correlation between the return on the stock and the return on the human capital. The term $\frac{\mu_S - r_f}{\gamma\sigma_S^2}$ is the solution in absence of human capital, which is well-known from Markowitz' original analysis and is also identical to the solution in Merton's basic multi-period portfolio model with constant investment opportunities, see Eq. (8.15). Human capital affects the optimal stock weight via the scaling term $1 + \ell_t$ and through the term $\ell_t \rho_{SL} \sigma_L / \sigma_S$, which adjusts for the extent to which the human capital replaces a stock investment.

Table 9.2 illustrates the optimal portfolio over a one-year period for frequently used parameter values. The riskfree rate is $r_f = 1\%$, the stock has an expected rate of return of $\mu_S = 6\%$ and a standard deviation of $\sigma_S = 20\%$. The standard deviation of relative changes in the human capital is $\sigma_L = 10\%$, and the correlation between the stock and the human capital is $\rho_{SL} = 0.1$. The numbers in Table 9.2 are to be read in the following way. For an agent with a relative risk aversion of $\gamma = 5$ and a human/financial wealth ratio of $\ell_t = 10$, the optimal decision is to invest 225% of current financial wealth in the stock, partly financed by a loan of 125% of current financial wealth. This levered stock investment has an expected rate of return of 12.3% (rounded from 12.25) and a standard deviation of 45%.

The table reveals that the weight of the stock is decreasing in the agent's degree of risk aversion γ . For a fixed γ , the optimal stock weight is decreasing in the ratio of human capital to financial wealth and thus typically decreasing over life in support of the typical "more stocks when young" advice. This feature is due to the fact that the term in the last

ℓ_t	$\gamma = 1$				$\gamma = 5$				$\gamma = 10$			
	stock	rf	exp	std	stock	rf	exp	std	stock	rf	exp	std
0	125	-25	7.3	25	25	75	2.3	5	13	87	1.6	3
1	245	-145	13.3	49	45	55	3.3	9	20	80	2.0	4
2	365	-265	19.3	73	65	35	4.3	13	28	72	2.4	6
5	725	-625	37.3	145	125	-25	7.3	25	50	50	3.5	10
10	1325	-1225	67.3	265	225	-125	12.3	45	88	12	5.4	18
20	2525	-2425	127.3	505	425	-325	22.3	85	163	-63	9.1	33
50	6125	-6025	307.3	1225	1025	-925	52.3	205	388	-288	20.4	78

Table 9.2: Optimal portfolios with human capital.

The table shows the percentages of financial wealth optimally invested in the stock and the riskfree asset, as well as the expectation and standard deviation of the financial return in percent. The assumed parameter values are $r_f = 1\%$, $\mu_S = 6\%$, $\sigma_S = 20\%$, $\sigma_L = 10\%$, and $\rho_{SL} = 0.1$.

bracket in (9.7) is positive with the assumed parameter values. The intuition is that the human capital resembles a riskfree investment much more than a stock investment so, to obtain the optimal overall risk profile, young agents (more precisely, those with large ℓ_t) short the riskfree asset and invest a lot in stocks.

Some of the portfolios in the table require substantial borrowing. However, banks might be reluctant to lend to individuals just because they expect to earn a lot of money in the future. In particular, if they want to spend the borrowed funds on stock investments. If borrowing constrained so that $\pi_{St} \leq 1$, 100% in stocks is optimal for all risk-tolerant and also more risk-averse investors with sufficient human capital relative to financial wealth (young and middle-aged agents). Figure 9.2 illustrates how the constrained optimal stock weight varies with the human-financial wealth ratio for different degrees of risk aversion (left panel) and age (right panel), where the translation to age follows the human capital calculation in the previous subsection.³ Young investors, even quite risk averse, should hold all their financial wealth in stocks. As they grow older, they should eventually start shifting gradually from stocks to bonds.

Many pension funds invest the retirement savings of each individual member following such a glidepath strategy starting with mostly stocks when the individual is young and then gradually decreasing the weight of stocks and increasing the weights of bonds as the individual ages. In the U.S., many individuals save for retirement through so-called 401(k) plans, and many of these 401(k) investors choose to invest their savings in target-date funds which have a built-in glidepath strategy. Target-date funds are offered by Vanguard, BlackRock, Fidelity, and other major investment management companies.

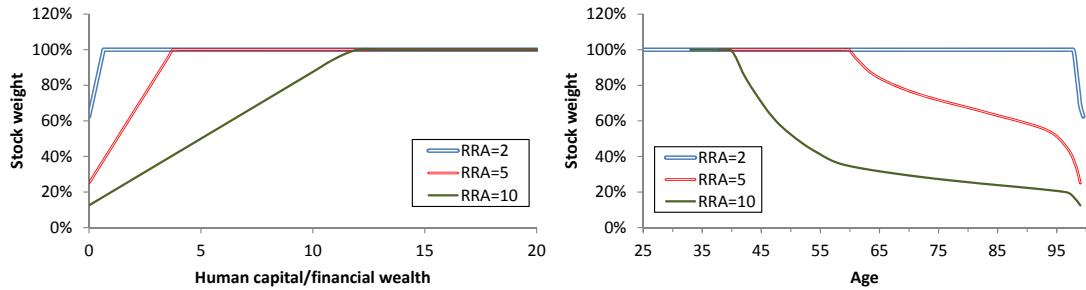
The impact of human capital on optimal investments is parameter dependent, however. Table 9.3 illustrates that results are different for investors with high risk aversion and either high income-stock correlation or high income uncertainty (or both). For $\gamma = 10$ and either an income-stock correlation of 0.4 (instead of 0.1) or a human capital standard deviation of 40% (instead of 10%), the term in the last bracket in (9.7) is negative so that the optimal stock weight is now decreasing in the human-financial wealth ratio ℓ_t . Consequently, very risk-averse agents should hold less stocks when young if their income is sufficiently risky or sufficiently stock-like. If such agents cannot short stocks, the optimal strategy is to have

³The human-to-financial ratio $\ell_t = 1, 2, 5, 10, 20, 50$ then corresponds roughly to ages of 97, 84, 55, 44, 35, and 26, respectively.

ℓ_t	$\gamma = 1$				$\gamma = 5$				$\gamma = 10$			
	stock	rf	exp	std	stock	rf	exp	std	stock	rf	exp	std
0	125	-25	7.3	25	25	75	2.3	5	13	88	1.6	3
1	230	-130	12.5	46	30	70	2.5	6	5	95	1.3	1
2	335	-235	17.8	67	35	65	2.8	7	-3	103	0.9	1
5	650	-550	33.5	130	50	50	3.5	10	-25	125	-0.3	5
10	1175	-1075	59.8	235	75	25	4.8	15	-63	163	-2.1	13
20	2225	-2125	112.3	445	125	-25	7.3	25	-138	238	-5.9	28
50	5375	-5275	269.8	1075	275	-175	14.8	55	-363	463	-17.1	73

Table 9.3: Optimal portfolios with riskier human capital.

The table shows the percentages of financial wealth optimally invested in the stock and the riskfree asset, as well as the expectation and standard deviation of the financial return in percent. The results in the table for both a case with a relatively high stock-income correlation and a case with a relatively high income volatility. More precisely, the assumed parameter values are $r_f = 1\%$, $\mu_S = 6\%$, $\sigma_S = 20\%$, and either (i) $\sigma_L = 10\%$, $\rho_{SL} = 0.4$ or (ii) $\sigma_L = 40\%$, $\rho_{SL} = 0.1$.

**Figure 9.2: Optimal stock weight with human capital.**

The figure shows the constrained optimal stock weight as a function of the human capital to financial wealth ratio (left panel) and age (right panel) for three different values of the relative risk aversion coefficient γ . The stock weight is restricted to the interval from 0% to 100%. The assumed parameter values are $r_f = 1\%$, $\mu_S = 6\%$, $\sigma_S = 20\%$, $\sigma_L = 10\%$, and $\rho_{SL} = 0.1$.

nothing in stocks early in life and only introduce stocks into the portfolio later in life when human capital has declined adequately. Figure 9.3 illustrates the life-cycle patterns in the optimal stock weight by showing the dependence of the weight on the human-financial wealth ratio. Here we have set the stock-income correlation to 0.5 in order to illustrate that the stock weight can be completely flat over life, which is the case for $\gamma = 5$.

9.1.4 More advanced models

It can be shown that (9.6) is identical to the dynamically optimal portfolio strategy of an unconstrained power utility investor in a continuous-time setting with constant investment opportunities, provided that the labor income is either riskfree or spanned by traded assets. Here “spanned” means that a portfolio of traded assets can be constructed so that the portfolio value is perfectly correlated with the labor income. Spanned income is unrealistic. Of course, the labor income of most individuals is exposed to macro-economic risks as unemployment rates and income growth rates vary over the business cycle. But a

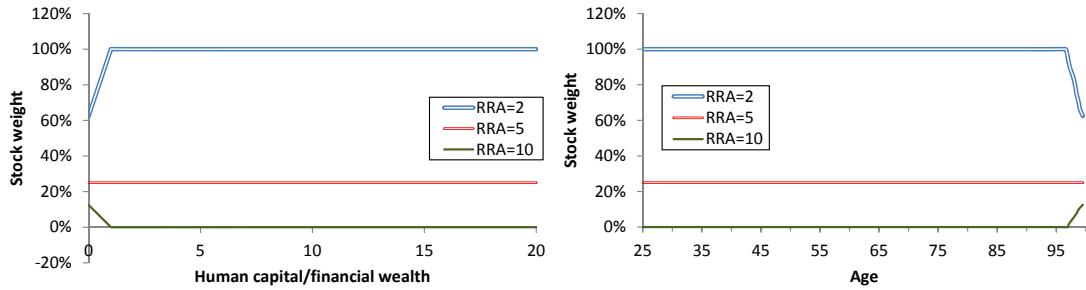


Figure 9.3: Optimal stock weight with stock-like human capital.

The figure shows the constrained optimal stock weight as a function of the human capital to financial wealth ratio (left panel) and age (right panel) for three different values of the relative risk aversion coefficient γ . The stock weight is restricted to the interval from 0% to 100%. We assume a relatively high stock-income correlation of $\rho_{SL} = 0.5$. The other parameter values are $r_f = 1\%$, $\mu_S = 6\%$, $\sigma_S = 20\%$, and $\sigma_L = 10\%$.

larger part of the income uncertainty is idiosyncratic, i.e., specific to the individual. While the aggregate income per capita in the U.S. has a volatility of about 1-3%, the volatility of an individual's income is considerably higher, say 10-20%, with a substantial variation across individuals.

When the income is neither riskfree nor spanned, the dynamic optimization problem does not have a simple closed-form solution, but has to be solved by rather advanced numerical solutions, i.e., computer-based algorithms.⁴ However, the simple portfolio strategy derived above comes close to the numerical solutions reported in various papers (see the discussion in Munk (2020)) as long as we stick to power utility and standard modeling assumptions on investment opportunities and labor income. See, for example, Heaton and Lucas (1997), Munk (2000), Viceira (2001), and Cocco, Gomes, and Maenhout (2005) for models with constant investment opportunities, whereas Munk and Sørensen (2010) and Lynch and Tan (2011) explore the effects of interactions between labor income and time-varying investment opportunities. As the simple mean-variance model, these papers report that with a near-zero income-stock correlation, young borrowing-constrained investors should invest 100% of their savings in stocks. These findings support the “more stocks when young” advice, but not because of the “stocks are safer in the long run” argument often accompanying the advice.

The suggested full investment in stocks when young is different from observed behavior. Especially among the young, a large share of individuals do not invest in stocks at all. There are several possible explanations. Branger, Larsen, and Munk (2019) show that a careful modeling of the risk and consequences of unemployment can lead to much lower optimal stock weights for young individuals, in some cases even a zero stock investment. Another explanation is transaction costs and implicit costs of learning about stocks and how to invest, cf., e.g., Vissing-Jørgensen (2002) and Alan (2006). Since borrowing-constrained young individuals often want to consume almost their current income, the amount of money they want to invest in stocks may not be large enough to cover such participation costs. The next section presents another plausible explanation, namely that stock investments are crowded out (i.e., replaced) by housing investments.

⁴Optimal unconstrained strategies have been derived in closed form for some settings with negative exponential utility and normally distributed income (Svensson and Werner 1993, Henderson 2005, Christensen, Larsen, and Munk 2012) and for settings with income being deterministic or perfectly correlated with financial asset prices (Hakansson 1970, Bodie, Merton, and Samuelson 1992).

The studies mentioned above all assume a single risky asset representing the stock market index. However, not all stocks are equally correlated with the income of a specific individual. To obtain the best overall diversification, the individual would prefer to tilt the portfolio towards stocks negatively correlated with her labor income and away from stocks highly positively correlated with labor income. It seems likely that the income of an individual is most often positively correlated with the price of the stocks of the company employing the individual. This suggests to put a lower (maybe even zero) weight on this stock in the individual's portfolio. However, the opposite is often seen. Many U.S. companies run a retirement savings fund for the employees of the company, and quite often this fund invests a substantial share of the savings in the company's own stock.

The infamous Texas-based energy company Enron illustrates how badly this may turn out. The company was apparently very successful in the 1990's and was named America's most innovative company six years in a row by the business magazine Fortune. The stock price of Enron was booming and reached its peak in August 2000. Enron's employee retirement fund invested a large share of the savings in Enron stocks and for a long period of time that served the employees well. But much of the apparent success was due to accounting fraud and after the public started learning about this during 2001, the stock price plunged. On December 2, 2001, the company was declared bankrupt, most employees were laid off, and the stocks were virtually worthless. Before this dramatic collapse, 58% of the retirement fund was invested in Enron stocks. The employees not only lost their jobs, but also the lion's share of their pension savings.

Finally, let us consider the impact of labor income on consumption. Suppose the individual could sell the rights to her future income stream and receive the human capital upfront instead. Then the individual could just add that to her initial financial wealth and spend her total wealth on consumption over the life cycle as she would do it in the no-income case. She would generally prefer a rather smooth consumption path over life, which may be quite different from her income stream. She would probably prefer consuming more than her current income in the earlier years of working life, but she may not be able to borrow the difference. Furthermore, because of the uncertainty about future income, she may decide having a buffer of wealth in case of unemployment, disability, or any other drop in income. Hence, due to borrowing constraints and buffer savings early-life consumption is lower than it would be in an ideal world. The extra savings (with returns) can then be spent on increased consumption later in life when constraints are less binding and future income is less uncertain. Various papers have shown that such features may lead to the observed hump-shaped life-cycle pattern in consumption, cf., e.g., Thurow (1969), Gourinchas and Parker (2002), and Cocco, Gomes, and Maenhout (2005).

9.2 Housing

9.2.1 Background

Housing plays a key role in individuals' consumption and investment decisions over the life cycle as illustrated by the following statistics. First, "housing services" is one of the most heavily-weighted items in the average basket of consumption goods in the United States. For example, in December 2021 the U.S. Consumer Price Index for All Urban Consumers applied a weight of 32.6% to the category *shelter*.⁵ Second, the importance of housing as an investment asset is illustrated by the home ownership rate, which in the U.S. has been in the range 62.9-69.2% in the period 1965-2021 and is also well above 50%

⁵See the homepage of the Bureau of Labor Statistics at <http://www.bls.gov/cpi/> for the U.S. data and the homepage of the OECD at <http://stats.oecd.org/> for international data.

in most other developed countries.⁶ Moreover, 27% of U.S. home sales in 2011 were purely for investment purposes.⁷ Third, housing wealth constitutes a large share of household assets: in 2010 the value of residential property owned by U.S. households was 36% of total household wealth, and for middle-income households the share is even larger.⁸

Housing has a dual role by serving both as an essential consumption good and an important investment for households. The consumption services offered by a house or an apartment can be obtained either by renting or by owning the residence. Annual rental rates are often estimated to be in the interval from 3% to 10% of the market value of the residence, but this rent-price ratio varies with the level of house price, interest rates, and other macroeconomic factors. Figure 9.4 shows that the average residential rent-to-price ratio in the U.S. has varied between 3% and 6% since 1960. During the housing boom in 1997-2006, where real U.S. home prices went up 85%, rents did not follow in lockstep with the market prices of housing units so the rent-to-price ratio dropped markedly. In the years following the 2008 financial crisis the rent-to-price ratio returned to around 5% more in line with pre-boom values, but has dropped again around 2020 due to another surge in prices.

By owning the residence, an investor gets exposure to real estate prices and may hope to earn capital gains upon a resale of the property. While owning the real estate, the investor can choose to live in it and enjoy consumption services or to rent out the residence and cash in rents. The owner generally has to cover property taxes and maintenance costs.

Investors can also obtain exposure to the housing market by investing in shares of REITs (real estate investment trusts), which are investment companies owning, renting out, and often managing a portfolio of real estate properties. Some REITs specialize in commercial real estate, others in residential real estate. The REITs investing in physical real estate are called equity REITs, whereas so-called mortgage REITs issue or take over existing mortgages to owners of real estate and receive the payments on those mortgages. Well-developed REIT markets exist in the United States, Australia, United Kingdom, France, Japan, and some other countries. Cotter and Roll (2015) study the risk and return characteristics of U.S. REITs. As explained by Ang (2014, Ch. 11), REIT returns exhibit only a low short-run correlation with returns on directly owned real estate (and a higher correlation with common stocks), but longer-term correlations are significantly higher, cf., e.g., Hoesli and Oikarinen (2012). Pagliari, Scherer, and Monopoli (2005) argue that after various relevant adjustments REIT returns and direct real estate returns are much more highly correlated even in the short run, and Lee, Lee, and Chiang (2008) and others report that REITs behave more and more like real estate and less and less like ordinary stocks. In some countries it is also possible to invest in financial securities linked to an index of house price for the entire country or a specified region.

How has real estate performed as an investment asset in the past? Figure 9.5 shows the development in an index of home prices in the United States since 1890. The figure is based on data published on the homepage of Professor Robert Shiller at Yale University (see <http://www.econ.yale.edu/~shiller/data.htm>), and is an updated version of Figure 2.1 in his book Shiller (2005). Home prices and construction costs are in real

⁶See the homepage of the Census Bureau at <http://www.census.gov/housing/hvs/> for U.S. data and Andrews and Sánchez (2011) for data from selected OECD countries.

⁷This number is taken from the Investment and Vacation Home Buyers Survey by the National Association of Realtors, cf. Choi, Hong, Kubik, and Thompson (2016).

⁸Data is based on the Survey of Consumer Finances, see Tables 8 and 9.1 in Bricker, Kennickell, Moore, and Sabelhaus (2012), and residential property includes both primary residence and other residential property but not equity in non-residential property. See also Campbell (2006, Figures 2 and 3) and Guiso and Sodini (2013).

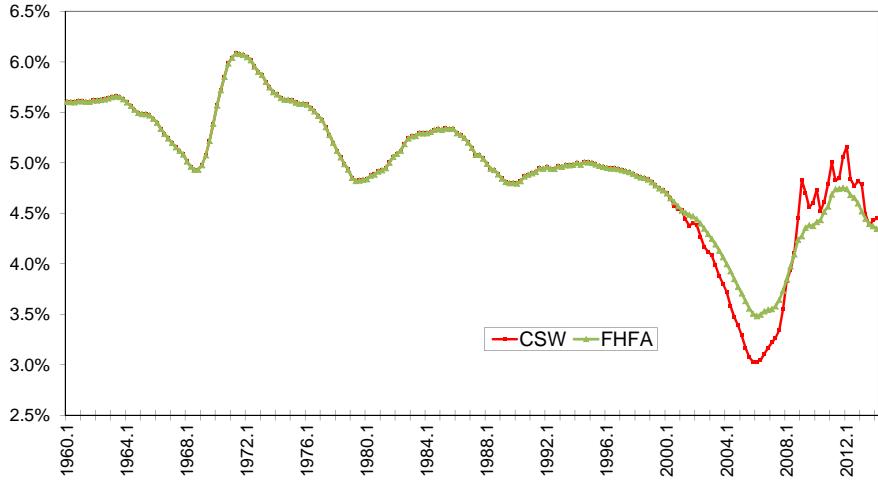


Figure 9.4: U.S. rent-to-price ratio, 1960-2014.

The figure shows the average ratio of estimated annual rents to house prices for the aggregate stock of housing in the United States. The rents are gross rents, not accounting for income taxes or depreciation. The data is with quarterly frequency from 1960Q1 to 2014Q1. The data and graph were taken from the homepage of the Lincoln Institute of Land Policy at <http://www.lincolninst.edu/subcenters/land-values/rent-price-ratio.asp> but are apparently not updated any longer. After 2000, the green curve employs a house price index of the Federal Housing Finance Agency (FHFA), whereas the red curve is based on the Macromarkets LLC national house price index (formerly the Case-Shiller-Weiss index).

terms. Obviously, home prices change substantially over some rather short periods. While current inflation-adjusted home prices are certainly above the price level in 1890, the geometric average growth rate is only around 0.3% per year over the 1890-2022 1.0% over the 1952-2022 period. This suggests that residential real estate is a mediocre investment, but remember that the owner can choose to rent out the property and receive a rent of maybe 5% or, alternatively, enjoy the benefits of living in the home.

The house prices used for these calculations do not take into account that the general quality of residential units has improved significantly over the years, so the price of a hypothetical constant-quality house would have suffered a severe drop. In that sense the price graph overstates the increase in home prices. Home-owners would have had to spend money and time on maintenance and improvements in order to experience the price path illustrated shown in the figure.

There are substantial differences in the development of house prices across geographical areas. Some urban areas can easily expand geographically, whereas others are physically restricted for example by the sea or mountains. Even within an area there can be large dispersion in price movements. The investment in any specific home therefore carries a sizeable idiosyncratic risk in addition to the overall housing market risk. The volatility of the nationwide U.S. home price index has a magnitude of 2-4%, while Goetzmann and Spiegel (2002) report city index volatilities of 3-6%. Estimates of the volatility of individual home prices are in the range 10-15%, cf., e.g., Case and Shiller (1989), Flavin and Yamashita (2002), and Bourassa, Haurin, Haurin, Hoesli, and Sun (2009). Assuming that the rental income exceeds the maintenance costs and property taxes by 3-4 percentage points, the low volatility of the nationwide index indicates that a diversified investment in

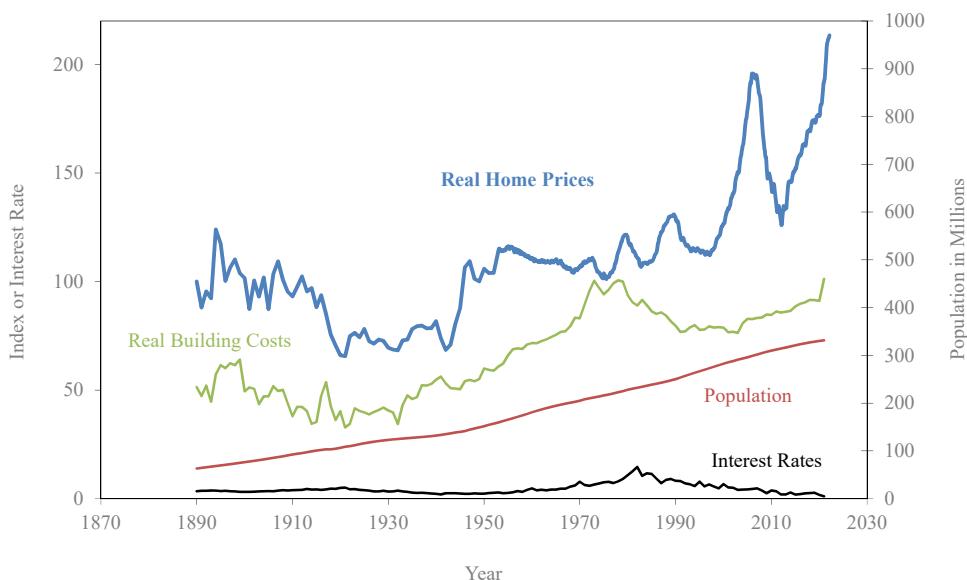


Figure 9.5: U.S. home prices, 1890-2022.

The figure shows an index of inflation-adjusted home prices in the U.S. (1890: index 100) together with an index of inflation-adjusted home construction costs, the long-term interest rate, as well as the population (right axis). The data is taken from the homepage of Professor Robert Shiller at Yale University.

the housing market is quite attractive. An investment in a specific home is less attractive due to the higher risk, but still seems comparable to a diversified investment in the stock market once the possible rental income or the consumption services from occupying the home are taken into account. And adding real estate to a stock investment further brings diversification benefits as stock prices and real estate prices are far from perfectly correlated. For example, there is a correlation of 0.25 between quarterly changes in the U.S. stock market and in a national U.S. home price index over the period 1953-2010.⁹

Intertemporal hedging concerns motivate additional investments in real estate. If you are not owning residential real estate, you have to rent a home and may fear future increases in rents that are due, for example, to increases in real estate market prices. One way to partially insure against increases in the price of housing consumption is to have a positive investment exposure to home prices, such that if the rent you pay go up, you also profit on your investments. Or, formulated differently, if you already own the home you want to live in for many, many years, why care about changes in rents or real estate prices?

A negative aspect of real estate investments is the high transaction costs, much higher than the costs of trading financial securities. Consequently, individuals tend to trade real estate infrequently. By purchasing a house or an apartment and living in that home, the housing consumption seems to be effectively locked in for a prolonged period of time, and the wealth exposure to real estate prices seems to be defined by this investment. But to some extent households can vary their housing position without physically transacting housing units. By spending money on home improvements or maybe even extending the

⁹The stock market is represented by the CRSP (center for Research in Security Prices) value-weighted market portfolio inclusive of the NYSE, AMEX, and NASDAQ markets. The home price index is the national Case-Shiller home price index with data taken from Robert Shiller's homepage at <http://www.econ.yale.edu/~shiller/data.htm>.

house by a newly built section, the consumption services can be increased. Furthermore, households may rent additional housing units (for example, a secondary home) or rent out part of their home, and in that way adjust housing consumption in their preferred direction. The investment exposure can be adjusted by investing in additional housing units or in REITs or other financial assets linked to house prices. To some extent housing consumption can thus be disentangled from housing investment and these positions can be adjusted without incurring large transaction costs.

Households' real estate investments are often financed by a mortgage offered by a bank or specialized mortgage institution, cf. Section 5.10.3, together with a relatively small down payment, and maybe some more traditional bank loans. The real estate serves as collateral for the mortgage so the issuer of the mortgage takes over the property in case the household cannot make the promised payments on the mortgage. Ordinary households can generally not borrow large amounts to invest in financial assets or to increase current consumption, but many households can borrow a large amount to invest in a home. Other things equal, this difference makes housing investments more attractive than stock investments, in particular to young individuals.

The mortgage issuer should have an interest in not offering mortgages to households with insufficient income or mortgages financing a home purchase at a price much higher than the price at which the home is likely to be resold. However, in the 1997-2006 housing boom in the United States, intense competition among mortgage issuers and bad risk management practices led to lax lending standards. The easy access to credit for even low-income families was contributing to the increase in housing demand and the run up of prices. Eventually a substantial share of households was not able to meet the required mortgage payments, which resulted in large losses in the mortgage institutions and many other financial institutions that held bonds and other securities linked to mortgages. A similar pattern was seen in other countries.

9.2.2 Housing in the extended mean-variance model

To investigate the role of housing in households' investment decisions, we add residential real estate to the mean-variance setting of Section 9.1.3.¹⁰ For notational simplicity we leave out time subscripts. Let r_H denote the rate of return on real estate or "housing" over the investment period with an expectation of μ_H and a standard deviation of σ_H . At the beginning of the period, the agent chooses the portfolio weights π_S and π_H of the stock and of housing, respectively, with the remaining financial wealth invested in the riskfree asset. This fits into our general model specification with

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_S \\ \pi_H \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} r_S \\ r_H \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_S \\ \mu_H \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_S^2 & \rho_{SH}\sigma_S\sigma_H \\ \rho_{SH}\sigma_S\sigma_H & \sigma_H^2 \end{pmatrix}, \quad \text{Cov}[\mathbf{r}, r_L] = \begin{pmatrix} \rho_{SL}\sigma_S\sigma_L \\ \rho_{HL}\sigma_H\sigma_L \end{pmatrix},$$

where the ρ 's denote the various correlations as indicated by the subscripts. In our illustrations below we assume the parameter values listed in Table 9.4. In particular, we assume a 4% expected annual return on residential real estate, which again is the expected real price appreciation plus the rental rate less taxes and maintenance costs. The house price volatility is 10%, and the slightly positive pairwise correlations between stock prices,

¹⁰Flavin and Yamashita (2002) and Pelizzon and Weber (2009) include housing in a mean-variance framework, but assume that the housing investment position is exogenously given and ignore human capital.

Symbol	Description	Baseline value
r_f	Riskfree rate	0.01
μ_S	Expected stock return	0.06
σ_S	Stock price volatility	0.20
μ_H	Expected housing return	0.04
σ_H	House price volatility	0.10
σ_L	Human capital volatility	0.10
ρ_{SH}	Stock-house correlation	0.20
ρ_{SL}	Stock-human capital correlation	0.10
ρ_{HL}	House-human capital correlation	0.10

Table 9.4: Parameter values assumed in illustrations.

The table lists the parameter values for the riskfree rate, the stock index, real estate, and human capital in the examples and illustrations, unless mentioned otherwise.

house prices, and labor income are all in line with empirical studies.

In this case we can write the optimal unconstrained portfolio weights in (9.6) as

$$\pi_S = \frac{1}{\gamma(1 - \rho_{SH}^2)\sigma_S} (1 + \ell) \left(\frac{\mu_S - r_f}{\sigma_S} - \rho_{SH} \frac{\mu_H - r_f}{\sigma_H} \right) - \ell \frac{\sigma_L \rho_{SL} - \rho_{SH} \rho_{HL}}{\sigma_S \sqrt{1 - \rho_{SH}^2}}, \quad (9.8)$$

$$\pi_H = \frac{1}{\gamma(1 - \rho_{SH}^2)\sigma_H} (1 + \ell) \left(\frac{\mu_H - r_f}{\sigma_H} - \rho_{SH} \frac{\mu_S - r_f}{\sigma_S} \right) - \ell \frac{\sigma_L \rho_{HL} - \rho_{SH} \rho_{SL}}{\sigma_H \sqrt{1 - \rho_{SH}^2}}. \quad (9.9)$$

Again, the speculative demands are scaled due to the presence of human capital and the portfolio weights are subsequently adjusted for the extent to which the human capital resembles a stock and a real estate investment, respectively.

Table 9.5 shows optimal portfolios for different combinations of the risk aversion coefficient γ and the human-financial wealth ratio ℓ . As found in Section 9.1.3, agents with low risk aversion or with medium-to-high risk aversion and a significant human capital want to borrow money to boost their investment in the risky assets. Human capital works like an inherent investment primarily in the riskfree asset due to the low correlations of human capital with the risky assets. Hence, the larger the human capital, the more the agent borrows and invests in the risky assets. Real estate dominates the risky portfolio. The tangency portfolio has 28% in stocks and 72% in real estate due to real estate having a larger Sharpe ratio than stocks. Despite stocks and real estate having identical correlations with human capital, the income hedge portfolio has 1/3 in stocks and 2/3 in real estate because of real estate having a standard deviation half the size of stocks. With human capital, the income hedge portfolio is subtracted from the (magnified) investment in the tangency portfolio, so the income hedge causes an increase in the real estate to stock ratio. For example, with a relative risk aversion of 5 the ratio is $52/20 = 2.6$ without human capital and $927/332 \approx 2.8$ with a human-financial wealth ratio of 20. By comparing Table 9.5 to Table 9.2, we see that the introduction of real estate reduces the optimal weight in the stock index and in the riskfree asset (for most investors the latter means: increases borrowing).

Holding real estate gives easy access to loans through mortgages, while stock investments generally do not, at least not to the same extent. Suppose that you can borrow at most a fraction $1 - \kappa$ of the value of the real estate you own. This corresponds to the constraint

$$\pi_S + \kappa \pi_H \leq 1 \quad (9.10)$$

ℓ	$\gamma = 1$			$\gamma = 5$			$\gamma = 10$		
	stock	house	rf	stock	house	rf	stock	house	rf
0	99	260	-259	20	52	28	10	26	64
1	194	513	-606	35	96	-31	16	44	41
2	289	765	-953	51	140	-91	21	61	17
5	573	1521	-1994	98	271	-269	39	115	-53
10	1047	2781	-3728	176	490	-566	67	203	-170
20	1995	5302	-7197	332	927	-1159	124	380	-405
50	4839	12865	-17603	801	2240	-2941	296	911	-1108

Table 9.5: Optimal unconstrained portfolios.

The table shows percentages of financial wealth optimally invested in stock, real estate, and riskfree asset. The baseline parameter values listed in Table 9.4 are assumed.

on portfolio weights. We take $\kappa = 0.2$ corresponding to an 80% loan-to-value limit as our benchmark. Panel A of Table 9.6 shows the optimal portfolio for various combinations of the relative risk aversion and the human-financial wealth ratio. Several things are worth noticing. First, a levered house investment is very attractive for investors who are young/middle-aged or relatively risk tolerant. Secondly, non-participation in the stock market is optimal for young investors. Thirdly, the optimal stock weight is increasing or hump-shaped over life. Fourthly, the optimal stock weight can be non-monotonic in risk aversion which with the assumed parameter values is the case for a human-financial wealth ratio of 1, 2, or 5. This phenomenon occurs because the agent compares stocks to a levered house investment and stocks are less risky than a levered house investment. Hence, when increasing the risk aversion, the agent gradually shifts from a levered house investment to stocks and eventually to the riskfree asset.

To better understand the impact of the access to collateralized borrowing, Panel B of Table 9.6 lists optimal portfolios in the case of a 60% loan-to-value limit, whereas Panel C assumes no borrowing at all. Note that a portfolio weight written in blue (red) is larger (smaller) than the corresponding weight in the baseline case of Panel A. The young investors' appetite for financial investments with high risk (and high expected return) implies that if mortgages are available, they prefer housing investments with a mortgage to an unlevered stock investment. However, if borrowing is prohibited, the stock is more attractive than the house because of the stock's higher risk and expected return so that the young or risk-tolerant households optimally invest their entire financial wealth in stocks. The extent to which a housing investment gives access to borrowing is thus essential for the optimal portfolio, especially for the young or risk-tolerant households. In line with intuition, a reduction in the loan-to-value limit (i.e., an increase in κ) decreases (or leaves unchanged) the portfolio weight of the house and the borrowed amount, whereas the stock weight can vary non-monotonically. Finally, Panel D assumes that the borrowing rate is 2%, whereas the lending rate is still 1%. Of course, portfolios not involving borrowing are unchanged (high risk aversion, low human-financial wealth ratio). Some agents who were borrowing in the baseline case are now neither borrowing nor lending (for $\gamma = 5$, $\ell = 1$ and $\gamma = 10$, $\ell = 5$). Other agents are borrowing less, whereas the agents with relatively low risk aversion and high human capital still borrow as much as possible and invest nothing in stocks. Again, the overall qualitative patterns in how the optimal portfolio weights vary with the level of risk aversion and the human-financial wealth ratio remain unchanged.

ℓ	$\gamma = 1$			$\gamma = 5$			$\gamma = 10$		
	stock	house	rf	stock	house	rf	stock	house	rf
Panel A: Baseline case with max 80% LTV, $\kappa = 0.2$									
0	52	240	-192	20	52	28	10	26	64
1	13	434	-347	35	96	-31	16	44	41
2	0	500	-400	51	140	-91	21	61	17
5	0	500	-400	50	250	-200	39	115	-53
10	0	500	-400	16	420	-336	60	200	-160
20	0	500	-400	0	500	-400	32	340	-272
50	0	500	-400	0	500	-400	0	500	-400
Panel B: max 60% LTV, $\kappa = 0.4$									
0	33	167	-100	20	52	28	10	26	64
1	4	241	-145	35	96	-31	16	44	41
2	0	250	-150	47	133	-80	21	61	17
5	0	250	-150	30	174	-105	39	115	-53
10	0	250	-150	3	242	-145	36	159	-95
20	0	250	-150	0	250	-150	12	220	-132
50	0	250	-150	0	250	-150	0	250	-150
Panel C: no borrowing, $\kappa = 1$									
0	62	38	0	20	52	28	10	26	64
1	100	0	0	31	69	0	16	44	41
2	100	0	0	38	62	0	21	61	17
5	100	0	0	60	40	0	31	69	0
10	100	0	0	95	5	0	43	57	0
20	100	0	0	100	0	0	67	33	0
50	100	0	0	100	0	0	100	0	0
Panel D: Higher borrowing than lending rate, $r_{\text{bor}} = 2\%$, $r_{\text{len}} = 1\%$									
0	68	160	-128	20	52	28	10	26	64
1	45	274	-219	30	70	0	16	44	41
2	22	388	-310	42	83	-25	21	61	17
5	0	500	-400	69	154	-123	31	69	0
10	0	500	-400	51	244	-195	50	100	-50
20	0	500	-400	15	424	-339	66	172	-138
50	0	500	-400	0	500	-400	30	352	-282

Table 9.6: Optimal portfolios with borrowing constraints.

The table shows percentages of financial wealth optimally invested in stock, real estate, and riskfree asset. The baseline parameter values listed in Table 9.4 are assumed. In Panels B, C, and D the numbers in blue are larger than in the baseline case of Panel A, numbers in red are smaller, whereas the remaining numbers are unchanged.

9.2.3 More advanced models

The mean-variance model above illustrates a number of aspects of households' life-cycle portfolio decisions, but does not produce the true dynamically optimal decisions. Several papers have numerically solved for the optimal life-cycle decisions in more advanced models that take housing transactions costs, renting and owning decisions, portfolio and borrowing constraints, and consumption into account. See, e.g., [Cocco \(2005\)](#) and [Yao and Zhang \(2005\)](#). These papers generally confirm the qualitative conclusions of the mean-variance model. In particular, young individuals should invest nothing or little in stocks and instead invest in residential real estate serving as collateral for a loan. Hence, housing may explain the observed low stock market participation of young individuals. Later in life, stocks and bonds enter the optimal portfolio, but housing continues to dominate.

Home investments are most often partly financed by a mortgage. Since mortgages come in a large variety, a home buyer also faces decisions regarding mortgage type and maturity, in addition to the amount to be borrowed. A key distinction is between fixed-rate mortgages and adjustable-rate mortgages. At least seen in isolation, an adjustable-rate mortgage is more risky as the interest rate to be paid by the borrower may increase, but on the other hand short-term interest rates are most of the time below the long-term interest rates, so the borrower faces a trade-off. Furthermore, to the extent that the labor income and investment returns of the individual follow the level of short-term interest rates, the adjustable rate mortgage is really not that risky, but offers a form of built-in hedge: if labor income or investment returns drop, the payments of the mortgage would also tend to fall. Optimal mortgage decisions have been studied in formal models by [Campbell and Cocco \(2003\)](#) and [van Hemert \(2010\)](#), among others.

In addition to the papers mentioned above, various aspects of the link between housing and life-cycle consumption and portfolio choice have been discussed by, e.g., [Kraft and Munk \(2011\)](#) [Attanasio, Bottazzi, Low, Nesheim, and Wakefield \(2012\)](#), [Fischer and Stamos \(2013\)](#), [Corradin, Fillat, and Vergara-Alert \(2014\)](#), and [Kraft, Munk, and Wagner \(2018\)](#). Much work is still ahead to enhance our understanding of the dynamics of real estate prices and their correlation structure with labor income and stock markets, as well as the optimal individual decisions in a realistic model encompassing both consumption, investment, and housing over the life cycle.

9.3 Saving for retirement

As public pension payouts typically fall short of pre-retirement income, individual workers should think about building up savings during working life and then decumulate these savings in retirement in order to smooth consumption over the life cycle. How much should you save for retirement?

Let us begin with a simple setting that ignores uncertainty. Suppose your annual income after tax is constant and equal to I . You work for T_W years and then live on being retired for another T_R years. At the end of each year you save a fraction s of your after-tax income. The annual savings sI are invested in a portfolio of financial assets. The annual rate of return of the portfolio is $r > 0$ which is assumed to be known. Then the value of your savings at the end of your working life is going to be

$$F = sI(1+r)^{T_W-1} + sI(1+r)^{T_W-2} + \cdots + sI = sI \frac{(1+r)^{T_W} - 1}{r}. \quad (9.11)$$

When retired, you gradually sell out of your portfolio to finance a constant annual payout

Public pens	Rate of return on savings					
	0	0.01	0.02	0.03	0.04	0.05
0.00	33.33	26.96	21.30	16.48	12.51	9.35
0.25	25.00	20.22	15.98	12.36	9.38	7.01
0.50	16.67	13.48	10.65	8.24	6.26	4.68
0.75	8.33	6.74	5.33	4.12	3.13	2.34

Table 9.7: Required saving rate. Work for $T_W = 40$ years, retired for $T_R = 20$ years. Constant annual after-tax income.

over the T_R years (an annuity) equal to

$$A = F \frac{r}{1 - (1 + r)^{-T_R}} = sI \frac{(1 + r)^{T_W} - 1}{1 - (1 + r)^{-T_R}}. \quad (9.12)$$

Suppose that, in addition, you receive a state pension of P per year. Let b denote the state pension relative to the pre-retirement income, i.e. $P = bI$.

Before retirement, the annual income after savings is $(1 - s)I$ which can be spent on consumption. In retirement, you can spend $P + A$, the sum of state pension and the payout from your own retirement savings. If you aim to have the same level of consumption in retirement as before retirement, you need to choose the savings rate s so that

$$(1 - s)I = P + A, \quad (9.13)$$

that is,

$$(1 - s)I = bI + sI \frac{(1 + r)^{T_W} - 1}{1 - (1 + r)^{-T_R}}. \quad (9.14)$$

The solution is

$$s = \frac{1 - b}{1 + \frac{(1+r)^{T_W}-1}{1-(1+r)^{-T_R}}} = (1 - b) \frac{(1 + r)^{T_R} - 1}{(1 + r)^{T_R + T_W} - 1}. \quad (9.15)$$

Table 9.7 shows the required saving rate s from (9.15) in percent of income for different combinations of b , the ratio of public pension to pre-retirement income, and r , the annual rate of return on savings. We consider an individual working for $T_W = 40$ years and being retired for $T_R = 20$ years. With zero return on savings and zero state pension, you need to save 1/3 of your income since you will be retired 1/3 of your adult life (20 years out of 60). The higher the state pension and the higher the return on savings, the lower the fraction of labor income you need to save for retirement. With a state pension of 50% of labor income and a 3% return on savings, you need only save 8.24% of your income. If your life expectancy $T_W + T_R$ increases, you either need to save a bigger fraction of your income or work for more years (or a combination of both).

This simple framework ignores many important real-life aspects, such as uncertainty about lifetime, returns, and income. In the preceding subsections we have already discussed how the uncertainty about labor income and the correlation between labor income and financial assets affect optimal portfolios over the life cycle. In particular, a glidepath strategy of investing a lot in stocks when young and then gradually lowering the weight of stocks and increasing the weight of bonds seems optimal for many individuals, and the supply of and demand for target-date fund and other life-cycle investment products are

increasing. With the portfolio changing with age, the expected return and the risk on savings are also going to vary over life. Of course, these observations may suggest how the savings are to be allocated across financial assets at different ages, but do not answer the question how much an individual should save for retirement. To answer this question we would need to set up and solve one of the more advanced life-cycle utility maximization problems outlined in Section 9.1.4.

An important ingredient in such models is lifetime uncertainty which complicates the planning of retirement saving considerably. You may live shorter than expected and die with unused savings being bequeathed to your heirs. Or you may live longer than expected so that you run out of money and end up with a low consumption at old age. If you save so much that you can uphold a decent standard of living even if living very long, you will most likely die with substantial unused savings.

Lifetime uncertainty can be managed through annuities where a group of individuals pool their savings and share their lifetime risks. Each participant regularly receives a certain payout for as long as the participant lives. Upon the death of a participant, the remaining balance of the participant is effectively distributed among the surviving participants' accounts. By annuitizing your retirement savings, you can thus eliminate the risk of outliving your wealth. With well-functioning and competitive annuity markets, most individuals should benefit from annuitizing at least part of the retirement savings. Of course, as public pensions are paid out until death, they already work as an annuity, so the demand for additional annuitization of savings depend on the generosity of public pensions and also on the risk attitudes and life expectancy of each individual.

Some countries have mandatory retirement saving schemes covering large groups of workers. Each worker is then required to contribute a certain fraction of income each month to a pension fund. The fund is investing the savings on behalf of its members. When the worker retires, the fund makes regular payouts to the worker, often annuity-style payouts until death. While members of the same fund can have heterogeneous preferences and income dynamics, the fund typically requires the same contribution rate from all members, makes payouts following the same schedule to all retirees, and sometimes even invests the savings of all members in the same portfolio irrespective of their age and risk tolerance.

For further discussion and specific models of optimal saving rates, annuitization decisions, the design of multi-member pension plans, and other challenges related to retirement saving, we refer the reader to Benartzi and Thaler (2007), Dahlquist, Setty, and Vestman (2018), and Larsen and Munk (2023).

9.4 Conclusion

Chapter 8 showed that, under some assumptions, a multi-period investor should hold a portfolio in which each asset has a constant portfolio weight over time. Among other things, this result relies on the investor having no income stream and holding only financial assets, and these assumptions are clearly unrealistic for most individual or household investors.

This chapter has shown that, for most individuals, the presence of labor income generally implies that a large share of financial wealth should be invested in stocks (and other risky assets). Moreover, this stock weight should follow a glide path over the life of the individual, starting out at 100% (or more, if possible) when young and then, at some appropriate age, beginning to decrease with age, to end up at a constant and fairly low level in retirement.

Residential real estate appears to be a relatively attractive asset from a risk-return

perspective and also offers owners access to inexpensive loans. Our theoretical analysis suggests that residential real estate should play a dominant role in the overall portfolios of young individuals, maybe even crowding out stock investments. Later in life, stocks and bond should enter the portfolio, but housing still plays a significant role.

Finally, the chapter has presented some simple computational frameworks for thinking about how much an individual should save for retirement.

Various of the papers mentioned in this chapter compare the outcome of theoretical models to observed household decisions. Guiso and Sodini (2013) survey the empirical household finance literature. Some of the broad findings are that many individuals save too little for retirement, invest too little in risky assets, and do not diversify their portfolio sufficiently.

9.5 Exercises

Exercise 9.1. Suppose that you have just turned $t = 30$ years old with no savings, but with a good job that paid you $L_t = \$50,000$ after taxes in the preceding year. You expect your income to grow by $g = 1\%$ (in real terms) per year for as long as you keep working. You realize that you have to save for retirement and have decided to save a certain fraction $k = 10\%$ of your income at the end of each year. You will start your retirement savings account immediately by contribution $k \times L_t = \$5,000$. The next contribution is $k \times L_{t+1} = k \times L_t \times (1 + g) = \$5,050$ in exactly one year from now. The savings are invested in financial assets, and you expect an annual rate of return of $r = 4\%$. The returns are added to your retirement savings account at the end of each year. At the end of the first year you thus receive a return of $r \times k \times L_t = 0.04 \times \$5,000 = \$200$, which brings the balance of your account to $\$5,000 + \$200 = \$5,200$ plus the new contribution of $\$5,050$, that is $\$10,250$.

- (a) What is your total savings at retirement, assuming that you retire exactly when you turn $T = 60$ years? What if $T = 65$ or $T = 70$?
- (b) How sensitive are your answers to (a) to the values of g , k , and r ?

Suppose that you expect to live until you are $\bar{T} = 80$ years old. When you retire, your savings are distributed as an annuity of your expected remaining lifetime, still assuming that the part yet not paid out is invested at a return of r .

- (c) What is the annuity payment you obtain if you retire at the age of $T = 60$ years? What if $T = 65$ or $T = 70$?
- (d) Answer (c) again for $\bar{T} = 85$.
- (e) How sensitive are your answers to (c) and (d) to the values of g , k , and r ?
- (f) The above computations ignore risk. What types of risk are relevant for these problems?

Exercise 9.2. Consider Merton's basic model for optimal long-term investments. Assume that you can invest only in a riskfree asset and the stock market index. The riskfree asset gives a continuously compounded rate of return of $r_f = 0.01 = 1\%$ per year. The gross return on the stock market index over any T -year period is given by

$$R_T^{\text{stock}} = \exp \left\{ \left(\mu - \frac{1}{2} \sigma^2 \right) T + \sigma \varepsilon \sqrt{T} \right\},$$

where ε is a standard normally distributed random variable. The annualized volatility is $\sigma = 0.2 = 20\%$ and the annualized expected return is given by $\mu = 0.07 = 7\%$ per year.

- (a) What is the optimal investment strategy for an investor with a constant relative risk aversion of $\gamma = 5$?
- (b) Discuss the validity of the following argument:

“There is ample empirical evidence that stocks typically outperform bonds over periods of ten years or more, whereas over shorter periods anything can happen. Therefore, long-term investors should hold more stocks than short-term investors.”

Now we introduce labor income into the model. Let I_t denote the after-tax labor income in year t , which we assume is paid out at the end of the year. Maria is currently standing at the end of year t and expects to earn a labor income over the next $T = T_1 + T_2$ years. Maria's human capital at the end of year t is denoted by L_t and is defined as the present value of the future income to be received, i.e., the income in years $t+1, t+2, \dots, t+T$. If the appropriate discount rate for future income is r , we can formally write this as

$$L_t = \sum_{s=1}^T \frac{\mathbb{E}_t[I_{t+s}]}{(1+r)^s}.$$

Over the first T_1 years she expects her annual labor income to grow at a rate of G per year, whereas over the remaining T_2 years she expects a zero income growth rate.

- (c) Explain why Maria's human capital at the end of year $t+T_1$ is going to be

$$L_{t+T_1} = I_{t+T_1} \frac{1 - (1+r)^{-T_2}}{r},$$

where I_{t+T_1} is her labor income received in year $t+T_1$.

- (d) Explain why Maria's current human capital, i.e. at the end of year t , is given by

$$L_t = I_t \frac{1+G}{r-G} \left[1 - \left(\frac{1+G}{1+r} \right)^{T_1} \right] + I_t \left(\frac{1+G}{1+r} \right)^{T_1} \frac{1 - (1+r)^{-T_2}}{r}.$$

Suppose now that $T_1 = T_2 = 15$, $G = 2\%$, $Y_t = \$50,000$, and that Maria's future income is risk free.

- (e) What is then the appropriate discount rate? What is Maria's current human capital, i.e. at the end of year t ?
- (f) Suppose Maria's financial wealth at the end of year t is \$50,000 and that her relative risk aversion is $\gamma = 5$. What is then Maria's optimal investment in stocks and in the riskfree asset?

Exercise 9.3. Consider Merton's basic model for optimal long-term investments. Assume that you can only invest in a riskfree asset and the stock market index. The riskfree asset gives a continuously compounded rate of return of $r_f = 0.01 = 1\%$ per year. The gross return on the stock market index over any T -year period is given by

$$R_T^{\text{stock}} = \exp \left\{ \left(\mu - \frac{1}{2} \sigma^2 \right) T + \sigma \varepsilon \sqrt{T} \right\},$$

where ε is a standard normally distributed random variable. Assume that the annualized volatility is $\sigma = 0.2 = 20\%$ and the annualized expected return is given by $\mu = 0.05 = 5\%$ per year.

- (a) Suppose that over a period of T years, you invest your entire wealth either in the stock market index or in the riskfree asset. Show that the probability that the gross stock return R_T^{stock} is more than two times the gross riskfree return over the T -year period is equal to

$$\text{Prob} \left(R_T^{\text{stock}} > 2e^{r_f T} \right) = N \left(\frac{\left(\mu - r_f - \frac{1}{2} \sigma^2 \right) \sqrt{T}}{\sigma} - \frac{\ln 2}{\sigma \sqrt{T}} \right),$$

where $N(\cdot)$ denotes the cumulative probability distribution function of the standard normal distribution. Compute this probability for $T = 1, 2, 5, 10, 20, 40$ and discuss the results.

- (b) What is the optimal investment strategy for an investor with a constant relative risk aversion of $\gamma = 5$? The findings in question (a) could suggest that, the longer your investment horizon, the more you should invest in the stock market. Why is that not the case in Merton's basic model?

Now we introduce labor income into the model. We consider an investor with a remaining working horizon of either 5 years or 25 years. Suppose the labor income of the investor is known to be \$40,000 per year and thus completely risk free.

- (c) What is the human capital of the investor with a remaining working horizon of 5 years?
What is the human capital of the investor with a remaining working horizon of 25 years?

Suppose the investor has a relative risk aversion of $\gamma = 5$ and a financial wealth of \$20,000. Furthermore, suppose that she wants to split her total wealth (the sum of financial wealth and human capital) into the stock market index and the riskfree asset in line with Merton's basic model, cf. the answer to question (b) above.

- (d) How does she invest her financial wealth if she has a remaining working horizon of 5 years?
How does she invest her financial wealth if she has a remaining working horizon of 25 years?
Explain why the remaining working horizon has this effect on the optimal investment of the financial wealth.

Exercise 9.4. Angela has a wealth of \$400,000 that she wants to invest over a period of T years. She has decided to use Merton's basic model for long-term investment, and she believes she has a constant relative risk aversion of $\gamma = 2$. The yield curve is flat at an interest rate level of 1% (continuously compounded), and Angela expects the yield curve to stay unchanged. Angela wants to invest only in a riskfree asset (a government bond) and an exchange-traded fund mimicking the domestic stock market index. The estimates of the parameters for the domestic stock market index are $\mu = 0.05$ (the log of the expected gross return per year) and $\sigma = 0.20$ (volatility or standard deviation per year).

- (a) If Angela's investment horizon is $T = 10$ years, how should she invest her wealth according to Merton's model? What if $T = 20$ years?
- (b) Suppose again that $T = 10$ and that Angela has invested in the optimal portfolio identified in (a). She does not trade in the financial market over the next month. Suppose that during this month the return on the domestic stock market index was -10%. Should she then rebalance her portfolio? If yes, exactly how should she do that?

Angela suddenly realizes that she should take her labor income into account when making her investment decision. In the past year, her after-tax labor income was \$40,000. She will receive labor income at the end of each year throughout her investment horizon of T years. She believes her future labor income is riskfree and will grow by 2% per year.

- (c) Suppose that $T = 10$ years. What is Angela's human capital? What is now the optimal way to invest her financial wealth of \$400,000?
- (d) Suppose that $T = 20$ years. What is Angela's human capital? What is now the optimal way to invest her financial wealth of \$400,000?
- (e) Compare your answers to (c) and (d) and explain the differences. Would you expect to see similar differences if the future income is not riskfree?

Exercise 9.5. Larry considers how to invest his financial wealth of \$100,000. He knows that he should take his human capital and its riskiness into account when figuring out his optimal investments, and he has decided to use the mean-variance model extended with human capital for this purpose. He estimates that his human capital is \$400,000 and has a standard deviation of 0.10.

First, he considers investing in the TOP10 stock market index (via an exchange-traded fund) or the riskfree asset or a combination of the two. The TOP10 stock market index consists of the 10 most valuable listed companies. Larry has decided that he does not want to take short positions or borrow money. The riskfree rate is $r_f = 0.01$, whereas the TOP10 stock index has an expected return of $\mu_S = 0.05$ and a standard deviation of $\sigma_S = 0.2$. The correlation between the TOP10 stock index and Larry's human capital is 0.1.

- (a) Determine Larry's optimal portfolio of the TOP10 index and the riskfree asset if he has a relative risk aversion of $\gamma = 3$. Answer the same question if $\gamma = 5$. Show the formula that you apply to find these portfolio weights.
- (b) Briefly explain which parameters or variables that influence how human capital affects the optimal portfolio.

Larry works for the large company MegaCorp whose stocks are also listed on the stock exchange, but they are not included in the TOP10 index. Larry now considers including MegaCorp stocks in his portfolio. The MegaCorp stocks seem quite attractive with an expected return of 0.1 and a standard deviation of 0.3, and the correlation between the return on MegaCorp and the return on the TOP10 index is only 0.2. Naturally, Larry's human capital is relatively sensitive to the well-being of MegaCorp, and Larry estimates that the correlation is 0.8 between his human capital and MegaCorp returns. Larry still does not want to take short positions or borrow money.

- (c) If you ignore Larry's human capital, what would be the optimal portfolio of the TOP10 index, MegaCorp stocks, and the riskfree asset if he has a relative risk aversion of $\gamma = 3$? Answer the same question if $\gamma = 5$.
- (d) Now take Larry's human capital into account. Determine his optimal portfolio of the TOP10 index, MegaCorp stocks, and the riskfree asset if he has a relative risk aversion of $\gamma = 3$. Answer the same question if $\gamma = 5$. Discuss your findings.

Exercise 9.6. Stella has a financial wealth of \$100,000, and she is considering how to invest this wealth. She has decided to invest in a riskfree asset and an exchange-traded fund (ETF) tracking the stock market, and she can invest in both without any constraints. She knows that she should take the magnitude and risk characteristics of her human capital into account. To determine her optimal portfolio, she will use the extended mean-variance model with an annual riskfree rate of $r_f = 0.02$ and stock market parameters $\mu_S = 0.07$ and $\sigma_S = 0.15$. Her relative risk aversion is $\gamma = 5$.

Stella expects to work for 30 more years. In the most recent year, her annual after-tax income was \$40,000, and she expects that to increase by 2% every year until retirement. The standard deviation of her human capital is 0.1 and the correlation between her human capital and the stock market is 0.2.

- (a) Calculate Stella's human capital if the appropriate discount rate for future income is 4%. What is her human capital if the appropriate discount rate is 10%? Which factors determine what the appropriate discount rate is?
- (b) How should Stella invest her financial wealth if she is using a discount rate of 4%? What if she is using a discount rate of 10%?

Stella thinks that the 10% discount rate is appropriate and she invests her financial wealth as calculated above with that discount rate.

Now consider Stella's situation one year later. Over this year, Stella did not actively change her portfolio. During the year, the return on the stock market was 20%. As expected, Stella's labor income during the year was 2% higher than in the previous year. Her consumption during the year was exactly equal to her after-tax income.

- (c) Before any rebalancing of her portfolio, what is now the value of Stella's position in the stock market ETF? What is now her total financial wealth and her human capital?
- (d) How should Stella now optimally invest her financial wealth? Precisely how should she rebalance her current portfolio so that it is optimal?

CHAPTER 10

The Capital Asset Pricing Model

The preceding chapters have focused on the optimal investment decisions of individual investors. Among the inputs to the decision problem are the expected returns of the available assets. For fixed expected dividends of the assets, this implies that the decision-maker takes the prices of the assets as given. But there is an obvious feedback from investment decisions to prices. Prices are set so that the total demand of investors equals the total supply. If we assume a certain supply of assets, the models of investors' asset demand should allow us to make conclusions about the equilibrium prices or, equivalently, the equilibrium expected returns of the assets.

This chapter develops and studies the Capital Asset Pricing Model, which is the most famous model of the equilibrium expected returns of financial assets. The Capital Asset Pricing Model is usually just referred as the CAPM (pronounced “cap-em”). The model was suggested and analyzed by [Treynor \(1961\)](#), [Sharpe \(1964\)](#), [Lintner \(1965\)](#), and [Mossin \(1966\)](#). The CAPM was originally developed in a one-period economy and builds on Markowitz' mean-variance portfolio theory.

Section [10.1](#) states and proves the basic version of the CAPM, and discusses some of its assumptions and implications. Section [10.2](#) summarizes the results of empirical tests of the CAPM, which are not particularly supportive. Some variations of the basic CAPM are outlined in Section [10.3](#). The special consumption-based version of the CAPM is presented in Section [10.4](#).

10.1 The basic CAPM

10.1.1 Statement and proof of the CAPM

The CAPM involves the so-called **market portfolio**, which is the portfolio of all risky assets in the economy. The value of the market portfolio is simply the total value of all risky assets. The weight of any individual asset in the market portfolio equals the ratio of (1) the value of all the issued units of that asset to (2) the value of the market portfolio. Hence, the weights are given by the market capitalization shares. Table [10.1](#) gives a simple example, when the market consists of only three assets. In this example, the market portfolio weights are 20% for asset X, 30% for asset Y, and 50% for asset Z.

Here is the line of thinking that leads to the CAPM. If investors agree on the riskfree rate and the efficient frontier of risky assets, they agree on the composition of the tangency

asset	units	price per unit	market cap	market weights
X	500	\$20	\$10,000	0.2 = \$10,000/\$50,000
Y	750	\$20	\$15,000	0.3 = \$15,000/\$50,000
Z	5,000	\$5	\$25,000	0.5 = \$25,000/\$50,000
Sum			\$50,000	1

Table 10.1: An example of the market portfolio.

The table gives a simple example of the market portfolio in a market with only three assets.

portfolio. Hence, everybody invests only in some combination of the riskfree asset and the common tangency portfolio of risky assets. In equilibrium, asset prices and thus expected returns have to be set so that the total demand of investors equals the supply of assets. Therefore, the tangency portfolio must consist of all the risky assets in the economy and thus the tangency portfolio is identical to the market portfolio. By combining this insight with the property (7.41) of the tangency portfolio, we obtain an expression for the equilibrium expected return on any individual asset.

Why is the tangency portfolio identical to the market portfolio? Suppose that the market portfolio is as given in Table 10.1, but that the tangency portfolio consists of 30% in X, 20% in Y, and 50% in Z. Since the market price of both X and Y is \$20, this implies that every investor wants 1.5 unit of X for every unit of Y. But, in fact there are fewer units of X than Y available in the market (500 vs. 750), so the market is not in equilibrium. The excess demand for X drives up the price of X. The excess supply of Y drives down the price of Y. (We implicitly assume that asset supply is fixed.) This implies both that the market weights adjust and that the optimal portfolio weights of investors change due to the implied modification of the expected returns on X and Y. The market is only in equilibrium when the relative asset weights in investors' portfolios match their weights in the market portfolio.

The market portfolio surely has a positive Sharpe ratio. Given that the market portfolio plays the role as the tangency portfolio, the efficient frontier of risky assets in the (σ, μ) -diagram consists of the upward-sloping straight line starting at the point $(0, r_f)$ corresponding to the riskfree asset and going through the point $(\text{Std}[r_m], E[r_m])$ corresponding to the market portfolio. This line is called the **Capital Market Line**. Its slope equals the Sharpe ratio of the market portfolio, $(E[r_m] - r_f) / \text{Std}[r_m]$.

The CAPM can now be stated in the following way.

Theorem 10.1

In a one-period economy, suppose that

- (i) a riskfree asset exists,
- (ii) all investors have mean-variance preferences and are not subject to any portfolio constraints,
- (iii) all investors have homogeneous beliefs, i.e., they agree on the riskfree rate r_f and on the efficient frontier of risky assets.

Then, when the market is in equilibrium, the following holds:

- (a) The tangency portfolio equals the market portfolio of all risky assets.

- (b) Each investor optimally combines the riskfree asset and the market portfolio.
(c) For any risky asset i , the equation

$$\mathbb{E}[r_i] - r_f = \beta_i (\mathbb{E}[r_m] - r_f) \quad (10.1)$$

holds, where r_i is the rate of return on asset i , r_m is the rate of return of the market portfolio, and β_i is the market beta of asset i , which is defined as

$$\beta_i = \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]}. \quad (10.2)$$

- (d) For any portfolio weight vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$, the expected excess return on the portfolio is

$$\mathbb{E}[r_p] - r_f = \beta_p (\mathbb{E}[r_m] - r_f), \quad (10.3)$$

where

$$\beta_p = \frac{\text{Cov}[r_p, r_m]}{\text{Var}[r_m]} = \sum_{i=1}^N \pi_i \beta_i. \quad (10.4)$$

Note that the beta of the market portfolio itself equals 1:

$$\beta_m = \frac{\text{Cov}[r_m, r_m]}{\text{Var}[r_m]} = \frac{\text{Var}[r_m]}{\text{Var}[r_m]} = 1.$$

Also note that we can rewrite the beta in terms of the correlation as follows:

$$\beta_i = \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} = \frac{\text{Corr}[r_i, r_m] \text{Std}[r_i] \text{Std}[r_m]}{\text{Var}[r_m]} = \text{Corr}[r_i, r_m] \frac{\text{Std}[r_i]}{\text{Std}[r_m]}. \quad (10.5)$$

An interesting implication of the CAPM is that the Sharpe ratio of an individual asset divided by the Sharpe ratio of the market portfolio equals the correlation between the asset and the market:

$$\frac{\text{SR}_i}{\text{SR}_m} = \text{Corr}[r_i, r_m], \quad (10.6)$$

which is to be shown in Exercise 10.2. As the correlation is less than or equal to one, it follows that individual assets cannot have a higher Sharpe ratio than the market portfolio if the CAPM holds.

Proof

The claim (a) follows from the considerations explained above. The claim (b) is an easy consequence of (a) and the two-fund separation result in Theorem 7.7. To show (c), recall from Theorem 7.6 that the tangency portfolio has the property that

$$\frac{\mathbb{E}[r_i] - r_f}{\text{Cov}[r_i, r_{\tan}]} = \frac{\mathbb{E}[r_{\tan}] - r_f}{\text{Var}[r_{\tan}]}$$

or, slightly rewritten, that

$$E[r_i] - r_f = \frac{\text{Cov}[r_i, r_{\tan}]}{\text{Var}[r_{\tan}]} (E[r_{\tan}] - r_f)$$

for every risky asset i . By (i), the tangency portfolio is the market portfolio and therefore

$$E[r_i] - r_f = \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} (E[r_m] - r_f)$$

for every risky asset i . Due to the definition of the market beta in (10.2), this is exactly the CAPM equation (10.1).

To show Part (d), first recall from (2.22) that the portfolio return is given by $r_p = \sum_{i=1}^N \pi_i r_i$. The covariance of the portfolio return with the market return is

$$\text{Cov}[r_p, r_m] = \text{Cov} \left[\sum_{i=1}^N \pi_i r_i, r_m \right] = \sum_{i=1}^N \pi_i \text{Cov}[r_i, r_m],$$

cf. (3.45), so the market beta of the portfolio is

$$\beta_p = \frac{\text{Cov}[r_p, r_m]}{\text{Var}[r_m]} = \frac{\sum_{i=1}^N \pi_i \text{Cov}[r_i, r_m]}{\text{Var}[r_m]} = \sum_{i=1}^N \pi_i \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} = \sum_{i=1}^N \pi_i \beta_i.$$

From (3.43), the expected portfolio return is

$$E[r_p] = E \left[\sum_{i=1}^N \pi_i r_i \right] = \sum_{i=1}^N \pi_i E[r_i].$$

By substituting in (10.1) on the right-hand side, we get

$$\begin{aligned} E[r_p] &= \sum_{i=1}^N \pi_i (r_f + \beta_i (E[r_m] - r_f)) = \sum_{i=1}^N \pi_i r_f + \sum_{i=1}^N \pi_i \beta_i (E[r_m] - r_f) \\ &= r_f \sum_{i=1}^N \pi_i + (E[r_m] - r_f) \sum_{i=1}^N \pi_i \beta_i = r_f + \beta_p (E[r_m] - r_f), \end{aligned}$$

where the last equality is due to the fact that the portfolio weights sum to one and due to (10.4).

Here is an alternative, graphics-based proof of (c), still relying on the observation that the tangency portfolio has to be the market portfolio under the listed assumptions. Consider a portfolio of asset i and the market portfolio. A fraction w of the total investment is invested in asset i so the fraction invested in the market portfolio is $1 - w$. (Of course, asset i is also part of the market portfolio.) The expectation and standard deviation of the return on the portfolio are

$$\mu(w) = w E[r_i] + (1 - w) E[r_m], \tag{10.7}$$

$$\sigma(w) = \sqrt{w^2 \text{Var}[r_i] + (1 - w)^2 \text{Var}[r_m] + 2w(1 - w) \text{Cov}[r_i, r_m]}, \tag{10.8}$$

cf. Equations (4.2)–(4.4). We know from Section 4.1 that by varying w , the points $(\sigma(w), \mu(w))$ trace out a hyperbola in the usual diagram. This hyperbola goes through the point corresponding to the tangency/market portfolio since $w = 0$ gives that portfolio, and as the hyperbola is produced by feasible portfolios of risky assets it must be to the right of the curved mean-variance frontier of risky assets.

Figure 10.1 illustrates the situation. The blue curve is the efficient frontier of risky assets. The red wedge is the efficient frontier of all assets, which is tangent to the blue curve in the point corresponding to the tangency portfolio, which under our assumptions is identical to the market portfolio. The green curve shows the pairs of expected return and standard deviation that can be obtained by combining the market portfolio and asset i .

If we change w a little, what is the change in the expected return and standard deviation of the portfolio? Differentiation with respect to w gives

$$\mu'(w) = E[r_i] - E[r_m] \quad (10.9)$$

and

$$\begin{aligned} \sigma'(w) &= \frac{1}{2\sigma(w)} (2w \operatorname{Var}[r_i] - 2(1-w) \operatorname{Var}[r_m] + 2(1-2w) \operatorname{Cov}[r_i, r_m]) \\ &= \frac{1}{\sigma(w)} (w \operatorname{Var}[r_i] - (1-w) \operatorname{Var}[r_m] + (1-2w) \operatorname{Cov}[r_i, r_m]) \\ &= \frac{1}{\sigma(w)} (\operatorname{Cov}[r_i, r_m] - \operatorname{Var}[r_m] + w [\operatorname{Var}[r_i] + \operatorname{Var}[r_m] - 2 \operatorname{Cov}[r_i, r_m]]). \end{aligned} \quad (10.10)$$

The ratio $\mu'(w)/\sigma'(w)$ is the marginal effect of a small change in w on the mean-to-standard deviation ratio of the portfolio. Graphically, this is the slope of the line tangent to the hyperbola in the point $(\sigma(w), \mu(w))$.

In particular, for $w = 0$, the ratio $\mu'(0)/\sigma'(0)$ gives the slope of the tangent to the hyperbola through $(\sigma(0), \mu(0))$. We have

$$\mu'(0) = E[r_i] - E[r_m]$$

and

$$\sigma'(0) = \frac{1}{\sigma(0)} (\operatorname{Cov}[r_i, r_m] - \operatorname{Var}[r_m]) = \frac{1}{\operatorname{Std}[r_m]} (\operatorname{Cov}[r_i, r_m] - \operatorname{Var}[r_m]),$$

using the fact that $\sigma(0) = \sqrt{\operatorname{Var}[r_m]} = \operatorname{Std}[r_m]$. Hence, the slope is

$$\frac{\mu'(0)}{\sigma'(0)} = \frac{E[r_i] - E[r_m]}{\operatorname{Cov}[r_i, r_m] - \operatorname{Var}[r_m]} \operatorname{Std}[r_m]. \quad (10.11)$$

But we also know that the capital market line is a tangent to the mean-variance frontier of risky assets and goes through the point corresponding to the tangency/market portfolio. The slope of the capital market line is $(E[r_m] - r_f) / \operatorname{Std}[r_m]$. In any given point, the tangent to a smooth curve is unique, so the two tangents identified above are identical. In particular, the slopes are identical, i.e.,

$$\frac{E[r_i] - E[r_m]}{\operatorname{Cov}[r_i, r_m] - \operatorname{Var}[r_m]} \operatorname{Std}[r_m] = \frac{E[r_m] - r_f}{\operatorname{Std}[r_m]}.$$

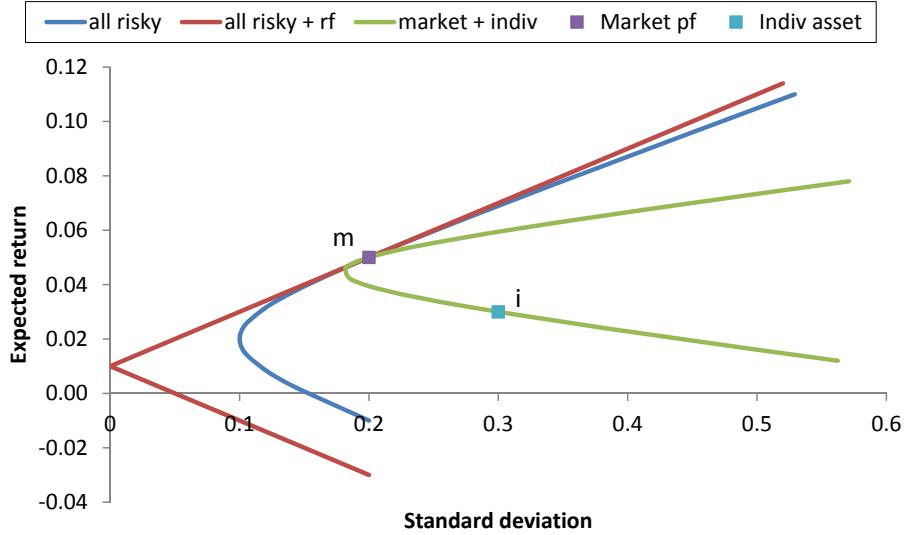


Figure 10.1: Proof of the CAPM.

The riskfree rate is 1%, the market portfolio (purple square) has an expected return of 5% and a standard deviation of 20%, and asset i (blue square) has an expected return of 3% and a standard deviation of 30%. The correlation between the market portfolio and asset i is 0.2. The efficient frontier of risky assets (blue curve) is drawn assuming the minimum-variance portfolio has an expected return of 2% and a standard deviation of 10%.

Solving for $E[r_i]$, we get

$$\begin{aligned} E[r_i] &= \frac{E[r_m] - r_f}{(\text{Std}[r_m])^2} (\text{Cov}[r_i, r_m] - \text{Var}[r_m]) + E[r_m] \\ &= \frac{E[r_m] - r_f}{\text{Var}[r_m]} (\text{Cov}[r_i, r_m] - \text{Var}[r_m]) + E[r_m] \\ &= (E[r_m] - r_f) \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} - (E[r_m] - r_f) + E[r_m] \\ &= \beta_i (E[r_m] - r_f) + r_f, \end{aligned}$$

where β_i is given by (10.2). Now the CAPM-equation (10.1) follows immediately.

Recall that for an investor with a linear mean-variance trade-off, Eq. (7.55) shows the optimal fraction of wealth invested in the tangency portfolio. Under the CAPM assumptions, the tangency portfolio is the market portfolio, so such an investor optimally weighs the market portfolio by

$$w_j = \frac{E[r_m] - r_f}{\gamma_j \text{Var}[r_m]}, \quad (10.12)$$

where γ_j measures the relative risk aversion of investor j . Stated slightly imprecisely, the average investor must hold the market portfolio, otherwise the market would not be in

equilibrium. Hence,

$$1 = \frac{E[r_m] - r_f}{\bar{\gamma} \text{Var}[r_m]} \Rightarrow E[r_m] - r_f = \bar{\gamma} \text{Var}[r_m], \quad (10.13)$$

where $\bar{\gamma}$ is the average relative risk aversion across investors. The market risk premium is thus increasing in the amount of risk (the variance) and average aversion to risk, as we would expect. If we substitute (10.13) back into (10.12), we find that

$$w_j = \frac{\bar{\gamma}}{\gamma_j}, \quad (10.14)$$

which says that your investment decision is fully determined by how your risk aversion relates to the average risk aversion across investors. Of course, this statement is based on a number of assumptions.

The next example illustrates how to use the CAPM for portfolios.

Example 10.1

Suppose the riskfree rate is $r_f = 5\%$ and the expected market return is $E[r_m] = 9\%$. You are considering investing in two stocks ‘XYZ’ and ‘UTL’ and in gold. The market betas of the three assets are $\beta_{\text{XYZ}} = 1.2$, $\beta_{\text{UTL}} = 0.5$, and $\beta_{\text{gold}} = -0.2$, respectively. Assuming that the CAPM holds, what is the expected rate of return on a portfolio with 50% in XYZ, 30% in UTL, and 20% in gold?

Here is one way to answer this question. First, use the CAPM equation (10.1) to find the expected returns on the three assets:

$$\begin{aligned} E[r_{\text{XYZ}}] &= 5\% + 1.2 \times (9\% - 5\%) = 9.8\%, \\ E[r_{\text{UTL}}] &= 5\% + 0.5 \times (9\% - 5\%) = 7\%, \\ E[r_{\text{gold}}] &= 5\% - 0.2 \times (9\% - 5\%) = 4.2\%. \end{aligned}$$

Then the expected rate of return on the portfolio is the weighted average

$$E[r_p] = 0.5 \times 9.8\% + 0.3 \times 7\% + 0.2 \times 4.2\% = 7.84\%.$$

Alternatively, compute the beta of the portfolio using (10.4),

$$\beta_p = 0.5 \times 1.2 + 0.3 \times 0.5 + 0.2 \times (-0.2) = 0.71,$$

and then use the CAPM equation (10.1) to find

$$E[r_p] = 5\% + 0.71 \times (9\% - 5\%) = 7.84\%.$$

10.1.2 Relevant risk and performance measures according to the CAPM

According to the CAPM, the risk premium on any risky asset (the expected rate of return above the riskfree rate) equals the product of the market beta of the asset and the market risk premium. Hence, the market beta is the only asset-specific determinant of the asset’s risk premium. From the definition of the beta in (10.2), we see that it is really the covariance of the asset’s return with the market return that determines the risk premium. In this sense *the correct risk measure for each individual asset is its market*

beta or, equivalently, its covariance with the market return. For example, the standard deviation or variance of the asset return is in itself unimportant, and so is the covariance of the asset return with anything else than the market return.

Why is the covariance with the market return the appropriate risk measure for individual assets? In Eq. (7.45), we computed the marginal change in the portfolio variance when changing the weight on an individual asset a little bit. Under the assumptions of the CAPM, each investor holds the market portfolio and the precise relation is then

$$\frac{\partial \text{Var}[r_m]}{\partial \pi_i} = 2 \text{Cov}[r_i, r_m]. \quad (10.15)$$

If you would consider having a bit more of asset i than its share of the market portfolio, the covariance $\text{Cov}[r_i, r_m]$ measures the increase in your risk. If, and only if, the expected return $E[r_i]$ satisfies the CAPM equation (10.1), no investor would invest more or less in asset i than its market share, and hence the demand for asset i equals the supply.

By definition the average investor holds the market portfolio. An asset with a positive covariance with the market portfolio is, other things equal, quite unattractive to the average investor. The asset tends to provide high returns when the market is already giving a high return, and it typically provides low returns when the market is down. The asset cannot offer any protection against market downturns. Hence, to attract investors, the asset must have a fairly low price, and thus a fairly high expected return.

Conversely, an asset with a negative covariance with the market portfolio constitutes a hedge against bad times. The asset tends to give high returns when the investors appreciate them the most, that is when the market as a whole is going down. Hence, such an asset is attractive, and it will have a high price, and thus a low expected return. In fact, risk-averse investors would be willing to accept a lower expected return than the riskfree rate because the asset acts as an insurance.

If the market beta β_i is the appropriate risk measure for any individual asset i , the risk-return tradeoff is not really captured by the Sharpe ratio $(E[r_i] - r_f)/\text{Std}[r_i]$ as defined in (3.19), but rather by the so-called **Treynor ratio**

$$\text{TR}_i = \frac{E[r_i] - r_f}{\beta_i}. \quad (10.16)$$

Obviously, if the CAPM is true, the Treynor ratio of any asset would equal $E[r_m] - r_f$. If the CAPM does not hold, the deviation of the excess expected return is known as the asset's **alpha**,

$$\alpha_i = E[r_i] - r_f - \beta_i (E[r_m] - r_f). \quad (10.17)$$

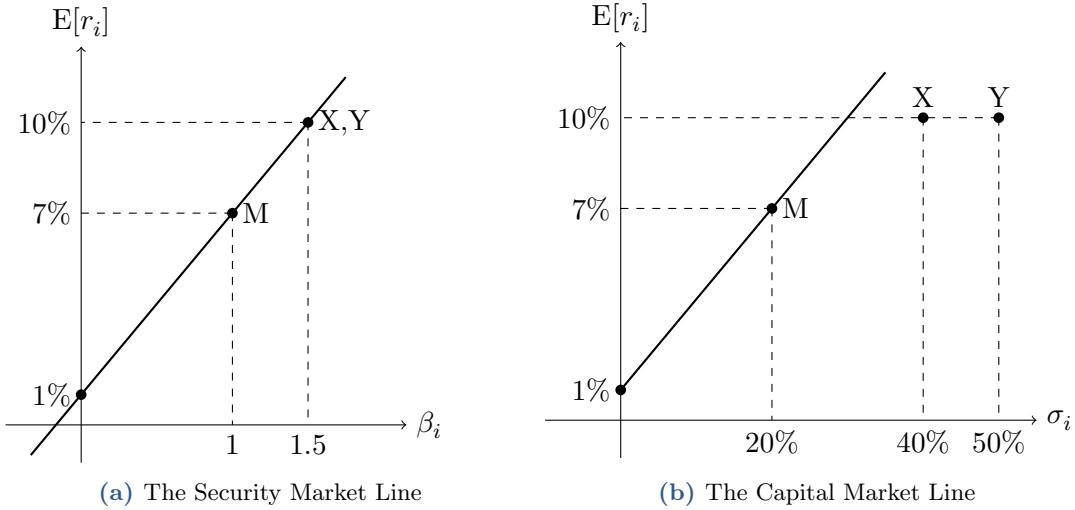
With a non-zero alpha, the Treynor ratio becomes

$$\text{TR}_i = \frac{\alpha_i}{\beta_i} + E[r_m] - r_f.$$

10.1.3 The Security Market Line

In a diagram with market betas along the horizontal axis and expected returns along the vertical axis, the CAPM equation (10.1) corresponds to a straight line intercepting the vertical axis at r_f and having a slope equal to the market risk premium, $E[r_m] - r_f$. This line is called the **Security Market Line**.

Figure 10.2 compares the Security Market Line (SML) and the Capital Market Line (CML) in a simple numerical example. The riskfree rate is 1%, and the market portfolio

**Figure 10.2: CML vs. SML.**

The left graph shows the Security Market Line (SML), and the right graph shows the Capital Market Line (CML). M denotes the market portfolio, whereas X and Y represent individual assets.

has an expected rate of return of 7%, a standard deviation of 20%, and—as always—a beta of 1. The equation of the SML is thus $E[r_i] = 1\% + \beta_i \times 6\%$ as shown in the left panel. The CML is located in the diagram with standard deviation along the horizontal axis, as illustrated in the right panel. The CML meets the vertical axis at the riskfree rate and goes through the point corresponding to the market portfolio. The slope of the CML equals the Sharpe ratio of the market portfolio, $(7\% - 1\%) / 20\% = 0.3$ so the equation for the CML is thus $E[r_i] = 1\% + 0.3 \times \sigma_i$.

Let X and Y denote individual stocks. X has a standard deviation of 40% and a market beta of 1.5, whereas Y has a standard deviation of 50% and also a market beta of 1.5. If the assets are priced according to the CAPM, they are located on the SML and have expected returns of $1\% + 1.5 \times 6\% = 10\%$. As we can see in the right panel, both X and Y are located to the right of the CML. Only efficient portfolios—combinations of the market portfolio and the riskfree asset—are located on the CML. An efficient portfolio with an expected return of 10% would have a standard deviation of only 30%. This is the case for a portfolio with 150% in the market portfolio and −50% in the riskfree asset. Both X and Y carry some non-systematic risk and are thus not on the efficient frontier.

If the point $(\beta_i, E[r_i])$ corresponding to some asset i plots on the Security Market Line, the asset is priced in accordance with the CAPM. If the point plots below [above] the line, the asset is overpriced [underpriced] according to the CAPM. The vertical distance between the point and the line is exactly the alpha of the asset, cf. Eq. (10.17). As a simple example, consider Figure 10.3. Again, the riskfree rate is assumed to 1%, and the expected rate of return on the market portfolio is 7%. The equation for the SML is thus $E[r_i] = 1\% + \beta_i \times 6\%$. Asset A has a market beta of 0.5, so the expected rate of return should be 4% if the CAPM holds. If the expected rate of return is really 6%, asset A has an alpha of $6\% - 4\% = 2\%$. Asset B has a market beta of 1.5, which according to the CAPM corresponds to an expected rate of return of 10%. If the true expected return is 9%, asset B has an alpha of $9\% - 10\% = -1\%$.

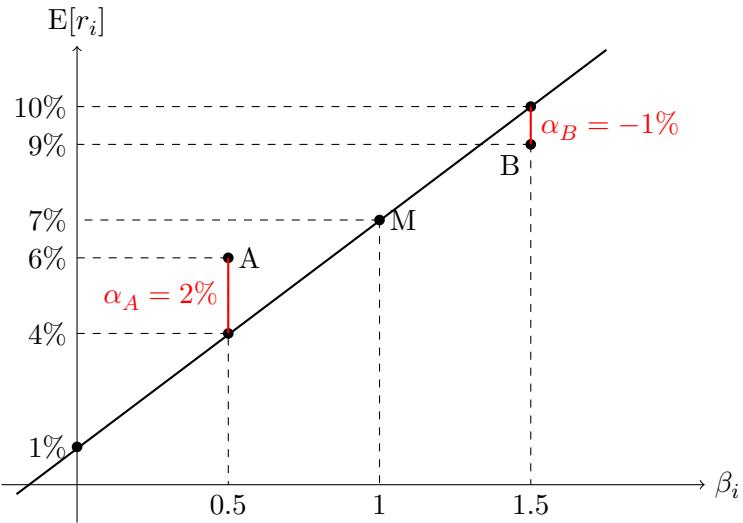


Figure 10.3: SML and alpha.

The upward-sloping black line is the Security Market Line (SML). Assets corresponding to points on the SML are priced in accordance with the CAPM. Assets corresponding to points above the SML (like asset A) are underpriced by the CAPM and have a positive alpha. Assets corresponding to points below the SML (like asset B) are overpriced by the CAPM and have a negative alpha.

10.1.4 Optimal investments according to the CAPM

According to the CAPM, any investor should hold a combination of the riskfree asset and the market portfolio of all risky assets. If we approximate the market portfolio by a broad stock market index, you just need to hold a mix of the riskfree asset and an ETF tracking that index. The portfolios of different investors differ only with respect to the portfolio weights of the market portfolio and the riskfree asset. More risk-averse investors naturally invest less in the market portfolio and more in the riskfree asset.

The average investor must hold the market portfolio so if you are an average investor, the CAPM investment prescription makes good sense. What if you are different from the average investor? Well, under the assumptions of the CAPM, investors only differ in their degree of risk aversion. Investors are required to have the same investment horizon and the same beliefs about investment opportunities. These assumptions are quite restrictive and unrealistic, and if we relax these assumptions, the optimal portfolio may very well depend on the investment horizon and the beliefs of the investor. Furthermore, investors might face different constraints which may also explain differences in their portfolios.

Later in this chapter and in subsequent chapters we consider some extensions of the CAPM that may lead to more variations in optimal portfolios across investors. However, it is important to remember that, whether one or the other model is correct, the average investor must hold the market portfolio, so deviations in your portfolio from the market portfolio should be explained by how you differ from the average investor.

10.1.5 Determining the market beta of an asset

The market beta of asset i is defined in (10.2) as its return covariance with the market, $\text{Cov}[r_i, r_m]$, divided by the return variance of the market, $\text{Var}[r_m]$. How do we determine these numbers?

One suggestion is to set up a number of possible future scenarios with associated probabilities. In each scenario the returns of asset i and the market portfolio have to be specified. Given this assumed joint probability distribution of the random variables r_i and r_m , you can then calculate $\text{Cov}[r_i, r_m]$ and $\text{Var}[r_m]$ and, thus, β_i . See Example 3.4 for further information. However, setting the scenario-specific returns and the probabilities is far from straightforward and some degree of subjective guessing cannot be avoided. The resulting beta-value is sensitive to the specific assumptions and will thus be associated with significant uncertainty.

A more common approach is to rely on historical return observations. Given a time-series of pairwise observations (r_{it}, r_{mt}) , we can estimate the sample covariance and the sample market variance and then calculate a beta-estimate from these sample moments. If you regress stock returns r_{it} on market returns r_{mt} , the estimate of the regression coefficient is, in fact, equal to the sample covariance between r_i and r_m divided by the sample variance of r_m . The regression is often implemented using excess returns, i.e.

$$r_{it} - r_{ft} = \alpha_i + \beta_i (r_{mt} - r_{ft}) + \varepsilon_{it}, \quad t = 1, \dots, T, \quad (10.18)$$

where ε_{it} is a mean-zero noise term. While the riskfree rate is known over the next period, it can still vary from period to period. In fact, the CAPM claims that the expected excess return on asset i , $E[r_i] - r_f = E[r_i - r_f]$, equals the product of β_i and the expected excess return on the market, $E[r_m] - r_f = E[r_m - r_f]$. And since with a known r_f we have

$$\text{Cov}[r_i - r_f, r_m - r_f] = \text{Cov}[r_i, r_m], \quad \text{Var}[r_m - r_f] = \text{Var}[r_m],$$

we can also express the theoretical market beta of asset i as

$$\beta_i = \frac{\text{Cov}[r_i - r_f, r_m - r_f]}{\text{Var}[r_m - r_f]}. \quad (10.19)$$

In terms of excess returns, the beta estimate is then

$$\hat{\beta}_i = \frac{\sum_{t=1}^T (r_{it} - r_{ft} - \overline{r_i - r_f})(r_{mt} - r_{ft} - \overline{r_m - r_f})}{\sum_{t=1}^T (r_{mt} - r_{ft} - \overline{r_m - r_f})^2}, \quad (10.20)$$

where T is the number of return observations used and

$$\overline{r_i - r_f} = \frac{1}{T} \sum_{t=1}^T (r_{it} - r_{ft}), \quad \overline{r_m - r_f} = \frac{1}{T} \sum_{t=1}^T (r_{mt} - r_{ft})$$

are the sample averages of the excess returns. The beta estimate in (10.20) is identical to the estimate of the slope β_i in the simple linear time-series regression (10.18). For a given value of α_i and β_i , we can form a time series of residuals

$$\varepsilon_{it} = r_{it} - r_{ft} - [\alpha_i + \beta_i (r_{mt} - r_{ft})], \quad t = 1, \dots, T,$$

and the estimates of α_i and β_i are the values that minimize the sum of squared residuals $\sum_{t=1}^T \varepsilon_{it}^2$. It can be shown that this produces the β_i -estimate stated in (10.20), whereas the estimate of α_i is

$$\hat{\alpha}_i = \overline{r_i - r_f} - \hat{\beta}_i \times \overline{r_m - r_f}. \quad (10.21)$$

According to the CAPM, α_i should not be significantly different from zero. With these

estimates, the residuals are

$$\hat{\varepsilon}_{it} = r_{it} - r_{ft} - [\hat{\alpha}_i + \hat{\beta}_i (r_{mt} - r_{ft})],$$

and the estimate of $\text{Var}[\varepsilon_i]$ is then simply the sample variance of the time series $\hat{\varepsilon}_{i1}, \dots, \hat{\varepsilon}_{iT}$. In Excel, the linear regression can be performed by choosing ‘Data Analysis’ under the ‘Data’ tab, and then choosing ‘Regression’. Most analysts in the financial industry seem to base the estimates either on 60 monthly return observations or a year of daily return observations.

Example 10.2

Let us perform the regression (10.18) for Microsoft stocks using 60 monthly observations of excess returns from January 2019 to December 2023. The data and the results of the regression analysis can be found in the Excel file `MicrosoftRegressions.xlsx` in the supplementary material to these lecture notes. The excess market returns and the riskfree rates are downloaded from the homepage of Professor Kenneth French, see http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, and the monthly returns on Microsoft are from CRSP, all downloaded on April 17, 2024. The circles in Figure 10.4 represent the 60 pairs of excess returns, and the line shown is the best straight line describing the relation between Microsoft and the market in the linear-regression sense.

The regression output in Excel is shown in Figure 10.5 with comments linked to the more important cells. The main findings are:

1. The estimate of β is 0.8223.

This quantifies the sensitivity of Microsoft to the market. If the excess market return increases by one percentage point, the excess return on Microsoft increases on average by 0.8223 percentage points.

The p-value associated with the estimate is shown as 0.0000 and is really 8.14×10^{-11} , meaning that it is extremely unlikely that the true β is zero. This is also reflected by the 95% confidence interval for β , which is [0.6146, 1.0300] and thus far from zero.

As you expect stocks to show some sensitivity to market movements, zero is not really an interesting benchmark for β -values. You can test whether the beta is statistically significant from any given value β^* by comparing the test statistic

$$t\text{-stat} = \frac{\hat{\beta}_i - \beta^*}{\text{standard error of } \hat{\beta}_i} \quad (10.22)$$

to the typical cut-off level of 2. For example, to test whether Microsoft’s β is different from one, we compute

$$t\text{-stat} = \frac{0.8223 - 1}{0.1037} = -1.7129.$$

As the absolute value of the test statistic is smaller than 2, we conclude that β is not significantly different from 1. This can also be seen from the fact that the 95% confidence interval includes 1.

2. The estimate of α is 1.3655% per month.

This is the part of the excess expected return on Microsoft stocks that cannot be attributed to the sensitivity to general stock market movements. Using simple annual-

ization, this corresponds to a substantial extra return of 16.4% per year.

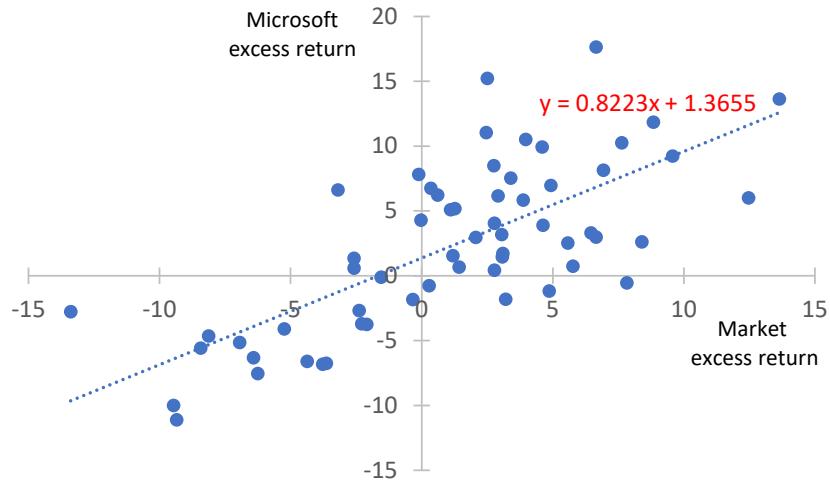
The alpha estimate is even statistically significant which is relatively rarely seen in analyses like this. The p-value associated with the estimate is 0.0231, which means that if the true α was zero, there would only be a 2.31% probability of obtaining the estimate we got. Moreover, the 95% confidence interval for the true α goes from 0.1945% to 2.5365%.

The significantly positive α indicates that Microsoft has performed better than predicted by the CAPM over the 60-month period used. A stock with a positive α is generally a great investment opportunity. However, bear in mind that the large α for Microsoft is just an estimate. This past outperformance of Microsoft might not continue in the future. Maybe lots of unexpected good news about Microsoft came out in the estimation period, and this might be different in any future period.

Focusing on the beta estimate, the wide confidence interval means a lot of uncertainty about the true, unknown market beta of Microsoft. To improve precision, we could include extra data points. Microsoft stocks have been exchange traded since mid March 1986. When using all monthly returns from January 1987 to December 2023, the point estimate of beta is 1.1677 with a 95% confidence interval ranging from 1.0116 to 1.3238. As expected, the confidence interval is narrower when using 378 observations than when using 60 observations, but in this case only slightly so because the earlier returns of Microsoft were more volatile and gave a lower R^2 in the regression. Also note that Microsoft and its market exposure may have changed over the years, so including older observations may contaminate the estimation of the current beta value. In particular, the beta estimate based on the recent 60 months falls outside the confidence interval stemming from the long sample. This suggests that Microsoft's beta has changed over time, which is confirmed by Figure 10.6. The blue curve shows the trailing 60-month beta estimate, i.e., the beta estimate based on the preceding 60 months. Clearly, the 60-month beta estimate has varied considerably, reaching a maximum of 1.73 in January 2001 and a minimum of 0.73 in July 2021. Using the most recent 60 months in the estimation is generally considered a reasonable compromise between the number of observations and the relevance of them but, obviously, this is debatable.

In the example, the confidence interval for Microsoft's beta based on a recent 60-month period goes from 0.6146 to 1.0300. While this interval may seem wide, it is narrow compared to other individual stocks, which may be due to the fact that Microsoft is a large stock (in April 2024, Microsoft is the largest stock in the S&P 500 with a weight of 7.2%) and that the regression works well for Microsoft in terms of a large R^2 , meaning that the residuals are fairly small. If you run a similar regression for other stocks, you often get less precise estimates. For example, running the monthly regression for the Kraft-Heinz Company (in April 2024, number 271 in the S&P 500 ranked by market capitalization) for the same 2019-2023 period, the beta estimate of 0.6549 comes with a wide 95% confidence interval of [0.2571, 1.0528], and the alpha estimate of -0.3500% has a wide confidence interval of [-2.5932%, 1.8932%].

For portfolios you will often obtain a larger precision. As one example, consider a regularly updated portfolio of the 10% stocks with largest market capitalization (see Section 6.6.2 for more on the portfolio formation). Based on monthly return observations downloaded from the homepage of Professor Kenneth French, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, the beta-estimate of

**Figure 10.4: Microsoft vs. market.**

The graph shows 60 pairs of excess monthly returns in percent on Microsoft and the U.S. stock market over the period from January 2019 to December 2023.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.7211
R Square	0.5200
Adjusted R Square	0.5117
Standard Error	4.4302
Observations	60
ANOVA	
df	
Regression	1
Residual	58
Total	59
Coefficients	
Intercept	1.3655
Mkt-RF	0.8223

Annotations explaining the output:

- Correlation between excess returns on Microsoft and market**: Points to the Multiple R value (0.7211).
- 52% of the variation in excess returns on Microsoft is explained by market movements. Equals "Regression-SS" divided by "Total-SS" (see below). Also equal to the square of the correlation.**: Points to the R Square value (0.5200).
- Estimate of the residual std dev.**: Points to the Standard Error value (4.4302).
- = beta^2 * Var[r_m] * (T-1) where T is number of observations**: Points to the Regression SS value (1233.0541).
- Estimate of residual variance**: Points to the Residual SS value (1138.3255).
- = (T-1) * sample variance of Microsoft = sum of squares of deviations from average**: Points to the Total SS value (2371.3796).
- Estimate of beta**: Points to the Intercept coefficient (1.3655).
- Estimate of alpha**: Points to the Mkt-RF coefficient (0.8223).
- t-stat equals coefficient estimate divided by its standard error. Coefficient is significantly different from 0 if t-stat is larger than 2.**: Points to the t Stat values (2.3342 and 7.9263).
- p-value = probability of obtaining the estimated parameter value if the true parameter value is 0.**: Points to the P-value values (0.0231 and 0.0000).
- Bounds of the 95% confidence interval for the true parameter value**: Points to the Lower 95% and Upper 95% values (0.1945, 2.5365 and 0.6146, -1.0300).

Figure 10.5: Microsoft vs. market.

The Excel output from a regression of excess monthly Microsoft returns on excess monthly U.S. stock market returns over the period from January 2019 to December 2023.



Figure 10.6: Microsoft's market beta.

For each month, the blue curve shows the market beta estimate of Microsoft based on the preceding 60 months of returns. The solid orange line shows the beta estimate and the dotted orange lines the 95% confidence interval based on the full return sample from January 1987 to December 2023.

this large-stock portfolio is 0.9550 with the 95% confidence interval ranging from 0.9191 to 0.9909 if you use the 60 observations from January 2019 to December 2023. The running 60-month beta estimate varies much less for the large-stock portfolio than for Microsoft. Note, however, that the large-stock portfolio has a large overlap with the market portfolio used on the right-hand side of the regression, which can be problematic. As another example, take a portfolio of stocks in the manufacturing industry. For the 60 months up to December 2023, the beta estimate is 1.1394 with the confidence interval going from 1.0158 to 1.2631, which is a little narrower than the interval for Microsoft's beta over the same period. The increase in precision obtained by going from a single stock to a portfolio depends heavily on the nature of the portfolio.

The main purpose of a beta estimate is to predict how the returns on a stock is going to covary with the market returns in a future period. Maybe the regression-based beta estimate is not the best predictor of the future beta? Many analysts adjust the regression estimate $\hat{\beta}_i$ of the market beta using a simple rule of the form

$$\tilde{\beta}_i = w \times \hat{\beta}_i + (1 - w) \times 1 = 1 + w \times (\hat{\beta}_i - 1), \quad (10.23)$$

which is just a weighted average of the original estimate and a value of one. The beta-estimate is thus moved closer to one. Estimates above one are reduced, estimates below one are increased. The weight $w = 2/3$ was apparently first used by Merrill Lynch, the investment bank which is now part of Bank of America, and this weight seems popular and is used, for example, in Bloomberg's database. Maybe this weight is inspired by prior research of Blume (1971).

The adjustment is motivated as follows. First, as indicated by Figure 10.6, the market beta of a stock is likely to vary over time. You could expect the market beta to move towards one over time. This would be consistent with the typical life cycle of companies. A young company tends to sell a few highly specialized products or services, and the market beta is then determined by the sensitivity of the demand for these products or services (and the costs of producing them) to the overall economy and thus the stock

market.¹ When the company grows and matures, it often starts diversifying internally by having a larger and more diverse range of products and selling them to a broader audience. Hence, the company becomes more like the “average” company with a market beta closer to one. Admittedly, a trend towards one is not obvious in the case of Microsoft depicted in Figure 10.6.

Secondly, the further the beta-estimate is away from one, the more likely is it that the estimate is wrong, and that the true beta is closer to one. The largest beta-estimates tend to be over-estimated and the smallest to be under-estimated. Why? We know that the average beta is one. Think of a world in which all stocks had a true beta of one. Based on a time series of stock returns, we would still obtain some beta-estimates above one (they are over-estimated) and some beta-estimates below one (they are under-estimated). Even if the true betas are spread around one, there is a tendency that the largest beta-estimates are over-estimated and the smallest are under-estimated.

While (10.23) adjusts the estimated market beta towards one, you could also consider adjusting it towards the average market beta estimate across all stocks or just the stocks in the industry the company belongs to, cf. [Vasicek \(1973\)](#). [Rosenberg and Guy \(1976a, 1976b\)](#) find that, in addition to the past beta, the industry classification and various firm characteristics can contribute to better predictions of future betas, whereas [Abell and Krueger \(1989\)](#) report that conditioning on macroeconomic variables might also help.

For more on how to estimate betas, see the analysis and discussion in [Hollstein and Prokopczuk \(2016\)](#) and [Welch \(2022\)](#).

10.1.6 CAPM-implied expected returns

As discussed in Section 3.7.2 and elsewhere, time-series average returns of individual assets are very imprecise estimates of the true expected returns. The CAPM can guide our search for reasonable expected return estimates. Of course, if you know the market beta of the asset, or have a precise estimate of it, you can calculate the expected return on the asset directly from the CAPM relation

$$\mathbb{E}[r_i] = r_f + \beta_i (\mathbb{E}[r_m] - r_f).$$

But as seen in Example 10.2 above, traditional beta estimates are often very imprecise.

Here is an alternative approach for determining the expected return on an asset. Recall from Section 7.3 that if risky asset returns are normally distributed, $\mathbf{r} \sim N(\boldsymbol{\mu}, \underline{\Sigma})$, the optimal portfolio vector for an investor with a linear mean-variance tradeoff is

$$\boldsymbol{\pi}^* = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}), \quad (10.24)$$

where r_f is the riskfree rate and γ is the investor’s relative risk aversion. If the CAPM is true, the average investor optimally holds the market portfolio. As above, let $\bar{\gamma}$ denote the average relative risk aversion of investors. Then the vector of market portfolio weights $\boldsymbol{\pi}_{\text{mkt}}$ satisfies the relation

$$\boldsymbol{\pi}_{\text{mkt}} = \frac{1}{\bar{\gamma}} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}). \quad (10.25)$$

If we assume that the variance-covariance matrix is known, we can solve the above equation

¹In addition to the beta of the assets or sales, the beta of a company’s stock is also affected by the leverage of the company.

for μ , and the solution

$$\mu_{\text{mkt}} = r_f \mathbf{1} + \bar{\gamma} \sum \pi_{\text{mkt}} \quad (10.26)$$

is then the vector of expected returns that the average investor must assume in order to find the current market portfolio weights optimal. For an individual asset i , the CAPM-implied expected return estimate is

$$\mu_{i,\text{mkt}} = r_f + \bar{\gamma} \sum_{j=1}^N \Sigma_{ij} \pi_{j,\text{mkt}}, \quad (10.27)$$

where Σ_{ij} is asset i 's return covariance with asset j and $\pi_{j,\text{mkt}}$ is asset j 's market weight.

How do we know the average risk aversion $\bar{\gamma}$? Assuming that we know the expectation and the variance of the market portfolio return, we can back out $\bar{\gamma}$ from (10.13) as

$$\bar{\gamma} = \frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]}. \quad (10.28)$$

When applying the above formulas, remember from Section 7.3 that to avoid confusion you should represent returns as decimals, e.g. a 4% expected return must be written as 0.04.

Example 10.3

Consider a hypothetical stock market where only five stocks—called AA, BB, CC, DD, and EE—are traded and they have identical market capitalization, i.e., each stock has a market weight of 0.2. Suppose the standard deviations and the correlation matrix are as shown in Table 10.2. The variance-covariance matrix is easily calculated from this information.

If this market is in a CAPM equilibrium, what must the expected stock returns be? Suppose the one-year riskfree rate is 1% and the average risk aversion is $\bar{\gamma} = 2$. (For example, this is consistent with the market return having an expectation of 9% with a standard deviation of 20%—numbers in line with the empirical estimates presented in Section 6.5.) Then the vector of expected stock returns follows from an application of (10.26) with the results shown in the right-most gray-shaded column in the table. The stock AA has a low standard deviation of 20% and zero correlation with other stocks, so it only needs an expected return of 2.60% to achieve the 20% market weight. Stock BB has the same standard deviation, but high correlation with several other stocks in the market portfolio so it has to offer a higher expected return of 10.28%. Stock CC has the same correlations as stock BB, but a higher standard deviation, so the expected return on CC must be as high as 30.76%. Stock DD is like stock BB except that it has a negative correlation with stock EE which makes it relatively more attractive. Hence, the expected return on DD is only 6.12%, compared to BB's expected return of 10.28%. Finally, stock EE must offer an expected return of 16.04%.

The implied expected returns depend on which stocks you include in the calculations. Suppose you ignore stock EE and consider only the first four stocks. Each of these stocks will then have a market share of 25% of the part of the market included. The resulting implied stock returns are now shown in the right-most blue-shaded column. For example, the implied expected return on stock DD is now 9.40% instead of the 6.12% in the previous calculation. Since we ignore stock EE, we also ignore the negative correlation of DD with

Stock	StdDev	Correlations					Mkt weights and exp returns			
		AA	BB	CC	DD	EE	π_{mkt}	μ_{mkt}	π_{mkt}	μ_{mkt}
AA	0.2	1.0	0.0	0.0	0.0	0.0	0.2	0.0260	0.25	0.0300
BB	0.2	0.0	1.0	0.8	0.8	0.8	0.2	0.1028	0.25	0.0940
CC	0.6	0.0	0.8	1.0	0.8	0.8	0.2	0.3076	0.25	0.2860
DD	0.2	0.0	0.8	0.8	1.0	-0.5	0.2	0.0612	0.25	0.0940
EE	0.4	0.0	0.8	0.8	-0.5	1.0	0.2	0.1604	N.A.	N.A.

Table 10.2: Inputs and outputs in Example 10.3.

The left part of the table shows standard deviations and correlations of the five stocks in Example 10.3. The right part of the table shows the weights of the assets in the market portfolio and their CAPM-implied expected returns.

EE, so now the diversification potential in DD seems smaller and, therefore, it must offer a higher expected return.

The above procedure for backing out return expectations is often used as part of the so-called Black-Litterman model for active portfolio management that will be outlined in Section 13.2. In that model the market-neutral views on expected returns implied by the CAPM are combined with the manager's subjective views on expected returns.

10.2 The empirical performance of the CAPM

10.2.1 Test methodologies

Roll (1977) emphasized that, strictly speaking, it is impossible to test the CAPM since the true market portfolio is unobservable. This point is now known as "Roll's critique." The true market portfolio should include all risky assets and thus also important assets as real estate and human capital, cf. the discussions in Sections 9.1 and 9.2. In particular the human capital is unobservable and difficult to estimate. Both empirical tests and practical applications of the CAPM have to apply an observable proxy for the market portfolio. The vast majority of tests and applications use a stock market index as a proxy for the market portfolio. You might argue that if applications are using the stock market index to represent the market portfolio, it makes sense that empirical tests also do it so that they really test the model as how it would be used in practice.

The CAPM claims a certain relation between expected returns and betas, but we cannot observe expected returns or betas. In tests, the expected return on an asset or a portfolio is approximated by the average return over a number of periods, and the true beta is approximated by an estimated beta. Even if the CAPM is true, we will not observe the same perfect relation between average returns and estimated betas. But we can test statistically whether the deviations are so small that the CAPM most likely is true or the deviations are so big that the CAPM most likely is wrong.

Many empirical tests of the CAPM and related models apply a two-step procedure. In the first step betas of the assets used in the test are estimated from the time-series regression (10.18), often based on 60 monthly observations as in Example 10.2. The second step is a cross-sectional regression of the average excess returns of the test assets on their betas:

$$\overline{r_i - r_f} = \gamma_0 + \gamma_1 \beta_i + \varepsilon_i, \quad i = 1, 2, \dots, N. \quad (10.29)$$

According to the CAPM, we should find $\gamma_0 \approx 0$ and $\gamma_1 \approx \overline{r_m - r_f}$. In many tests an additional asset-specific right-hand side variable is included in the second step, i.e.,

$$\overline{r_i - r_f} = \gamma_0 + \gamma_1 \beta_i + \gamma_2 X_i + \varepsilon_i, \quad i = 1, 2, \dots, N. \quad (10.30)$$

If the CAPM is correct, we should find $\gamma_2 \approx 0$. The variable X_i could be the square of the beta, the residual variance of the asset estimated in the first step, or some firm-characteristic like the size (market capitalization) or the book-to-market ratio.

While the test procedure seems straightforward, there are a number of practical problems. First, returns are often highly volatile or noisy. Hence, both the beta and especially the average return are measured imprecisely, cf. the discussion in Section 3.7.2. The beta estimates of portfolios can be more precise, as explained above, so a partial solution is to use portfolios instead of individual assets. To get a good estimate of the slope of the beta-return relation, portfolios with dispersed betas should be used. One way to proceed is to perform the first-step regression on each individual asset, rank the assets according to the beta-estimate, and then form, say, 10 portfolios based on the estimated betas with the first portfolio including the stocks with the 10% lowest beta-estimates, etc., so that the tenth portfolio includes the stocks with the 10% highest beta-estimates. The cross-sectional regression is then performed on these 10 portfolios rather than the individual assets. This reduces the number of data points in the second regression but, on the other hand, the inputs to that regression are more reliable.

A related second problem is that the right-hand side variable in the second regression, the β_i , is estimated and thus involve measurement errors. It is well known in statistics that the measurement error in the explanatory variable causes the estimate of the slope to be biased downwards and the intercept to be biased upwards, i.e., the relation appears flatter than it really is. The problem is essentially that, as explained above, betas that are estimated to be high are often overestimated, and betas that are estimated to be low are often underestimated. By using portfolios in the second-step regression, this problem is reduced, but not eliminated. Nevertheless, it remains a fundamental problem in tests and applications of the CAPM that past betas are relatively poor predictors of future betas and, generally, it seems difficult to predict future betas, cf. the discussion and references in Ang (2014, Sec. 10.4.3).

Furthermore, when including right-hand side variables in addition to the beta-estimate, multicollinearity becomes a concern. If the additional variable is correlated with beta, it is hard to disentangle the effects of the two variables from each other, and the estimates of the coefficients γ_1 and γ_2 may be biased, which makes it hard to evaluate whether these coefficients are as predicted by the CAPM.

Fama and MacBeth (1973) introduced a variation to the above procedure in which the cross-sectional regression is performed at each time period, i.e. for every t , they estimate

$$r_{it} - r_{ft} = \gamma_{0t} + \gamma_{1t} \beta_i + \varepsilon_{it}, \quad i = 1, 2, \dots, N. \quad (10.31)$$

Here, β_i is the estimate of the beta of asset i at time t , and they compute that from a time-series regression like (10.18), but using a rolling window of the most recent 60 months. After running (10.31) for all t , they have a time series of estimates of γ_{0t} and γ_{1t} , and they then compare the average intercept $\bar{\gamma}_0 = (\sum_{t=1}^T \gamma_{0t})/T$ to zero and the average slope $\bar{\gamma}_1 = (\sum_{t=1}^T \gamma_{1t})/T$ to the average market risk premium $\overline{r_m - r_f}$.

Another test approach is simply to test for a set of assets or portfolios whether the estimate of α_i in the time-series regression (10.18) is significantly different from zero. See Gibbons, Ross, and Shanken (1989) for an analysis of this approach.

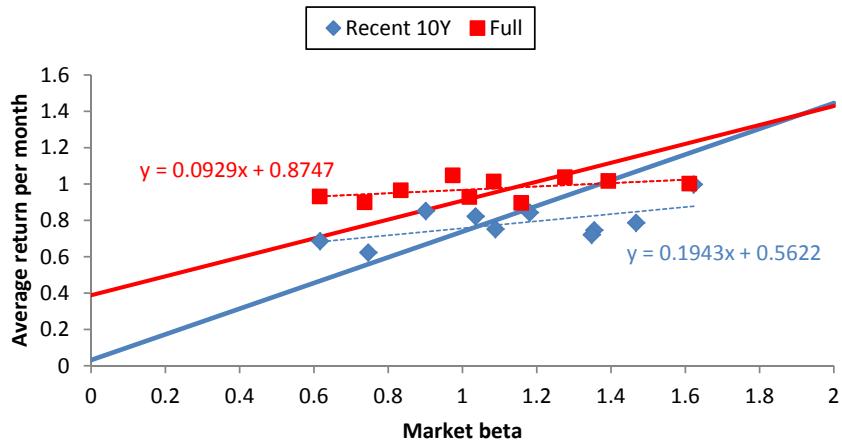


Figure 10.7: CAPM-test: beta-sorted portfolios.

The figure shows the relation between market-betas and average returns. The red squares show the combination of the beta estimate and the average return on ten beta-sorted portfolios as explained in the text. The solid red line shows the relation between beta and expected return suggested by the CAPM using the average observed riskfree rate and excess market return. Monthly observations from July 1963 to August 2017 are used. The blue diamonds and the blue solid line show the same relations but using only data from September 2007 to August 2017.

10.2.2 Test results

Many of the early tests by Black (1972), Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), and others were somewhat supportive of the CAPM, but did find that the relationship between betas and average returns was flatter than predicted by the CAPM. In fact, Haugen and Heins (1975) even found average returns to be decreasing in market betas. Including later return data, Fama and French (1992) and others found that the relationship was virtually flat, prompting many to conclude that “beta was dead.”

Figure 10.7 provides an updated version of these tests based on monthly returns from July 1963 to August 2017 downloaded from the homepage of Professor Kenneth French. Ten portfolios are formed based on the betas of individual stocks estimated from the past 60 monthly returns. The 10% stocks with the lowest beta-estimates are allocated to the first decile portfolio, etc. Then the subsequent monthly returns on each portfolio are calculated. We consider value-weighted portfolios. Portfolios are rebalanced at the end of June each year. A beta for each portfolio is estimated from regressing all its monthly excess returns on the excess market returns. These betas are nicely increasing from portfolio 1 to 10 as you would expect.

The red squares in Figure 10.7 shows the combinations of betas and average returns of the ten beta-sorted portfolios over the full sample period. The dashed red line is the best straight line fitting the squares. As you can see, this line is upward sloping in accordance with the prediction of the CAPM, but it is almost flat. The solid red line shows the relation between beta and expected return predicted by the CAPM if you use the sample average of the one-month riskfree rate as the intercept and the sample average of the monthly excess market returns as the slope. The empirical relation is much flatter than the theoretical relation. Low-beta portfolios appear to have positive α , i.e., to provide a larger average return than can be explained by their market exposure. In contrast, high-beta portfolios

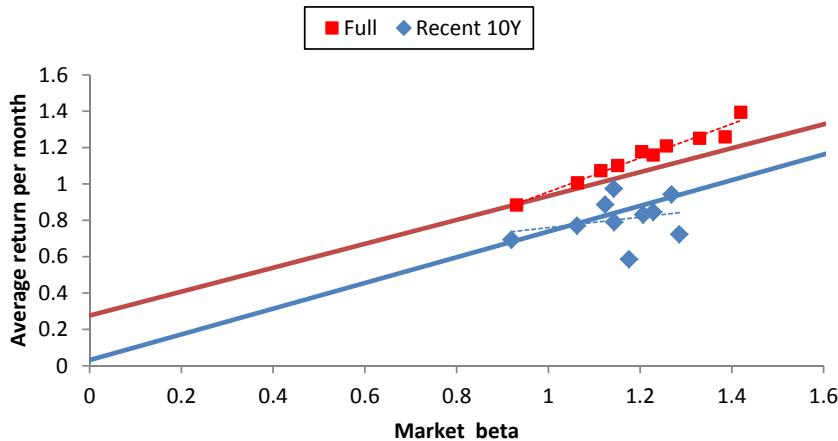


Figure 10.8: CAPM-test: size-sorted portfolios.

The figure shows the relation between market-betas and average returns. The red squares show the combination of the beta estimate and the average return on ten size-sorted portfolios as explained in the text. The solid red line shows the relation between beta and expected return suggested by the CAPM using the average observed riskfree rate and excess market return. Monthly observations from July 1926 to August 2017 are used. The blue diamonds and the blue solid line show the same relations but using only data from September 2007 to August 2017.

have negative α and thus provide lower average returns than expected given their market exposure. The blue diamonds and lines are equivalent to the red, but are based only on the most recent 10 years of observations. While the conclusions from empirical tests will often depend on the sample period used, this does not seem to be the case here.²

Figure 10.8 is similar to Figure 10.7 but is based on decile portfolios sorted according to the market capitalization of each stock. The data period spans July 1926 to August 2017. For the full sample, the estimated portfolio beta is declining monotonically with the market cap as can be seen from Table 10.3. The portfolio of the smallest 10% of stocks has an estimated beta of 1.42, whereas the portfolio of the largest 10% of stocks has a beta of 0.931. In fact, the latter portfolio is the only having a beta less than 1, but the stocks in this portfolio are so big compared to the stocks in the other portfolios that the value-weighted beta of all stocks still equals 1 as required. In the full sample, the average return is also declining with market cap so the positive relation between beta and average return holds as seen by the red squares in Figure 10.8, but the best linear relation is somewhat steeper than predicted by the CAPM, cf. the solid red line. Almost all of the good long-run performance of small-stock portfolios can be explained by the large market exposure of these portfolios as captured by their large beta-estiamte. The large-stock decile portfolio has a slightly negative alpha, whereas all other size portfolios have a small or modestly positive alpha.

Chapter 6 presented evidence that the relation between market capitalization and average return has become less clear in more recent years. This is confirmed by the right part of Table 10.3 and the blue diamonds and lines in Figure 10.8. The small-stock portfolios

²The conclusion could also depend on the data frequency. Indeed, Kothari, Shanken, and Sloan (1995) demonstrated that if annual returns are used instead of monthly returns, the relation between betas and average returns is stronger and thus more supportive of the CAPM, at least for the sample period and assets included in their study.

	Full sample			Recent 10 years		
	Avg ret	Beta	Alpha	Avg ret	Beta	Alpha
Smallest 10%	1.393	1.420	0.184	0.586	1.176	-0.277
Decile-2	1.258	1.386	0.071	0.723	1.285	-0.217
Decile-3	1.250	1.330	0.100	0.943	1.268	0.015
Decile-4	1.209	1.258	0.106	0.832	1.207	-0.053
Decile-5	1.159	1.229	0.075	0.846	1.229	-0.053
Decile-6	1.177	1.204	0.110	0.975	1.142	0.136
Decile-7	1.101	1.151	0.068	0.791	1.143	-0.049
Decile-8	1.073	1.114	0.064	0.887	1.124	0.061
Decile-9	1.006	1.064	0.031	0.771	1.063	-0.012
Largest 10%	0.884	0.931	-0.005	0.694	0.919	0.012

Table 10.3: Statistics on size-sorted portfolios.

The table shows the average monthly return and the estimates of β_i and α_i for ten portfolios sorted on the market capitalization of stocks. Monthly observations are used. The full sample covers the period from July 1926 to August 2017, so the recent 10 year period is September 2007 to August 2017.

still tend to have larger betas, but nevertheless they have delivered mediocre returns in the recent 10-year period, resulting in negative alphas. Overall, the CAPM seems to do a decent job in explaining return differences across size portfolios.

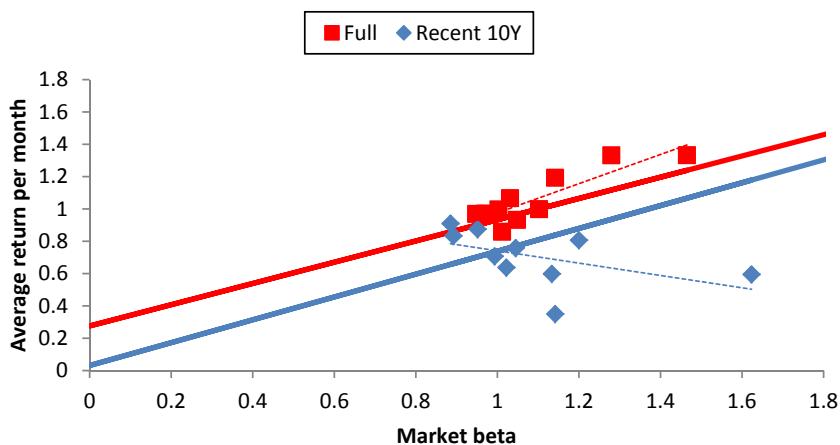
Another stylized fact emphasized in Section 6.6 is that average stock returns depend on the ratio of the book value of equity to the market value of equity of the issuing company. In the period 1926-2017, stocks with a high book-to-market ratio (i.e. value stocks) have delivered a much larger return than stocks with a low book-to-market ratio (i.e. growth stocks), but the relative performance of the two groups of stocks have varied substantially over time with a smaller average return difference in recent decades. This is confirmed by the average return numbers listed in Table 10.4, where stocks have been sorted into ten portfolios based on their book-to-market ratio, with annual rebalancing of the portfolios. The table also reveals that value stocks tend to have larger betas than growth stocks, so the long-run relation between betas and average returns lends some support to the CAPM. However, the average returns on the value stocks have been even higher than what the beta can explain, leading to a positive alpha. In contrast, the portfolio of stocks with the lowest book-to-market ratios have underperformed relative to the CAPM as indicated by the slightly negative alpha. In the most recent 10-year period, the patterns are almost reversed with the growth portfolio achieving the highest average return in spite of the lowest beta, which of course implies a positive alpha. The value stocks have disappointed in the recent period, especially given their high beta. Figure 10.9 illustrates the relation between betas and average returns for the book-to-market portfolios. In particular, the recent sample is challenging for the CAPM but is, on the other hand, also based on much fewer observations so that statistical significance is harder to detect.

These findings are in line with published empirical studies. Fama and French (2006) and Ang and Chen (2007) report that in the period 1926-1963 the value premium on U.S. stocks could be almost perfectly explained by the value stocks having higher market betas than the growth stocks. But for a post-1963 sample, value stocks have lower market betas than growth stocks so that the CAPM cannot explain why value stocks have offered higher returns. In the above analysis, the market beta of each portfolio was assumed to be constant through the sample period, but this is far from correct for portfolios sorted on book-to-market. We return to this issue and its implications for CAPM testing in

	Full sample			Recent 10 years		
	Avg ret	Beta	Alpha	Avg ret	Beta	Alpha
Lo 10	0.859	1.011	-0.082	0.875	0.951	0.171
Decile-2	0.970	0.948	0.071	0.834	0.892	0.172
Decile-3	0.972	0.970	0.058	0.910	0.885	0.253
Decile-4	0.933	1.048	-0.032	0.638	1.022	-0.116
Decile-5	0.998	1.003	0.063	0.758	1.045	-0.011
Decile-6	1.068	1.031	0.114	0.708	0.993	-0.026
Decile-7	0.999	1.102	-0.001	0.350	1.142	-0.488
Decile-8	1.194	1.141	0.167	0.599	1.134	-0.234
Decile-9	1.333	1.279	0.216	0.807	1.200	-0.072
Hi 10	1.334	1.465	0.095	0.595	1.623	-0.583

Table 10.4: Statistics on book-to-market sorted portfolios.

The table shows the average monthly return and the estimates of β_i and α_i for ten portfolios sorted on the book-to-market ratio of stocks. Monthly observations are used. The full sample covers the period from July 1926 to August 2017, so the recent 10 year period is September 2007 to August 2017.

**Figure 10.9: CAPM-test: book-to-market sorted portfolios.**

The figure shows the relation between market-betas and average returns. The red squares show the combination of the beta estimate and the average return on ten portfolios sorted on the book-to-market ratio as explained in the text. The solid red line shows the relation between beta and expected return suggested by the CAPM using the average observed riskfree rate and excess market return. Monthly observations from July 1926 to August 2017 are used. The blue diamonds and the blue solid line show the same relations but using only data from September 2007 to August 2017.

	Full sample					Recent 10 years				
	Grow	BM2	BM3	BM4	Value	Grow	BM2	BM3	BM4	Value
<i>Average monthly return</i>										
Small	0.850	0.972	1.272	1.459	1.633	0.321	0.759	0.633	0.816	0.707
ME2	0.912	1.200	1.261	1.340	1.519	0.991	0.998	0.944	0.773	0.700
ME3	0.986	1.185	1.200	1.301	1.411	0.840	0.969	0.916	0.972	0.915
ME4	0.998	1.028	1.137	1.243	1.314	0.991	0.901	0.605	0.862	0.648
Large	0.901	0.910	0.977	0.933	1.230	0.853	0.755	0.714	0.267	0.701
<i>Beta</i>										
Small	1.630	1.408	1.372	1.269	1.379	1.290	1.243	1.202	1.146	1.243
ME2	1.265	1.227	1.198	1.212	1.377	1.234	1.177	1.182	1.159	1.413
ME3	1.248	1.124	1.122	1.159	1.376	1.182	1.162	1.156	1.174	1.262
ME4	1.092	1.078	1.114	1.155	1.418	1.093	1.120	1.228	1.125	1.268
Large	0.956	0.950	0.967	1.108	1.312	0.872	0.893	0.951	1.160	1.321
<i>Alpha</i>										
Small	-0.498	-0.230	0.094	0.349	0.450	-0.622	-0.151	-0.248	-0.026	-0.203
ME2	-0.196	0.118	0.197	0.267	0.338	0.087	0.135	0.077	-0.078	-0.330
ME3	-0.110	0.169	0.186	0.263	0.230	-0.026	0.116	0.068	0.110	-0.008
ME4	0.004	0.043	0.129	0.208	0.105	0.188	0.078	-0.294	0.035	-0.280
Large	-0.003	0.009	0.065	-0.072	0.091	0.205	0.092	0.011	-0.584	-0.264

Table 10.5: Statistics on book-to-market sorted portfolios.

The table shows the average return, the estimated beta, and the estimated CAPM alpha on 25 double-sorted portfolios. ‘BM’ refers to the book-to-market value of the equity of the company and ‘ME’ to the market capitalization. The left part of the table is based on the full sample covering 1926/07 to 2017/08, the right part only on the most recent 10 year subsample. The data were downloaded from the homepage of Professor Kenneth French, <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>

Section 10.3.3.

If we construct portfolios based on a two-dimensional sort on both market capitalization and the book-to-market ratio as explained in Section 6.6, we obtain the statistics shown in Table 10.5. Note that the average monthly returns displayed in the upper panel of the table are consistent with those in the upper panel of Table 6.13.

In the full sample, small value stocks provided the highest average return, whereas high-growth stocks of all sizes and large stocks with any book-to-market ratio, except the highest, have delivered the lowest average returns. Are the return differences explained by differences in market betas? No! In particular, small growth stocks with high betas exhibit poor average returns so they underperform compared to the CAPM as indicated by their negative alphas shown in the lower panel of the table. While small value stocks do have large betas, the average return more than compensates for the market exposure, leading to positive alphas. Note that the alphas of the five portfolios involving large stocks are all close to zero, so the apparent mispricing is concentrated in the smaller stocks and especially in the very small stocks.

In the recent 10 year sample, the patterns are again less clear. The small growth stocks exhibit a large negative alpha also in this subperiod, but value stocks of all sizes have produced modest returns despite having high betas, resulting in negative alphas.

Next, we turn to the momentum effect. As first identified by Jegadeesh and Titman (1993) and illustrated by updated statistics in Section 6.6, the average return on a stock over the next couple of months tend to positively related to its return over the past year.

	Full sample			Recent 10 years		
	Avg ret	Beta	Alpha	Avg ret	Beta	Alpha
Low PRIOR	0.320	1.558	-0.975	0.401	1.898	-0.972
PRIOR 2	0.696	1.328	-0.448	0.508	1.452	-0.550
PRIOR 3	0.756	1.175	-0.289	0.872	1.233	-0.030
PRIOR 4	0.882	1.093	-0.109	0.920	1.092	0.117
PRIOR 5	0.888	1.038	-0.067	1.135	1.052	0.360
PRIOR 6	0.948	1.030	-0.002	0.907	0.953	0.202
PRIOR 7	1.010	0.970	0.100	0.841	0.917	0.161
PRIOR 8	1.126	0.934	0.239	0.802	0.910	0.127
PRIOR 9	1.193	0.964	0.287	0.636	0.957	-0.071
High PRIOR	1.500	1.023	0.555	0.854	1.070	0.067

Table 10.6: Statistics on portfolios sorted on past returns.

The table shows the average monthly return and the estimates of β_i and α_i for ten portfolios sorted on the return over the past 2-12 months. Monthly observations are used. The full sample covers the period from January 1927 to August 2017, so the recent 10 year period is September 2007 to August 2017.

Stock-market winners tend to stay winners, losers tend to stay losers, at least in the short run.

Can these differences be explained by differences in market betas? Are recent winners more sensitive to market movements than recent losers? As demonstrated by Table 10.6 and Figure 10.10, the answer seems to be no. Similar to the above studies, stocks are sorted into ten portfolios, but now based on their prior returns over the period from 2 to 12 months ago. Then the value-weighted return on each portfolio over the next month is calculated. Portfolios are rebalanced every month.

Consider first the full 1927-2017 sample. Here the average monthly return is monotonically increasing, but the portfolio betas decreasing (except from portfolios 9 and 10), in the prior returns. This produces a steep, negative relation between the betas and the average returns, contradicting the CAPM. Loser portfolios have large negative alphas, winner portfolios have substantial positive alphas. In the more recent 10-year sample, results are similar although less clear-cut. Loser portfolios still stand out by having a large beta and a low average return, and thus a large negative alpha. Overall, the momentum effect in returns constitutes a major challenge to the CAPM.

As a final set of test assets, we take ten decile portfolios sorted on the return variance over the past 60 days as explained in Section 6.6.5. Recall that the main findings from that section was that stocks with high recent variance tend to deliver low subsequent returns compared to stocks with low recent variance. This is confirmed by the average returns for the ten variance-sorted portfolios shown in Table 10.7. Again such differences could be in line with the CAPM if they reflect differences in market betas. In fact, if we exclude the two portfolios with the highest variance stocks and consider the full sample, the average return is positively related to the beta estimate. But the high-variance portfolios have high market betas and low average returns. As illustrated by Figure 10.11, the best overall linear relation between betas and average returns is downward sloping. Portfolios of low-variance stocks have positive alphas, whereas portfolios of stocks with very high recent variance have large negative alphas. For the recent 10-year sample, the conclusions are similar.

As indicated by the above analysis, the CAPM apparently fails to explain return differences along some dimensions. Therefore, it is not surprising that many empirical tests

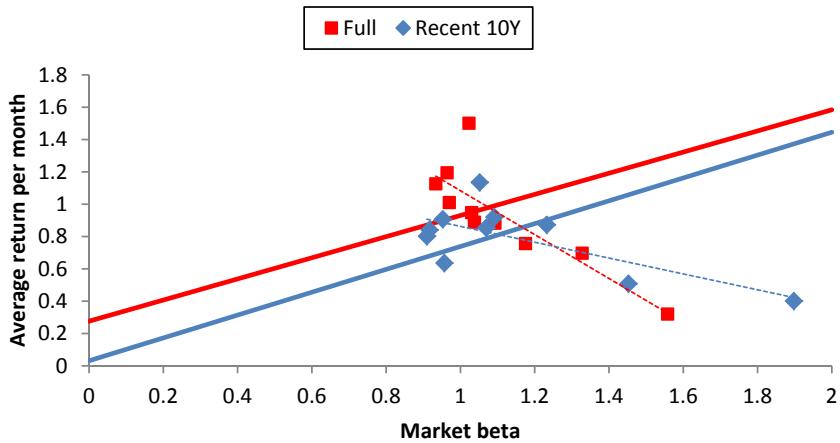


Figure 10.10: CAPM-test: portfolios sorted on past returns.

The figure shows the relation between market-betas and average returns. The red squares show the combination of the beta estimate and the average return on ten portfolios sorted on the return over the past 2-12 months as explained in the text. The solid red line shows the relation between beta and expected return suggested by the CAPM using the average observed riskfree rate and excess market return. Monthly observations from 1927/01 to 2017/08 are used. The blue diamonds and the blue solid line show the same relations but using only data from 2007/09 to 2017/08.

conclude that other variables than the market beta are significant in regressions with average asset or portfolio returns as the left-hand side variable. In other words, a number of specifications of X_i appear to be statistically significant in the cross-sectional regression (10.30) or when adding such an X_i -term to the Fama-MacBeth regression (10.31). Some obvious candidates for significant explanatory variables are variables related to the portfolio characteristics that CAPM cannot handle well, such as size, book-to-market, and momentum.

A few studies have tried to adjust for the fact that the stock market index is not the true market portfolio and ignores such important assets as human capital and real estate. Among others, Campbell (1996) and Jagannathan and Wang (1996) show that when a factor related to changes in human capital is added as an explanatory variable in the cross-sectional regression along with the stock index return, the model performs much better. These findings suggest that the apparent failure of the CAPM in empirical tests is due—at least partially—to the market proxy ignoring human capital.

Chapter 11 has much more information about models adding factors to the market return in order to better explain the cross section of expected returns.

10.3 Alternative versions of the CAPM

10.3.1 A CAPM with borrowing constraints

Suppose no riskfree asset exists. Each investor then optimally combines any two portfolios on the efficient frontier of risky assets. The market portfolio m consists of the total holdings of all investors and is thus a convex combination of investors' portfolios, where convex combination means a linear combination with non-negative weights summing to one. Being a combination of efficient portfolios, the market portfolio is also efficient.

Recall from Theorem 7.5 that every efficient portfolio (except from the minimum-

	Full sample			Recent 10 years		
	Avg ret	Beta	Alpha	Avg ret	Beta	Alpha
Low 10	0.856	0.582	0.165	0.735	0.579	0.294
Decile-2	0.978	0.783	0.182	0.957	0.860	0.318
Decile-3	0.992	0.879	0.146	0.801	0.960	0.091
Decile-4	0.975	0.987	0.073	0.807	1.165	-0.048
Decile-5	1.000	1.049	0.066	0.761	1.186	-0.108
Decile-6	1.105	1.176	0.105	0.652	1.335	-0.323
Decile-7	1.130	1.264	0.084	0.884	1.432	-0.160
Decile-8	1.138	1.399	0.022	0.699	1.495	-0.389
Decile-9	0.972	1.537	-0.216	0.379	1.624	-0.800
High 10	0.379	1.667	-0.877	0.487	1.805	-0.820

Table 10.7: Statistics on portfolios sorted on past return variance.

The table shows the average monthly return and the estimates of β_i and α_i for ten portfolios sorted on the variance of past returns. Monthly observations are used. The full sample covers 1963/07–2017/08, so the recent 10 year period is 2007/09–2017/08.

variance portfolio) has a zero-covariance partner portfolio also located on the efficient frontier. Let z indicate the efficient zero-covariance partner of the market portfolio, cf. Figure 10.12. As explained in Theorem 7.5, the tangent to the efficient frontier in the point corresponding to the market portfolio intersects the vertical axis at $E[r_z]$ and, hence, the tangent has a slope of $(E[r_m] - E[r_z]) / \text{Std}[r_m]$.

We can obtain a different expression for the slope of the tangent by proceeding as in the alternative proof of property (c) in the standard CAPM of Theorem 10.1. The tangent to the efficient frontier is also tangent to the curve traced out by combinations of the market portfolio and any specific asset i and the slope of that tangent in the point corresponding to the market portfolio is

$$\frac{E[r_i] - E[r_m]}{\text{Cov}[r_i, r_m] - \text{Var}[r_m]} \text{Std}[r_m].$$

Equating the two slopes we get

$$\frac{E[r_i] - E[r_m]}{\text{Cov}[r_i, r_m] - \text{Var}[r_m]} \text{Std}[r_m] = \frac{E[r_m] - E[r_z]}{\text{Std}[r_m]}$$

from which

$$E[r_i] = E[r_z] + \beta_i (E[r_m] - E[r_z]) \quad (10.32)$$

follows. This is the **zero-beta CAPM** originally developed by Black (1972). Here $E[r_z]$ is not directly observable, but can be estimated from a time series of returns if it is stable over time. Obviously the zero-beta CAPM also implies a linear relation between market betas and expected returns, but the line has a different intercept and slope than the usual SML. In particular, the line is flatter—and thus fits the data better—if $E[r_z] > r_f$.

In closely related work, Frazzini and Pedersen (2014) build a model in which some investors are not allowed to use leverage. If they are relatively risk tolerant and thus want a portfolio with a relatively high systematic risk, they cannot just buy the market portfolio and lever it up. Instead, they have to overweight high-beta assets relative to their market weights and underweight low-beta assets. In equilibrium, the higher demand for high-beta assets causes them to have higher prices and thus lower expected returns than in

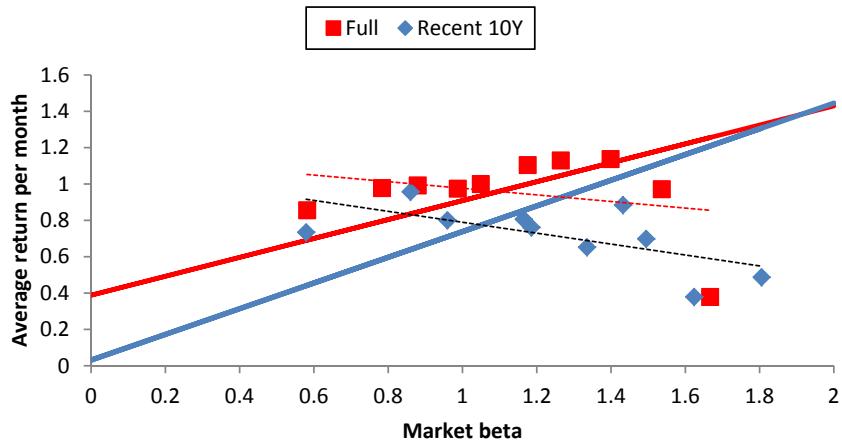


Figure 10.11: CAPM-test: variance-sorted portfolios.

The figure shows the relation between market-betas and average returns. The red squares show the combination of the beta estimate and the average return on ten variance-sorted portfolios as explained in the text. The solid red line shows the relation between beta and expected return suggested by the CAPM using the average observed riskfree rate and excess market return. Monthly observations from July 1963 to August 2017 are used. The blue diamonds and the blue solid line show the same relations but using only data from September 2007 to August 2017.

a corresponding unconstrained economy. Conversely, low-beta assets have higher expected returns. Hence, the model leads to a flatter security market line than the standard CAPM.

10.3.2 The liquidity-adjusted CAPM

Liquidity refers to the ease of trading an asset without significantly affecting its price. Low liquidity is referred to as illiquidity. Illiquidity is costly: A seller must accept a discount from the fair market value to obtain a quick sale. A buyer must pay a higher price to quickly establish a long position in an illiquid security.

The liquidity of an asset or the market for the asset depends on the direct transaction costs that have to be paid when buying or selling the asset. These costs include fees and commissions to brokers, accountants, banks, etc., as well as any tax payments caused by the trade. Even if the direct transaction costs may be small, the market may still be illiquid because of search frictions and asymmetric information. For some, in particular unusual or non-standardized, assets you have to search for a counterparty to take the opposite position of the trade you have in mind, and that search may take time and involve search costs. Maybe potential counterparties think that you know more about the true value of the asset than they do, and are therefore reluctant to trade. Finally, many potential investors may lack funding to implement the trade. Many investors rely on leverage and if credit is suddenly restricted, they are unable to implement the desired trades.

Some assets are more liquid than others. For example, exchange-traded stocks in large companies that are well-covered by analysts and media tend to be very liquid. You can almost always buy or sell a relatively low number of such stocks without a noticeable impact on the price of the stock. Most government bonds are also very liquid. In contrast, real estate is less liquid. To sell a certain building or a lot of land you may need to search for a buyer interested in that particular location, size, and quality, and since the number

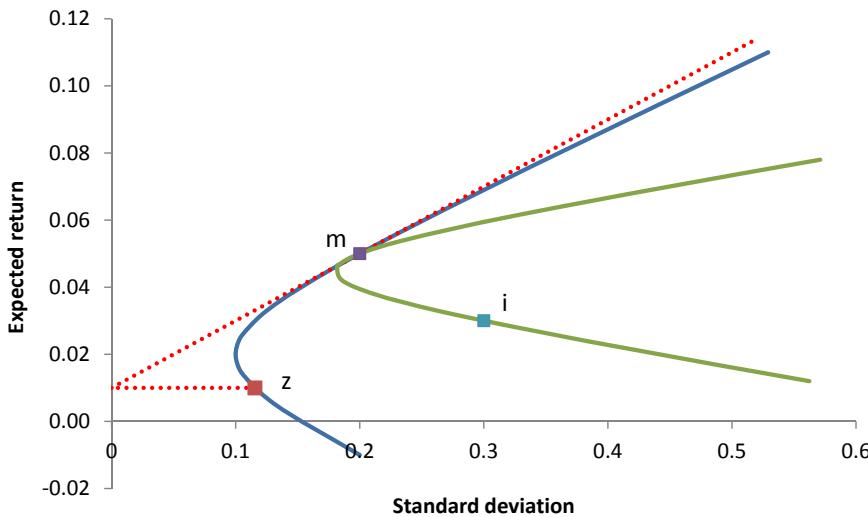


Figure 10.12: Proof of the zero-beta CAPM.

The market portfolio (purple square) has an expected return of 5% and a standard deviation of 20%, and asset i (blue square) has an expected return of 3% and a standard deviation of 30%. The correlation between the market portfolio and asset i is 0.2. The efficient frontier of risky assets (blue curve) is drawn assuming the minimum-variance portfolio has an expected return of 2% and a standard deviation of 10%. The zero-beta portfolio (red square) has an expected return of 1% and a standard deviation of 11.5%.

of potential buyers might be low, the search can take time and you may need to reduce the price below what you find reasonable. Among financial assets, private equity and municipal bonds are among the least liquid assets.

The liquidity of a given asset may vary over time. During the financial crisis that hit markets in 2008, assets like commercial papers and mortgage-backed securities that were previously considered very liquid suddenly turned illiquid.

How does illiquidity and uncertainty about future illiquidity affect the price or return of an asset? In standard models like the basic CAPM, individuals are assumed to be able to trade all financial assets at any point in time and without paying any transaction costs. For some financial assets transaction costs are so small that they can safely be neglected, but for other assets transaction costs are economically significant. Investors should care about returns net of transaction costs, and cross-sectional differences in liquidity will therefore induce cross-sectional differences in expected returns.

The asset pricing implications of liquidity risk is explained in the relatively simple liquidity-adjusted CAPM of Acharya and Pedersen (2005). They set up a model with overlapping generations of investors who live for one period. Each investor is born with a given endowment, purchases a portfolio of assets immediately, then liquidates the portfolio one period later, consumes the dividends and the proceeds from the sale, and leaves the economy. Investors are assumed to have CARA utilities (see Section 7.5.3). The financial market offers a perfectly liquid riskfree asset with a positive riskfree rate. The sale of a risky asset implies a transaction cost. Short-selling of the risky assets is prohibited. Under certain assumptions on dividends and transaction costs, the authors show that the expected periodic rate of return on any risky asset i (computed without adjusting for

transaction costs) is

$$\begin{aligned} E[r_i] - r_f &= E[c_i] + \lambda \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m - c_m]} + \lambda \frac{\text{Cov}[c_i, c_m]}{\text{Var}[r_m - c_m]} \\ &\quad - \lambda \frac{\text{Cov}[r_i, c_m]}{\text{Var}[r_m - c_m]} - \lambda \frac{\text{Cov}[c_i, r_m]}{\text{Var}[r_m - c_m]}, \end{aligned} \tag{10.33}$$

where $\lambda = E[r_m - c_m - r_f] > 0$. Here, r_m is the market return (without transaction costs), c_i is the transaction cost on asset i as a fraction of the price, and c_m is a weighted average of the relative transaction costs of the individual assets and thus a measure of the market illiquidity. Note that c_i and c_m refer to the costs of trading at the end of the period, and these costs are not known at the beginning of the period because of liquidity risk.

The excess expected return consists of the expected transaction cost and four terms compensating for risk. Each term is the product of a specific “beta” and the common market risk premium λ . The first beta corresponds to the usual CAPM market beta capturing the covariance of the asset’s return with the market return. The three new liquidity-related beta-terms have intuitive signs. An asset that tends to be very liquid (low c_i) when the general market is illiquid (high c_m) is very attractive, and therefore will have a high price and a low expected return. An asset that tends to provide high returns when the market is very illiquid will likewise have low expected returns. Finally, an asset that tends to be illiquid only when the market return is high is also relatively attractive and will thus offer low expected returns.

In an empirical study based on U.S. stocks, Acharya and Pedersen show that the liquidity-adjusted CAPM dominates standard CAPM empirically. They estimate that the expected or average transaction cost can explain around 3.5% of the expected return, whereas the three liquidity risk terms add up to around 1.1%. These estimates indicate that liquidity and liquidity risk are important elements in explaining the cross-sectional differences in expected returns.

Amihud, Mendelson, and Pedersen (2005) survey both the theoretical and the empirical literature on the impact of illiquidity on asset prices. More recent studies include Brunnermeier and Pedersen (2009), Hasbrouck (2009), and Brennan, Chordia, Subrahmanyam, and Tong (2012). Ang (2014, Ch. 13) has an excellent and extensive discussion of liquidity, the illiquidity risk premiums, and the role of illiquid assets in portfolios.

10.3.3 A multi-period version of the CAPM

As explained in Section 10.1, the CAPM was originally derived in a one-period setting. However, empirical tests of the CAPM are using returns from multiple periods. Most of these tests are implicitly assuming that the CAPM repeats itself period by period, that assets have constant expected returns and betas, and that the market risk premium is constant. Only under such assumptions can we use the average return as an approximation of the expected return, the sample beta as an approximation of the future beta, and the average market risk premium as an approximation of the future market risk premium.

In practice, all of these quantities vary over time to some extent. Maybe the CAPM holds for each period in the sense that

$$E_t [r_{i,t+1} - r_{f,t+1}] = \beta_{i,t} E_t [r_{m,t+1} - r_{f,t+1}], \quad i = 1, 2, \dots, N. \tag{10.34}$$

Here $r_{i,t+1}$ is the rate of return on asset i from time t to time $t + 1$, and similarly for the

State	Prob	Market risk premium	β_A	β_B
Good	0.5	6%	-1	3
Bad	0.5	10%	2	-1
Average		8%	0.5	1

Table 10.8: Time-varying betas and risk premium.

The table refers to a simple example where both the market beta of assets and the market risk premium vary with the state of the economy. See the explanations in the main text.

riskfree rate $r_{f,t+1}$ and the market return $r_{m,t+1}$. Furthermore, $\beta_{i,t}$ is given by

$$\beta_{i,t} = \frac{\text{Cov}_t [r_{i,t+1}, r_{m,t+1}]}{\text{Var}_t [r_{m,t+1}]}.$$

The t -subscripts indicate that the expectation, variance, and covariance are computed conditional on the information available at time t , i.e., at the beginning of the period.

Ignoring the variations in betas and the market risk premium may lead to perverse conclusions. Here is an example. Consider two assets labeled A and B. Their market betas and the market risk premium vary with the state of the economy. For simplicity, assume that the state of the economy is either ‘good’ or ‘bad’ and that the two states are equally likely. Suppose the possible values are as illustrated in Table 10.8. Based on the averages across states (as used in standard tests), the risk premium of A would be $0.5 \times 8\% = 4\%$, whereas the risk premium of B would be $1 \times 8\% = 8\%$. But, in fact, the average risk premium of A is

$$0.5 \times (-1) \times 6\% + 0.5 \times 2 \times 10\% = 7\%,$$

and the average risk premium of B is

$$0.5 \times 3 \times 6\% + 0.5 \times (-1) \times 10\% = 4\%.$$

The average risk premium is thus inversely related to the average beta, which seems to disqualify the CAPM. But, by construction, the CAPM is really valid in each period. Note that asset A has a beta which is positively covarying with the market risk premium, whereas the beta of B is negatively covarying with the market premium. Tests based on the overall averages underestimate the expected returns on assets with betas positively covarying with the market risk premium (like asset A) and overestimates the expected returns on assets with betas negatively covarying with the market premium (like asset B).

By ignoring the time-variation, the relation between betas and expected returns can therefore be misspecified. In standard tests of the CAPM, factors that proxy for the variations in betas and the market risk premium may prove significant when included as X_i in the cross-sectional regression (10.30).

Time variation in the market risk premium and market betas can partially explain the value effect. Petkova and Zhang (2005) show that value stocks have higher betas than growth stocks in bad times where the market risk premium is high, whereas value stocks have lower betas than growth stocks in good times where the market risk premium is low. In line with the above example, tests ignoring time variations underestimate the expected returns on value stocks and overestimate the expected returns on growth stocks. However, this mechanism is quantitatively not big enough to fully explain the observed

value premium in long data samples.

The multi-period CAPM (10.34) requires that investors at the beginning of each period agree on the efficient mean-variance frontier over the next period and that all investors invest in some combination of the tangency portfolio and the riskfree asset. Then again the tangency portfolio has to be identical to the market portfolio of risky assets, so that any investor at any point in time should simply hold a mix of the market portfolio and the riskfree asset. But, as explained in Section 8.3, investors with a longer-than-one-period investment horizon will generally *not* pick such a portfolio combination. They engage in intertemporal hedging to obtain protection against bad future states. For example, to hedge against low future interest rates—and thus low returns in subsequent periods—they will invest more in long-term bonds than what seems optimal in the short run. Why? Because should interest rates fall, the investors will make a profit on the bonds and thus enter the low-return scenario with a relatively large wealth.

Following this line of thinking, you should expect a higher demand for assets that allow investors to protect themselves against, for example, low future interest rates, low future risk premiums, low future labor income growth, or high unemployment risk. The equilibrium prices of such assets will be relatively high, and thus their expected returns relatively low. On the other hand, assets that are positively correlated with interest rates, risk premiums, and labor income growth are comparably unattractive, will have low equilibrium prices, and thus high expected returns. These ideas were formalized by Merton (1973a) in his Intertemporal CAPM. In standard cross-sectional CAPM regressions like (10.30), factors that covary with interest rates, risk premiums, labor income growth, etc., may therefore show up significant.

10.4 The Consumption-based CAPM

Investments move consumption opportunities across time and across states of the world. Individual investors ultimately care about the consumption they get out of their investments. Hence, any investor should value an asset by how much it contributes to her consumption at different points in time and in different states of the world. Investors would appreciate an extra dollar of returns more in “bad states” with low consumption than in “good states” with high consumption. The prices—and thus the expected returns—of financial assets will depend on how their returns covary with (aggregate) consumption. Rubinstein (1976), Lucas (1978), and Breeden (1979) derived the Consumption-based CAPM based on these insights.

10.4.1 Risk premium linked to marginal utility

We consider a one-period economy and refer to the beginning of the period as time 0 and the end of the period as time 1. Individuals living in this economy are concerned with their consumption both at time 0 and time 1. Each individual wants to consume as much as possible, but faces a budget constraint. Suppose that the individual has an initial wealth or endowment of e_0 .

At time 0, the individual invests part of her wealth in a portfolio of N financial assets. Let P_{i0} denote the price per unit of asset i , and let x_i denote the number of units of asset i purchased by the individual. The total price of the portfolio is then $\sum_{i=1}^N x_i P_{i0}$ and the individual consumes her wealth less her investments, that is, $c_0 = e_0 - \sum_{i=1}^N x_i P_{i0}$.

She keeps her portfolio until time 1 where she cashes in the dividends from the assets and sells the entire portfolio. This gives her a total income of $\sum_{i=1}^N x_i (D_i + P_i)$, where D_i is the dividend paid per unit of asset i and P_i is the ex-dividend unit price of asset i at

time 1. In addition, the individual might receive some income e from non-financial sources, e.g., labor income. If the individual does not care about future periods, she consumes the total income, i.e., her consumption at time 1 is $c = e + \sum_{i=1}^N x_i(D_i + P_i)$.

When making the portfolio decision at time 0, the individual does typically not know the assets' dividends D_i nor their time 1 prices P_i , just as the non-financial income e may be unknown. In other words, D_i , P_i , and e are random variables. Consequently, the time 1 consumption c is also a random variable.

We assume that the individual's utility is determined by $u(c_0) + e^{-\delta} E[u(c)]$, where u is an increasing and concave utility function as introduced in Section 7.5, and where δ is the subjective time preference rate of the individual. Note that the higher the δ , the lower the weight on the utility of time 1 consumption. Individuals are generally impatient in the sense that they prefer consuming a certain amount of goods today rather than at some future point in time. The parameter δ measures the degree of impatience. The decision problem that the individual faces at time 0 can be summarized as follows:

$$\begin{aligned} & \max_{x_1, \dots, x_N} \left\{ u(c_0) + e^{-\delta} E[u(c)] \right\} \\ \text{s.t. } & c_0 = e_0 - \sum_{i=1}^N x_i P_{i0}, \\ & c = e + \sum_{i=1}^N x_i (D_i + P_i), \end{aligned} \tag{10.35}$$

where the latter constraint determines the time 1 consumption for every possible realization of e , D_i , and P_i . The next theorem characterizes the solution.

Theorem 10.2

Let (c_0, c) denote the consumption plan corresponding to the optimal portfolio decision. Then, for any asset i ,

$$P_{i0} = E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} (D_i + P_i) \right] \tag{10.36}$$

and

$$E[r_i] - r_f = - \frac{\text{Cov}[r_i, u'(c)]}{E[u'(c)]} = - \frac{\text{Cov}\left[r_i, \frac{u'(c)}{u'(c_0)}\right]}{E\left[\frac{u'(c)}{u'(c_0)}\right]}, \tag{10.37}$$

where $r_i = \frac{P_i + D_i}{P_{i0}} - 1$ is the rate of return on asset i and r_f is the riskfree rate of return.

Proof

We substitute the constraints into the objective function and find the optimal portfolio by differentiating with respect to each x_i and equating the derivative by zero. The second-order condition for a maximum holds due to the concavity of the utility function. Note that

$$\frac{\partial c_0}{\partial x_i} = -P_{i0}, \quad \frac{\partial c}{\partial x_i} = D_i + P_i.$$

Hence, an application of the chain rule implies that the first-order condition with respect to x_i is

$$0 = u'(c_0) \frac{\partial c_0}{\partial x_i} + e^{-\delta} E \left[u'(c) \frac{\partial c}{\partial x_i} \right] = -P_{i0} u'(c_0) + e^{-\delta} E \left[u'(c) (D_i + P_i) \right],$$

which leads to (10.36).

We can re-express (10.36) in terms of the rate of return r_i as

$$E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} (1 + r_i) \right] = 1.$$

Recall from (3.38) that $E[X_1 X_2] = E[X_1] E[X_2] + \text{Cov}[X_1, X_2]$ for any two random variables X_1 and X_2 . Applying this result, the above expression can be rewritten as

$$(1 + E[r_i]) E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right] + \text{Cov} \left[r_i, \frac{e^{-\delta} u'(c)}{u'(c_0)} \right] = 1,$$

and isolating the expected return yields

$$E[r_i] = \frac{1}{E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right]} - \frac{\text{Cov} \left[r_i, \frac{e^{-\delta} u'(c)}{u'(c_0)} \right]}{E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right]} - 1.$$

This equation holds for all assets. In particular, it holds for the riskfree asset for which the covariance on the right-hand side is obviously zero,

$$r_f = \frac{1}{E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right]} - 1.$$

Combining the two preceding equations, we find that

$$E[r_i] = r_f - \frac{\text{Cov} \left[r_i, \frac{e^{-\delta} u'(c)}{u'(c_0)} \right]}{E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right]}.$$

Since $u'(c_0)$ is known at the beginning of the period, it works as a constant in both the covariance and the expectation, which implies that

$$\frac{\text{Cov} \left[r_i, \frac{e^{-\delta} u'(c)}{u'(c_0)} \right]}{E \left[\frac{e^{-\delta} u'(c)}{u'(c_0)} \right]} = \frac{\frac{e^{-\delta}}{u'(c_0)} \text{Cov} [r_i, u'(c)]}{\frac{e^{-\delta}}{u'(c_0)} E [u'(c)]} = \frac{\text{Cov} [r_i, u'(c)]}{E [u'(c)]}.$$

This shows the last equality in (10.37).

The relation (10.36) is a natural optimality condition. Fix a given portfolio x_1, \dots, x_N and the corresponding consumption plan (c_0, c) . Suppose you then invest in a small number ε_i extra units of asset i . This leads to a reduction in current consumption of $\varepsilon_i P_{i0}$, which is the costs of the extra investment. The change in utility of current consumption is³

$$u(c_0 - \varepsilon_i P_{i0}) - u(c_0) \approx -\varepsilon_i P_{i0} u'(c_0).$$

On the other hand, the extra investment allows an extra consumption of $\varepsilon_i(D_i + P_i)$ at time 1, which increases utility at time 1 by

$$u(c + \varepsilon_i(D_i + P_i)) - u(c) \approx \varepsilon_i(D_i + P_i)u'(c),$$

so at time 0 the increase in expected utility is approximately $\varepsilon_i E[e^{-\delta}(D_i + P_i)u'(c)]$. By definition, for the optimal portfolio there is no gain from investing slightly more or less in any asset, so the increase in expected utility from the extra consumption at time 1 has to exactly offset by the decrease in utility due to the reduced consumption at time 0. This requires

$$\varepsilon_i P_{i0} u'(c_0) = \varepsilon_i E \left[e^{-\delta}(D_i + P_i)u'(c) \right],$$

which is equivalent to (10.36).

Note that (10.37) resembles the CAPM equation (10.1), but according to (10.37) the relevant risk measure for an individual asset is the covariance of its return with the marginal utility of consumption. Risk-averse investors find assets with a positive covariance with marginal utility of consumption attractive. These assets tend to give high returns when the individual needs extra funds for consumption the most, i.e., when the individual's consumption is low and, hence, the marginal utility of extra consumption is large. Hence, such assets are in high demand and are traded at high prices, implying that the expected returns are low. In fact, the expected return is below the riskfree rate, but investors are willing to accept this since the asset provides insurance against low consumption. Conversely, an asset with a negative covariance with marginal utility tends to produce high returns when consumption is already high, i.e., marginal utility is low. Such an asset is traded at a low price, which corresponds to a high expected return. In a sense the asset can only attract investors by offering a high expected return.

The link between expected returns and marginal utility of consumption holds for any individual in the economy, at least under the stated assumptions. In order to apply such a link we need information about the consumption and utility function of the individual. But this information is generally unknown. Frequently, economists work with the concept of a representative individual. The idea is that prices in a hypothetical economy populated only by the representative individual are identical to the prices in the true economy with multiple individuals. The consumption of the representative individual has to equal the aggregate consumption of individuals in the true economy. Data on aggregate consumption are regularly published by statistical authorities.

To apply the relation between expected returns and marginal utility, we further have to know the utility function of the representative individual. Often the representative individual is simply assumed to have a utility function which seems reasonable for typical individuals in the true economy, for example a CRRA utility function with some appropriate level of the relative risk aversion coefficient. However, this assumption is not as innocuous as it may seem. Even if you think that all individual investors have CRRA utility, the relative risk aversion of the representative investor might not be constant. If

³The approximations in the following equations are precise in the limit $\varepsilon_i \rightarrow 0$.

we think of the representative individual as the average investor, it should be a weighted average with weights representing the relative wealth of the investors. After all, a poor individual invests very little in financial assets and thus have little influence on asset prices. If the stock market goes up, the investors with the largest portfolio weights in stocks benefit more than others. These are the most risk tolerant—the least risk averse—among investors. Therefore, the relative wealth share of the low risk aversion investors is bigger in good times than in bad times. Hence, the weighted average risk aversion across investors and thus the relative risk aversion of the representative individual are lower in good times than in bad times, i.e., they vary counter cyclically.⁴

10.4.2 Risk premium linked to consumption growth

Above we derived a relation between expected returns and the marginal utility of consumption. The term Consumption-based CAPM is typically used for a relation of the form

$$\mathbb{E}[r_i] = r_f + \eta \text{Cov}[r_i, g], \quad i = 1, 2, \dots, N. \quad (10.38)$$

Here $g = \frac{c}{c_0} - 1$ is the (state-dependent) relative growth rate of consumption over the period, and $\eta > 0$ is a risk premium associated with consumption growth. The appropriate risk measure of an individual asset is thus its covariance with consumption growth. As explained above, an asset that tends to provide high returns when consumption (and thus consumption growth) turns out to be high is not attractive to investors unless it offers a large expected return.

Given that (10.38) holds for all individual assets, the same relation must hold for all portfolios. In particular, it must hold for the so-called consumption-mimicking portfolio. This portfolio is (i) minimizing the variance of the difference between the portfolio return and the consumption growth and (ii) maximizing the absolute correlation with consumption growth. See, e.g., Munk (2013, Sec. 8.2.4) for details. From (10.38) it follows that the return r^c on the consumption-mimicking portfolio satisfies

$$\mathbb{E}[r^c] = r_f + \eta \text{Cov}[r^c, g].$$

It can be shown that

$$\text{Cov}[r^c, g] = \text{Var}[r^c],$$

so that

$$\eta = \frac{\mathbb{E}[r^c] - r_f}{\text{Cov}[r^c, g]} = \frac{\mathbb{E}[r^c] - r_f}{\text{Var}[r^c]}.$$

Substituting this back into (10.38), we find

$$\mathbb{E}[r_i] = r_f + \frac{\text{Cov}[r_i, g]}{\text{Var}[r^c]} (\mathbb{E}[r^c] - r_f).$$

Furthermore, it can be shown that

$$\text{Cov}[r_i, r^c] = \text{Cov}[r_i, g]$$

for any risky asset i . Consequently,

$$\mathbb{E}[r_i] = r_f + \frac{\text{Cov}[r_i, r^c]}{\text{Var}[r^c]} (\mathbb{E}[r^c] - r_f) = r_f + \beta_i^c (\mathbb{E}[r^c] - r_f), \quad (10.39)$$

⁴The argument is formalized by Chan and Kogan (2002).

where

$$\beta_i^c = \frac{\text{Cov}[r_i, r^c]}{\text{Var}[r^c]} \quad (10.40)$$

is referred to as the consumption beta of asset i . Equation (10.39) is exactly as the classical CAPM but with the return on a consumption-mimicking portfolio instead of the return on the market portfolio. An asset's consumption beta is the appropriate risk measure because each asset is evaluated by its contribution to consumption.

But how can we derive (10.38) in the first place? We need to make further assumptions or approximations. Below we consider various cases.

Case 1: Crude approximations for a general model. A first-order Taylor approximation of $u'(c)$ around c_0 gives that

$$u'(c) \approx u'(c_0) + u''(c_0)(c - c_0)$$

and thus

$$\frac{u'(c)}{u'(c_0)} \approx \frac{u'(c_0) + u''(c_0)(c - c_0)}{u'(c_0)} = 1 - \gamma(c_0)g,$$

where $\gamma(c_0) = -c_0u''(c_0)/u'(c_0)$ is the relative risk aversion of the individual evaluated at the time 0 consumption level. Now we have

$$\text{Cov}\left[r_i, \frac{u'(c)}{u'(c_0)}\right] \approx \text{Cov}[r_i, 1 - \gamma(c_0)g] = -\gamma(c_0) \text{Cov}[r_i, g]$$

and

$$\mathbb{E}\left[\frac{u'(c)}{u'(c_0)}\right] \approx 1 - \gamma(c_0) \mathbb{E}[g].$$

From (10.37), the expected excess return on a risky asset i is therefore

$$\mathbb{E}[r_i] - r_f \approx \frac{\gamma(c_0)}{1 - \gamma(c_0) \mathbb{E}[g]} \text{Cov}[r_i, g] \approx \gamma(c_0) \text{Cov}[r_i, g],$$

which is—approximately—of the desired form (10.38) with $\eta = \gamma(c_0)$, the relative risk aversion. The second approximation above assumes $\mathbb{E}[g] \approx 0$.

Case 2: The CRRA-lognormal model. Under the assumption that the individual has constant relative risk aversion and that the individual's consumption growth is lognormally distributed, the following result can be shown (see, e.g., Munk (2013, Thm. 8.1)):

Theorem 10.3

In a one-period setting with a riskfree asset, consider an individual with time preference rate δ and constant relative risk aversion γ . If the consumption growth of the individual is lognormally distributed,

$$\ln(1 + g) \equiv \ln\left(\frac{c}{c_0}\right) \sim N(\bar{g}, \sigma_g^2),$$

then the continuously compounded riskfree rate is

$$r_f^c \equiv \ln(1 + r_f) = \delta + \gamma \bar{g} - \frac{1}{2} \gamma^2 \sigma_g^2, \quad (10.41)$$

and the excess rate of return on any risky asset i is

$$\begin{aligned} E[r_i] - r_f^f &= -\text{Std}[r_i] \sqrt{e^{\gamma^2 \sigma_g^2} - 1} \text{Corr}[r_i, (c/c_0)^{-\gamma}] \\ &\approx \gamma \text{Cov}[r_i, g]. \end{aligned} \quad (10.42)$$

Ignoring the approximation, Eq. (10.42) is of the desired form (10.38) again with the risk premium identified as the relative risk aversion. The approximate expression is based on two approximations. First,

$$\sqrt{e^{\gamma^2 \sigma_g^2} - 1} \approx \gamma \sigma_g,$$

which stems from the approximation $e^x \approx 1 + x$ for $x \approx 0$. Secondly, a first-order Taylor approximation of the function $f(x) = x^{-\gamma}$ around 1 gives $f(x) \approx f(1) + f'(1)(x - 1) = 1 - \gamma(x - 1)$. With $x = c/c_0$ we get

$$\left(\frac{c}{c_0}\right)^{-\gamma} \approx 1 - \gamma \left(\frac{c}{c_0} - 1\right) = 1 - \gamma g$$

and, consequently,

$$\begin{aligned} \text{Corr}\left[r_i, \left(\frac{c}{c_0}\right)^{-\gamma}\right] &\approx \text{Corr}[r_i, 1 - \gamma g] = \frac{\text{Cov}[r_i, 1 - \gamma g]}{\text{Std}[r_i] \text{Std}[1 - \gamma g]} \\ &= \frac{-\gamma \text{Cov}[r_i, g]}{\text{Std}[r_i] \gamma \text{Std}[g]} = -\text{Corr}[r_i, g]. \end{aligned}$$

Case 3: Normally distributed returns and consumption. Assume c and r_i are jointly *normally* distributed. We need the following result called Stein's Lemma (for a proof see, e.g., Munk (2013, Lemma 8.1)):

If x and y are jointly normally distributed random variables and $g : \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable function with $E[|g'(y)|] < \infty$, then $\text{Cov}[x, g(y)] = E[g'(y)] \text{Cov}[x, y]$.

From Stein's Lemma, we get

$$\text{Cov}[r_i, u'(c)] = E[u''(c)] \text{Cov}[r_i, c].$$

From the definition of the consumption growth rate $g = \frac{c}{c_0} - 1$, we have

$$\text{Cov}[r_i, c] = c_0 \text{Cov}\left[r_i, \frac{c}{c_0}\right] = c_0 \text{Cov}[r_i, g].$$

Substituting the above into (10.37), we find

$$E[r_i] = r_f - \frac{\text{Cov}[r_i, u'(c)]}{E[u'(c)]} = r_f - \frac{c_0 E[u''(c)]}{E[u'(c)]} \text{Cov}[r_i, g], \quad (10.43)$$

which is of the form (10.38). Note that we need to have $E[|u''(c)|] < \infty$ in order to apply

Stein's Lemma and this is not a innocuous assumption. And, of course, the normality assumption is not completely realistic.

Case 4: Quadratic utility. If we assume quadratic utility

$$u(c) = -\frac{1}{2}(\bar{c} - c)^2 \quad \Rightarrow \quad u'(c) = \bar{c} - c,$$

we have

$$\mathbb{E}[u'(c)] = \bar{c} - \mathbb{E}[c], \quad \text{Cov}[r_i, u'(c)] = -\text{Cov}[r_i, c].$$

Therefore the expected return on a risky asset in (10.37) can be rewritten as

$$\mathbb{E}[r_i] = r_f + \frac{\text{Cov}[r_i, c]}{\bar{c} - \mathbb{E}[c]} = r_f + \frac{c_0}{\bar{c} - \mathbb{E}[c]} \text{Cov}[r_i, g] \quad (10.44)$$

which has the form (10.38) we were looking for. But of course quadratic utility is unreasonable as discussed already in Section 7.5.

The traditional CAPM equation for expected returns on individual stocks can be derived from the CCAPM equation (10.38). If we think of an average or representative agent, this agent must have a zero position in the riskfree asset, i.e., the representative agent cannot borrow or lend money. Why? The riskfree asset is in zero net supply, so if some investor is borrowing money, another investor must be lending the money. The average investor cannot borrow or lend. Consequently, the average investor must invest all her money in risky assets. Since the average investor cannot over- or under-weigh assets relative to the market weights, the average investor has to hold the market portfolio of risky assets. If we consider the decision problem (10.35) for the average investor, the amount invested $e_0 - c_0$ at the beginning of the period will grow to $(e_0 - c_0)(1 + r_m)$, where r_m is again the rate of return on the market portfolio. If we assume zero income, $e = 0$, end-of-period consumption is $c = (e_0 - c_0)(1 + r_m)$. The growth rate of consumption is thus

$$g = \frac{e_0 - c_0}{c_0}(1 + r_m) - 1,$$

so that the covariance in (10.38) becomes

$$\text{Cov}[r_i, g] = \text{Cov}\left[r_i, \frac{e_0 - c_0}{c_0}(1 + r_m) - 1\right] = \frac{e_0 - c_0}{c_0} \text{Cov}[r_i, r_m],$$

by using the covariance rules from Section 3.4. Hence,

$$\mathbb{E}[r_i] = r_f + \eta \frac{e_0 - c_0}{c_0} \text{Cov}[r_i, r_m] = r_f + \hat{\eta} \text{Cov}[r_i, r_m], \quad \hat{\eta} = \eta \frac{e_0 - c_0}{c_0}.$$

Since this holds for every asset i , it will also hold for the market portfolio itself so that

$$\mathbb{E}[r_m] = r_f + \hat{\eta} \text{Cov}[r_m, r_m] = r_f + \hat{\eta} \text{Var}[r_m] \quad \Rightarrow \quad \hat{\eta} = \frac{\mathbb{E}[r_m] - r_f}{\text{Var}[r_m]}.$$

Substituting this back into the previous equation, we get the traditional CAPM equation:

$$\mathbb{E}[r_i] = r_f + \frac{\mathbb{E}[r_m] - r_f}{\text{Var}[r_m]} \text{Cov}[r_i, r_m] = r_f + \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} (\mathbb{E}[r_m] - r_f).$$

10.4.3 The empirical performance of the Consumption-CAPM

The simple Consumption-based CAPM does not match the data well. Let us consider the aggregate-consumption version in Eq. (10.42) and ignore the approximation. Applying the relation to the stock market index, we should have that

$$E[r_m] = r_f + \gamma \text{Cov}[r_m, g] = r_f + \gamma \rho_{mg} \sigma_m \sigma_g.$$

Here r_m denotes the return on the index, σ_m is the standard deviation of the index return, ρ_{mg} the correlation between the index return and aggregate consumption growth, and γ is the relative risk aversion of the representative investor.

As discussed in Section 6.5, based on U.S. data from the second half of the 20th Century, $E[r_m] - r_f$ is around 8% per year and σ_m around 20%. Furthermore, σ_g is about 2%, and the sample correlation ρ_{mg} is around 0.2. Plugging in these numbers in the above equation, we get

$$0.08 = \gamma \times 0.2 \times 0.2 \times 0.02,$$

which requires that the relative risk aversion is 100, much higher than what seems reasonable. With $\gamma = 10$ and the above values of ρ_{mg} , σ_m , and σ_g , the equity premium $E[r_m] - r_f$ should only be around 0.8% according to the Consumption-CAPM. This is the so-called Equity Premium Puzzle identified by [Mehra and Prescott \(1985\)](#): the observed equity premium is much higher than it should be according to the Consumption-CAPM with apparently realistic estimates for the risk aversion and the covariance between the stock market and aggregate consumption.

There are several problems related to the data applied in such tests of the Consumption-CAPM. The historical equity premium of around 8% is probably significantly higher than the equity premium expected ex ante. First, there is a survivorship bias which may account for as much as 2-4%, cf. [Brown, Goetzmann, and Ross \(1995\)](#). Secondly, reductions in dividend taxation and trading costs and restrictions have favored stock holders and were probably not anticipated, and have thus led to higher-than-expected returns. Thirdly, the available aggregate consumption data may be of poor quality, cf., e.g., [Wilcox \(1992\)](#), and in particular be less volatile than actual consumption. In fact, the Consumption-CAPM performs much better if an “unfiltered” measure of aggregate consumption is used ([Kroencke 2017](#)) or if garbage growth is used instead of consumption growth ([Savov 2011](#)).

Furthermore, some statistical issues affect the conclusions of the typical tests of the Consumption-CAPM. First, the 8% equity premium is just a point estimate. As discussed earlier, expected returns cannot be estimated with great precision due to the sizeable return volatility. If the 8% represents the average of 50 annual return observations with a volatility of 20%, the 95% confidence interval of the true expected equity premium is roughly from 2.5% to 13.5%. Obviously, an equity premium of 2.5% is possible with a much lower relative risk aversion. Secondly, standard tests implicitly assume that the inputs to the Consumption-CAPM relation are constants over time, but they are most likely varying over time. As explained towards the end of Section 10.4.1, the relative risk aversion of the representative investor may very well vary counter cyclically, i.e., it is high in bad times and low in good times. If the return-consumption covariance moves in the same way, the standard tests will exaggerate the average relative risk aversion needed to explain the observed equity premium, cf. the discussion in Section 10.3.3.

As for the CAPM, there are many extensions of the Consumption-CAPM that perform much better than the basic version. Some extensions consider alternative investor preferences or alternative assumptions about the consumption dynamics or both. Other extensions explicitly model multiple consumption goods, for example distinguishing perishable

goods from durable goods or housing consumption from non-housing consumption. See Munk (2013, Ch. 9) for an overview of advanced consumption-based asset pricing models.

10.5 Exercises

Exercise 10.1. You have analyzed three stocks and have obtained the following estimates and information:

	Stock A	Stock B	Stock C
Stock price in \$	15.50	23.00	11.50
Expected return in %	9	11	13
Covariance with the market return in (%) ²	400	500	600

Furthermore, you have estimated the standard deviation on the market to be 25% and its expected return to be 12%. The riskfree rate is 6%. Assume that you invest in a portfolio with 200 shares of each stock.

- (a) Find the expected return and beta for the portfolio.
- (b) Is the portfolio correctly priced according the the CAPM? Why/why not?

Exercise 10.2. Show that when the CAPM holds, then Eq. (10.6) is satisfied, i.e., the correlation between the return on any asset i and the market return is equal to the ex ante Sharpe Ratio of asset i divided by the Sharpe Ratio of the market.

Exercise 10.3. If the CAPM holds, is the following situation possible? Explain in less than 40 words.

Portfolio	Expected return	Standard deviation	Correlation with the market portfolio
Market	8%	20%	1
A	7%	30%	0.6

Exercise 10.4. A market has two assets, A and B, with the following characteristics:

Asset	Expected return	Beta
A	8%	1.2
B	6%	0.6

If the market only has these two risky assets and a riskfree asset, and the CAPM holds, what must be the weights in the market portfolio?

Exercise 10.5. Consider a financial market in which the riskfree return over the next year is 2%. The CAPM holds in this market, and the expected return on the market portfolio over the next year is 10%. Stocks in the company XYZ offer an expected return of 18% over the next year. The correlation between the return on XYZ stocks and the return on the market portfolio is 0.6.

- (a) A given portfolio consisting of XYZ stocks and the riskfree asset has an expected return of 14% over the next year and the standard deviation of the return is 37.5%. Determine the weights of the two assets in the portfolio. What is the standard deviation of the return on XYZ stocks?
- (b) State an equation for the Security Market Line (SML) and sketch the line in a diagram. Explain briefly what the SML illustrates.
- (c) Determine the beta-value of XYZ and indicate the location of XYZ in your SML-diagram.
- (d) What is the standard deviation of the return on the market portfolio?

- (e) Sketch the Capital Market Line (CML) in a diagram under the assumption that the tangency portfolio is identical to the market portfolio. Indicate the location of XYZ in your CML-diagram. Explain briefly what the CML illustrates.

Exercise 10.6. Consider a financial market in which the CAPM holds. All returns mentioned in the following are over the next year. The riskfree return is 1%. The expected return on the market portfolio is 7% and the standard deviation of this return is 0.2 or 20%. Stocks in the company *Phantasia* offer an expected return of 10%. There is a correlation of 0.5 between the market return and the return on stocks in *Phantasia*.

- (a) What is the beta-value of the stocks of *Phantasia*? What is the standard deviation of the return of the stocks of *Phantasia*?
- (b) The return on a portfolio of a long position in the riskfree asset and a long position in the stocks of *Phantasia* has a standard deviation of 0.45 or 45%. What is the expected return on the portfolio?
- (c) Stocks in the company *Illusia* also have an expected return of 10%, but a correlation of 0.75 with the market portfolio. What is the standard deviation of the return of the stocks of *Illusia*? Explain why the market can be in equilibrium even though the stocks of *Phantasia* and *Illusia* have different standard deviations but the same expected returns.
- (d) Sketch diagrams showing the Security Market Line and the Capital Market Line. Indicate the location of the market portfolio, the stocks of *Phantasia*, and the stocks of *Illusia* in both diagrams.

Exercise 10.7. Consider a financial market in which the CAPM holds. The riskfree rate over the next year is 1%, and the market portfolio has an expected return of 6% and a standard deviation of 20%.

- (a) Stock A has an expected return of 11% and a standard deviation of 50%. What is the beta-value of stock A? What is the correlation between the return on stock A and the return on the market portfolio?
- (b) The return on stock B has a standard deviation of 80% and a correlation of 0.5 with the market return. What is the expected return of stock B? Compare with stock A.
- (c) The return on stock C has a standard deviation of 30%. You do not know its correlation with the market portfolio. What can you say about the expected return on stock C?

Exercise 10.8. (a) If the CAPM holds, is the following situation possible? Explain.

	Stock A	Stock B	Stock C
Expected return	0.068	0.092	0.128
Market beta	0.8	1.2	1.8

- (b) If the CAPM holds, is the following situation possible? Explain.

	Stock D	Stock E	Stock F
Expected return	0.101	0.112	0.140
Covariance with market portfolio	0.040	0.048	0.044

- (c) If the CAPM holds, is the following situation possible? Explain.

	Stock G	Stock H	Stock I
Expected return	0.070	0.094	0.138
Standard deviation of return	0.30	0.42	0.40

Exercise 10.9. Suppose that the CAPM holds. All returns mentioned in the following are over the next year. The riskfree rate of return is 1%. The market portfolio has a Sharpe ratio of 0.25, and the standard deviation of the market portfolio's rate of return is 20%. The rate of return on stocks in the company *Illuminati* has an expected value of 9% and a standard deviation of 40%.

- (a) What is the expected rate of return on the market portfolio?
- (b) What is the beta-value of the stocks of *Illuminati*?
- (c) What is the correlation between the rate of return on *Illuminati* stocks and the rate of return on the market portfolio?
- (d) The rate of return on a portfolio of a long position in the riskfree asset and a long position in *Illuminati* stocks has a standard deviation of 32%. What is the expected rate of return on the portfolio?
- (e) Stocks in the company *Hypothetics* also have an expected return of 9%, but a correlation of 0.4 with the market portfolio. What is the standard deviation of the return of the stocks of *Hypothetics*?
- (f) Explain why the market can be in equilibrium even though the stocks of *Hypothetics* and *Illuminati* have different standard deviations but the same expected returns.
- (g) Sketch diagrams showing the Security Market Line and the Capital Market Line. Indicate the location of the market portfolio, the stocks of *Hypothetics*, and the stocks of *Illuminati* in both diagrams.

Exercise 10.10. Go to the homepage of Professor Kenneth French and download monthly return observations for 10 industry portfolios as well as monthly observations of the riskfree rate and the excess market return. Using the most recent 60 months, calculate the average return and a beta estimate for each industry portfolio. In a diagram with beta along the horizontal axis and average return along the vertical axis, show where the 10 industry portfolios are located and draw the straight line best describing the beta-return relation. Do the industry portfolios seem to satisfy the CAPM?

CHAPTER 11

Factor models

A factor model explains the risk premium on any individual asset by the asset's exposure to a relatively low number of priced factors that are common for all assets. Differences in expected returns across assets should be due solely to the assets having different factor exposures. Moreover, in a factor model, all the return covariation across assets comes from the exposure of all returns to the common factors, and the return variation of an asset can be decomposed into a systematic component coming from the factor exposures and a non-systematic component coming from the asset-specific risk. The key challenges are to identify what the common factors should be and how large the risk premium associated with each factor is.

Obvious candidates for the factors are macroeconomic variables that are likely to affect more or less all assets, such as the GDP growth rate, oil prices, and recession indicators, and maybe adding some industry-specific factors. However, many factor models assume the factors are returns on specific portfolios mainly because returns can be observed frequently and precisely, whereas most macro variables are only published with a monthly or quarterly frequency and might even be revised later. The specific portfolios are sometimes chosen to mimic or track selected macro variables. Other popular factors are return differences motivated by observed anomalies relative to the CAPM, such as those we identified in Section 10.2. For example, the well-known Fama-French three-factor model adds to the market factor two return differences, namely the return difference between a portfolio of small stocks and a portfolio of large stocks, as well as the return difference between a portfolio of value stocks and a portfolio of growth stocks.

Section 11.1 introduces the general one-factor model framework and discusses its assumptions and consequences. Section 11.2 develops the Single-Index Model which is the most popular example of a one-factor model and can be seen as a close cousin of the Capital Asset Pricing Model discussed in Chapter 10. The assumption that a single factor is the source of the common variations across all risky assets is obviously restrictive, and the apparent empirical failure of the CAPM illustrated in the previous chapter suggests that more than one risk factor is priced. Section 11.3 introduces the general K -factor return-generating model and explores its properties. Section 11.4 explains the so-called Arbitrage Pricing Theory which leads to an equilibrium version of the multi-factor model stating that the risk premium on any individual asset can come only from its exposure to the factors. Section 11.5 presents the Fama-French three-factor model, as well as its

recent extension to five factors. Section 11.6 outlines the current state of the search for factors and discusses the robustness and practical usefulness of such factors. Finally, the implications of multi-factor models for portfolio choice are discussed in Section 11.7.

11.1 A general one-factor model

We use the same notation as in earlier chapters. In particular, N denotes the number of assets, and assets are labelled by $i = 1, 2, \dots, N$. The rate of return on asset i over a given period is still denoted by r_i .

11.1.1 Definition and basic properties

A one-factor model is the assumption that a random variable F exists so that

$$r_i = E[r_i] + \beta_i (F - E[F]) + \varepsilon_i, \quad i = 1, 2, \dots, N, \quad (11.1)$$

with the properties

- (i) $\text{Cov}[F, \varepsilon_i] = 0, \quad i = 1, 2, \dots, N,$
- (ii) $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i, j = 1, 2, \dots, N, \quad i \neq j.$

Here F is the common factor affecting all returns (except assets with $\beta_i = 0$), and ε_i is a random variable representing the asset-specific return uncertainty. By taking expectations on both sides of (11.1), we see that $E[\varepsilon_i] = 0$ by construction. The coefficient β_i is asset i 's factor beta or factor sensitivity. For an increase of ΔF in the factor, the return changes by $\beta_i \Delta F$.

Note that at this stage, the factor model is silent about what the expected return $E[r_i]$ should be, but is simply saying that unexpected returns on individual assets occur only due to an unexpected realization of the common factor or due to unexpected, return-relevant asset-specific news. Later we impose an equilibrium argument leading to a statement about what the expected asset returns should be.

We summarize some important properties of the one-factor model in a theorem:

Theorem 11.1

Suppose the one-factor model (11.1) holds for assets $i = 1, 2, \dots, N$.

(a) We have

$$\text{Cov}[r_i, F] = \beta_i \text{Var}[F], \quad (11.2)$$

$$\text{Var}[r_i] = \beta_i^2 \text{Var}[F] + \text{Var}[\varepsilon_i], \quad (11.3)$$

$$\text{Cov}[r_i, r_j] = \beta_i \beta_j \text{Var}[F]. \quad (11.4)$$

(b) The return r_p on a portfolio represented by the weight vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ satisfies

$$r_p = E[r_p] + \beta_p (F - E[F]) + \varepsilon_p, \quad (11.5)$$

where

$$\beta_p = \sum_{i=1}^N \pi_i \beta_i = \frac{\text{Cov}[r_p, F]}{\text{Var}[F]}, \quad \varepsilon_p = \sum_{i=1}^N \pi_i \varepsilon_i \quad (11.6)$$

with non-systematic portfolio risk

$$\text{Var}[\varepsilon_p] = \sum_{i=1}^N \pi_i^2 \text{Var}[\varepsilon_i]. \quad (11.7)$$

Proof

(a) The covariance of the return on asset i with the factor is

$$\begin{aligned} \text{Cov}[r_i, F] &= \text{Cov} [\text{E}[r_i] + \beta_i (F - \text{E}[F]) + \varepsilon_i, F] = \text{Cov}[\beta_i F + \varepsilon_i, F] \\ &= \beta_i \text{Cov}[F, F] + \text{Cov}[\varepsilon_i, F] = \beta_i \text{Var}[F], \end{aligned}$$

using the general properties of covariances stated in Section 3.4 as well as the above assumptions of the factor model.

Using rules from Section 3.4 and the assumption $\text{Cov}[F, \varepsilon_i] = 0$, the return variance is

$$\begin{aligned} \text{Var}[r_i] &= \text{Var} [\beta_i (F - \text{E}[F]) + \varepsilon_i] = \text{Var}[\beta_i F + \varepsilon_i] \\ &= \beta_i^2 \text{Var}[F] + \text{Var}[\varepsilon_i] + 2\beta_i \text{Cov}[F, \varepsilon_i] = \beta_i^2 \text{Var}[F] + \text{Var}[\varepsilon_i]. \end{aligned}$$

Since

$$r_i = \text{E}[r_i] + \beta_i (F - \text{E}[F]) + \varepsilon_i, \quad r_j = \text{E}[r_j] + \beta_j (F - \text{E}[F]) + \varepsilon_j$$

with $\text{Cov}[F, \varepsilon_i] = \text{Cov}[F, \varepsilon_j] = \text{Cov}[\varepsilon_i, \varepsilon_j] = 0$, we get

$$\begin{aligned} \text{Cov}[r_i, r_j] &= \text{Cov}[\beta_i (F - \text{E}[F]) + \varepsilon_i, \beta_j (F - \text{E}[F]) + \varepsilon_j] \\ &= \text{Cov}[\beta_i F + \varepsilon_i, \beta_j F + \varepsilon_j] \\ &= \text{Cov}[\beta_i F, \beta_j F] + \text{Cov}[\beta_i F, \varepsilon_j] + \text{Cov}[\varepsilon_i, \beta_j F] + \text{Cov}[\varepsilon_i, \varepsilon_j] \\ &= \beta_i \beta_j \text{Cov}[F, F] = \beta_i \beta_j \text{Var}[F], \end{aligned}$$

again using rules for covariances from Section 3.4.

(b) The portfolio return is

$$\begin{aligned} r_p &= \sum_{i=1}^N \pi_i r_i = \sum_{i=1}^N \pi_i (\text{E}[r_i] + \beta_i (F - \text{E}[F]) + \varepsilon_i) \\ &= \sum_{i=1}^N \pi_i \text{E}[r_i] + \sum_{i=1}^N \pi_i \beta_i (F - \text{E}[F]) + \sum_{i=1}^N \pi_i \varepsilon_i \\ &= \text{E} \left[\sum_{i=1}^N \pi_i r_i \right] + \left(\sum_{i=1}^N \pi_i \beta_i \right) (F - \text{E}[F]) + \sum_{i=1}^N \pi_i \varepsilon_i \\ &= \text{E}[r_p] + \beta_p (F - \text{E}[F]) + \varepsilon_p, \end{aligned}$$

which confirms (11.5). Note that

$$\beta_p = \sum_{i=1}^N \pi_i \beta_i = \sum_{i=1}^N \pi_i \frac{\text{Cov}[r_i, F]}{\text{Var}[F]} = \frac{\text{Cov} \left[\sum_{i=1}^N \pi_i r_i, F \right]}{\text{Var}[F]} = \frac{\text{Cov}[r_p, F]}{\text{Var}[F]}.$$

Finally, we get

$$\text{Var}[\varepsilon_p] = \text{Var} \left[\sum_{i=1}^N \pi_i \varepsilon_i \right] = \sum_{i=1}^N \pi_i^2 \text{Var}[\varepsilon_i]$$

using standard computational rules for variances as well as the assumption of the factor model that $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ whenever $i \neq j$.

Let us first consider the results in Part (a) of Theorem 11.1. Clearly, Eq. (11.2) implies that

$$\beta_i = \frac{\text{Cov}[r_i, F]}{\text{Var}[F]}. \quad (11.8)$$

Expressing the covariance as the product of the correlation and the standard deviations, we can rewrite the factor beta as

$$\beta_i = \frac{\text{Corr}[r_i, F] \text{Std}[r_i] \text{Std}[F]}{\text{Var}[F]} = \text{Corr}[r_i, F] \frac{\text{Std}[r_i]}{\text{Std}[F]}, \quad (11.9)$$

as we did in (10.5) for the CAPM.

The return variance in Eq. (11.3) is the sum of two terms. We refer to the first term $\beta_i^2 \text{Var}[F]$ as the **systematic risk** and to the second term $\text{Var}[\varepsilon_i]$ as the **non-systematic risk** of asset i . The non-systematic risk is also called the idiosyncratic risk or the asset-specific risk or, if the asset is a stock, the firm-specific risk. The one-factor model thus gives a decomposition of the return variance into a systematic and a non-systematic risk component. Note that a similar decomposition does not hold for standard deviations: the relation $\text{Std}[r_i] = \beta_i \text{Std}[F] + \text{Std}[\varepsilon_i]$ is *not correct*.

Eq. (11.4) shows that the covariances of all assets are fully determined by the factor variance and the factor beta's of the assets, i.e., the covariances arise only from the assets' sensitivity to the same factor.

Together, Eqs. (11.3) and (11.4) imply that the entire variance-covariance structure of all assets are determined by $2N + 1$ values, namely a factor beta and non-systematic risk for each asset, as well as the factor variance. Without imposing a factor model the variance-covariance matrix of N assets contains N variances and $N(N - 1)/2$ covariances. With $N = 10$, this involves $10 + 10 \times 9/2 = 55$ values, but with the one-factor model only $2 \times 10 + 1 = 21$ inputs are needed. With $N = 100$, the one-factor model reduces the number of inputs from $100 + 100 \times 99/2 = 5050$ to $2 \times 100 + 1 = 201$. With $N = 1000$, the reduction is from $1000 + 1000 \times 999/2 = 500500$ to $2 \times 1000 + 1 = 2001$ inputs.

If we need the full variance-covariance matrix of a large number of assets, for example to implement the mean-variance approach on individual assets, the reduction in the number of parameters to be estimated is an obvious benefit of imposing a factor structure on returns. Estimating all the pairwise covariances separately opens up for lots of estimation errors that might significantly influence the output of the mean-variance portfolio choice model. Very often the parameter estimates derived under some sensible return-generating model lead to better predictions of the future expectations, variances, and covariances of returns, see, e.g., the analysis and results of MacKinlay and Pastor (2000). Furthermore, the factor model divides the required parameters into macro quantities and firm-specific quantities, which allows the analysis to be decentralized.

Next, turn to Part (b) of Theorem 11.1. First of all, Eq. (11.5) confirms that the

same return structure holds for portfolios as for individual assets. For a large, balanced portfolio—meaning a portfolio of many assets with no portfolio weights being much larger than the rest—the non-systematic portfolio risk is small and often much smaller than the non-systematic risks of the assets in the portfolio. For example, if the portfolio is equally weighted so that $\pi_i = 1/N$ for all assets i , the non-systematic risk is

$$\text{Var}[\varepsilon_p] = \sum_{i=1}^N \frac{1}{N^2} \text{Var}[\varepsilon_i] = \frac{1}{N^2} \sum_{i=1}^N \text{Var}[\varepsilon_i] = \frac{1}{N} \overline{\text{Var}[\varepsilon_i]},$$

where $\overline{\text{Var}[\varepsilon_i]} = (\sum_{i=1}^N \text{Var}[\varepsilon_i])/N$ is the average non-systematic risk of the assets in the portfolio. Since the non-systematic return components of the assets are uncorrelated, these risks are reduced by diversification. Intuitively, the non-systematic return component will be positive for some assets and negative for other assets, so that a combination of the assets is less risky than the individual assets. The non-systematic risk of individual assets is also referred to as diversifiable risk.

If we formally let the number of assets in the portfolio grow to infinity, we see from the above equation that the non-systematic portfolio risk goes to zero:

$$\text{Var}[\varepsilon_p] \rightarrow 0 \quad \text{for } N \rightarrow \infty.$$

So if we had infinitely many assets, we could diversify away all the non-systematic risk. In contrast, the systematic risk remains. The factor beta of an equally-weighted portfolio is

$$\beta_p = \sum_{i=1}^N \frac{1}{N} \beta_i = \frac{1}{N} \sum_{i=1}^N \beta_i = \bar{\beta},$$

that is, equal to the average factor beta of the assets in the portfolio, which does not vanish just because we add more assets to the portfolio.

11.1.2 Alternative formulations of the one-factor model

We can reformulate Eq. (11.1) as

$$r_i = a_i + \beta_i F + \varepsilon_i, \tag{11.10}$$

where $a_i = E[r_i] - \beta_i E[F]$. Moreover, we can formulate the model in terms of the excess returns $r_i - r_f$. Just subtract r_f from both sides, and on the right-hand side move r_f inside the expected return which is possible since r_f is not a random variable. Then we get

$$r_i - r_f = E[r_i - r_f] + \beta_i (F - E[F]) + \varepsilon_i \tag{11.11}$$

or

$$r_i - r_f = a'_i + \beta_i F + \varepsilon_i, \tag{11.12}$$

where $a'_i = a_i - r_f = E[r_i - r_f] - \beta_i E[F]$.

The factor is the source of all joint variation in returns across stocks so good candidates are macroeconomic variables like the growth rate of either GDP, per-capita consumption, or industrial production. In order for the factor model to be of practical use, the factor must be easily observable, preferably with frequent updates. Hence, you might look for a financial variable, maybe the return on a specific portfolio. Section 11.2 presents an important example where the factor is the return on a stock market index. In fact it is

not a restriction to look for the factor only among returns. If you have some factor F in mind, which is not the return of some portfolio, you can construct a factor-mimicking portfolio and use the return on this portfolio as a factor. The factor-mimicking portfolio is the portfolio that gives the smallest variance of the difference between the factor and the portfolio value, which turns out also to be the portfolio maximizing the absolute value of the correlation between the factor value and the portfolio value.¹

11.1.3 Estimation of inputs

How do we obtain the relevant inputs to a factor model? Given a time series of observations of the factor and returns (or excess returns), we can estimate β_i and $\text{Var}[\varepsilon_i]$ from a linear regression of r_i (or $r_i - r_f$) on F . Here, we can use any of the above formulations of the factor model, for example

$$r_{it} - r_{ft} = a'_i + \beta_i F_t + \varepsilon_{it}. \quad (11.13)$$

This estimation procedure assumes that the factor structure holds throughout the sample period with constant coefficients a'_i and β_i .

11.2 The Single-Index Model

By far, the most famous one-factor model is the Single-Index Model or Market Model that was already foreseen in a footnote in [Markowitz \(1959\)](#), but was developed by Sharpe ([1963](#)). The model assumes that the factor equals the return on a market portfolio or, more specifically, a stock market index. In the expressions in the previous subsection, we can thus simply replace F by r_m , the return on the market index. The Single-Index Model is mainly applied to stocks and thus relates returns on individual stocks to the general stock market return.

11.2.1 The basics

The Single-Index Model is typically formulated in terms of excess returns as

$$r_i - r_f = \alpha_i + \beta_i (r_m - r_f) + \varepsilon_i, \quad (11.14)$$

where the assumptions are that $E[\varepsilon_i] = \text{Cov}[\varepsilon_i, r_m] = \text{Cov}[\varepsilon_i, \varepsilon_j] = 0$. In particular, the condition $E[\varepsilon_i] = 0$ implies that

$$\alpha_i = E[r_i] - r_f - \beta_i (E[r_m] - r_f), \quad (11.15)$$

which is therefore identical to the CAPM alpha introduced in [\(10.17\)](#). By substituting [\(11.15\)](#) into [\(11.14\)](#) and rearranging, we get

$$r_i = E[r_i] + \beta_i (r_m - E[r_m]) + \varepsilon_i,$$

which is on the form [\(11.1\)](#) of a one-factor model. Hence, the Single-Index Model is indeed just a specific one-factor model. Applying Theorem [11.1](#), we immediately get the following results:

¹See, e.g., [Munk \(2013, Sec. 4.6.2\)](#).

Theorem 11.2

Suppose the Single-Index Model (11.14) holds for assets $i = 1, 2, \dots, N$.

(a) We have

$$\text{Cov}[r_i, r_m] = \beta_i \text{Var}[r_m], \quad (11.16)$$

$$\text{Var}[r_i] = \beta_i^2 \text{Var}[r_m] + \text{Var}[\varepsilon_i], \quad (11.17)$$

$$\text{Cov}[r_i, r_j] = \beta_i \beta_j \text{Var}[r_m], \quad \text{for } j \neq i. \quad (11.18)$$

(b) The return r_p on a portfolio represented by the weight vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ satisfies

$$r_p - r_f = \alpha_p + \beta_p (r_m - r_f) + \varepsilon_p, \quad (11.19)$$

where

$$\alpha_p = \sum_{i=1}^N \pi_i \alpha_i, \quad \beta_p = \sum_{i=1}^N \pi_i \beta_i, \quad \varepsilon_p = \sum_{i=1}^N \pi_i \varepsilon_i \quad (11.20)$$

with non-systematic portfolio risk

$$\text{Var}[\varepsilon_p] = \sum_{i=1}^N \pi_i^2 \text{Var}[\varepsilon_i]. \quad (11.21)$$

In the Single-Index Model, we see directly from (11.14) that β_i is a measure of the sensitivity of asset i 's return to the market return: an increase in the index return by one percentage point should change the return on asset i by β_i percentage points. From (11.16), we get

$$\beta_i = \frac{\text{Cov}[r_i, r_m]}{\text{Var}[r_m]} = \text{Corr}[r_i, r_m] \frac{\text{Std}[r_i]}{\text{Std}[r_m]}, \quad (11.22)$$

in accordance with the definition (10.2) of the market-beta in the CAPM of the preceding chapter.

Eq. (11.17) decomposes the return variance of each asset into a systematic, market-related variance given by $\beta_i^2 \text{Var}[r_m]$ and a firm-specific variance given by $\text{Var}[\varepsilon_i]$. Note that the systematic variance component can be rewritten using (11.22) as

$$\beta_i^2 \text{Var}[r_m] = (\text{Corr}[r_i, r_m])^2 \frac{\text{Var}[r_i]}{\text{Var}[r_m]} \text{Var}[r_m] = (\text{Corr}[r_i, r_m])^2 \text{Var}[r_i].$$

Hence, the systematic component's share of the overall return variance is $(\text{Corr}[r_i, r_m])^2$, the squared correlation with the market return. Consequently, the firm-specific component's share of the overall return variance is $1 - (\text{Corr}[r_i, r_m])^2$.

Eq. (11.18) shows that the covariance between returns of different assets is due to the common variation with the market. Firm-specific risks are uncorrelated with the market and uncorrelated across assets so they are not affecting asset covariances. It can be shown that

$$\text{Corr}[r_i, r_j] = \text{Corr}[r_i, r_m] \text{Corr}[r_j, r_m], \quad \text{for } j \neq i,$$

i.e. the correlation between two stocks is equal to the product of the stocks' correlation with the market.

If all the assets in the market portfolio follow the Single-Index Model, then the market

portfolio itself has to satisfy (11.19), i.e. $r_m - r_f = \alpha_m + \beta_m(r_m - r_f) + \varepsilon_m$, which implies that

$$\alpha_m = 0, \quad \beta_m = 1, \quad \varepsilon_m = 0.$$

11.2.2 Estimation

We can estimate α_i , β_i , and $\text{Var}[\varepsilon_i]$ by the simple linear regression

$$r_{it} - r_{ft} = \alpha_i + \beta_i(r_{mt} - r_{ft}) + \varepsilon_{it}, \quad (11.23)$$

which is identical to the regression (10.18) used in Section 10.1.5 to estimate the CAPM market beta of an asset. We refer to that section for further information on linear regressions and how to execute them in Excel. The R^2 of the above regression is the ratio

$$R^2 = (\text{Corr}[r_i, r_m])^2 = \frac{\beta_i^2 \text{Var}[r_m]}{\text{Var}[r_i]}, \quad (11.24)$$

where the second equality is from (11.22). We see that R^2 is the fraction of the total return variance explained by the systematic risk exposure. Given the assumptions of the Single-Index Model, the regression does not only give us an estimate of β_i , but also estimates of α_i as well as of the systematic and the non-systematic risk of the asset as illustrated in the following example.

Example 11.1

Example 10.2 performed the regression (11.23) for Microsoft stocks using 60 monthly returns from January 2019 to December 2023. The regression is graphically depicted in Figure 10.4 and the commented Excel regression output can be seen in Figure 10.5.

As explained in Example 10.2, the regression leads to an α estimate of 1.3655% per month, which is large and even statistically significant. The estimate of β is 0.8223, but with a relatively wide confidence interval.

Now consider the information about the risk of Microsoft that we can deduce from the regression output in Figure 10.5. The estimate of the systematic risk can be derived from the regression output by dividing the 1233.0541 (in the SS-column; SS for sum of squares) by the 59 (in the df-column), which is the number of observations minus one. This leads to a systematic variance of 20.8992 with a unit of ‘squared percent’ or $(\%)^2$ since the returns are represented in percent and variances measure squared deviations from the mean return. Taking the square root, we get a systematic standard deviation of 4.5716% per month. Alternatively, you can compute the sample variance of the excess market return—in our case $30.9078(\%)^2$ —and multiply by the square of the beta-estimate. This leads to the same systematic variance (if you use non-rounded values in your calculations).

The number 19.6263 in the MS-column is an estimate of the non-systematic variance. Typically, the number is scaled by the ratio $(T - 2)/(T - 1)$, which in our case gives us a non-systematic variance of $19.6263 \times 58/59 = 19.2937$. Adding up the systematic variance of 20.8992 and the non-systematic variance of 19.2937, we get 40.1929, which is exactly the sample variance of Microsoft’s excess returns. The systematic variance constitutes 52.0% of the total variance, and this is identical to the R^2 of the regression.

Estimates of the α in the Single-Index Model (and other factor models) are very rarely

statistically significant, in particular when the model is applied to a single asset. This is basically because expected returns are so difficult to estimate precisely due to the large variability (or noise) in time series of returns, as we discussed in Section 3.7.2. Moreover, even if you would find a significantly positive historical estimate of α , do you believe in a positive α in the future? The apparent over-performance of a stock relative to the market in the data period may be due to unexpected positive firm-specific events that are unlikely to repeat themselves in the following years. Empirically, past α -estimates of individual stocks are poor predictors of future α -values. For these reasons, many analysts and investors are not using the historical estimate of α as a guidance for future expected returns. Instead, an α is typically estimated by a detailed *security analysis* of each asset. For a stock, this involves the very challenging task of estimating future cash flows from the issuing company to its stockholders, as well as the riskiness of these cash flows, and then discounting the cash flows back to obtain a present value of the stocks.

11.2.3 Portfolio choice with the Single-Index Model

Suppose that all asset returns can be described by the Single-Index Model and that all assets have zero alpha as in the CAPM. Further suppose that you can invest directly in the market portfolio, for example through an ETF. How should your portfolio combine the market portfolio with individual assets? It should not: if you want to maximize the Sharpe ratio of your portfolio, just invest in the market portfolio, not in individual assets.

Theorem 11.3

Suppose that the Single-Index Model holds, the market portfolio is tradeable, and all individual assets have zero alpha. Then the Sharpe ratio is maximized by investing only in the market portfolio. Individual assets thus enter the portfolio with exactly the weight they have in the market portfolio.

Of course, this result is consistent with the CAPM conclusions from Theorem 10.1 that each investor should combine the riskfree asset and the market portfolio, but the CAPM involves assumptions about investor preferences and beliefs that we do not impose in the Single-Index Model. Hence, we need a separate proof.

Proof

With zero alphas, the rate of return on each risky asset satisfies $r_i = r_f + \beta_i(r_m - r_f) + \varepsilon_i$ and the expected rate of return is $E[r_i] = r_f + \beta_i(E[r_m] - r_f)$. Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ denote a portfolio weight vector with elements summing to one. Then the expected excess return

on the portfolio is

$$\begin{aligned}
E[r_p] - r_f &= \sum_{i=1}^N \pi_i E[r_i] - r_f = \sum_{i=1}^N \pi_i (E[r_i] - r_f) \\
&= \sum_{i=1}^N \pi_i \beta_i (E[r_m] - r_f) = (E[r_m] - r_f) \sum_{i=1}^N \pi_i \beta_i \\
&= \sigma_m \text{SR}_m \sum_{i=1}^N \pi_i \beta_i = \sigma_m \text{SR}_m \beta_p = \sigma_m \text{SR}_m \rho_{pm} \frac{\sigma_p}{\sigma_m} = \text{SR}_m \rho_{pm} \sigma_p,
\end{aligned}$$

where SR_m is the Sharpe ratio of the market portfolio, ρ_{pm} is the correlation between the portfolio and the market, and σ_p and σ_m are the standard deviations of the portfolio and the market, respectively. We have used Eqs. (11.20) and (11.22) for the portfolio beta. Now the Sharpe ratio of the portfolio is

$$\text{SR}_p = \frac{E[r_p] - r_f}{\sigma_p} = \rho_{pm} \text{SR}_m. \quad (11.25)$$

Since the correlation by definition is less than or equal to one, we see that it is impossible to find a portfolio with a Sharpe ratio larger than the Sharpe ratio of the market portfolio.

Here is an alternative approach. Let us consider whether a deviation from the market portfolio can lead to a larger Sharpe ratio. From Chapter 7 we know that with a certain set of risky assets the maximal Sharpe ratio is attained by the tangency portfolio derived from those assets. The tangency portfolio weight vector is given by Eq. (7.34). Suppose you can invest only in the market portfolio and some other risky asset i . If we think of the market portfolio as the first asset and asset i as the second asset, then the vector of excess expected returns and the variance-covariance matrix are

$$\boldsymbol{\mu} - r_f \mathbf{1} = \begin{pmatrix} \mu_m - r_f \\ \mu_i - r_f \end{pmatrix} = \begin{pmatrix} \mu_m - r_f \\ \beta_i(\mu_m - r_f) \end{pmatrix}, \quad \underline{\Sigma} = \begin{pmatrix} \sigma_m^2 & \beta_i \sigma_m^2 \\ \beta_i \sigma_m^2 & \beta_i^2 \sigma_m^2 + \sigma_{\varepsilon,i}^2 \end{pmatrix}.$$

The determinant of the variance-covariance matrix is $\sigma_m^2[\beta_i^2 \sigma_m^2 + \sigma_{\varepsilon,i}^2] - \beta_i^2 \sigma_m^4 = \sigma_m^2 \sigma_{\varepsilon,i}^2$, and the inverse of the matrix is thus

$$\underline{\Sigma}^{-1} = \frac{1}{\sigma_m^2 \sigma_{\varepsilon,i}^2} \begin{pmatrix} \beta_i^2 \sigma_m^2 + \sigma_{\varepsilon,i}^2 & -\beta_i \sigma_m^2 \\ -\beta_i \sigma_m^2 & \sigma_m^2 \end{pmatrix}.$$

Consequently,

$$\begin{aligned}
\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1}) &= \frac{1}{\sigma_m^2 \sigma_{\varepsilon,i}^2} \begin{pmatrix} \beta_i^2 \sigma_m^2 + \sigma_{\varepsilon,i}^2 & -\beta_i \sigma_m^2 \\ -\beta_i \sigma_m^2 & \sigma_m^2 \end{pmatrix} \begin{pmatrix} \mu_m - r_f \\ \beta_i(\mu_m - r_f) \end{pmatrix} \\
&= \frac{1}{\sigma_m^2 \sigma_{\varepsilon,i}^2} \begin{pmatrix} \sigma_{\varepsilon,i}^2 (\mu_m - r_f) \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\mu_m - r_f}{\sigma_m^2} \\ 0 \end{pmatrix}.
\end{aligned}$$

Scaling so that the weights add up to one, we see that the tangency portfolio consists of the market portfolio only and has a zero position on asset i .

Asset	Weight	Return
Underpriced stock	1	$\alpha_i + r_f + \beta_i(r_m - r_f) + \varepsilon_i$
Market portfolio	$-\beta_i$	$-\beta_i r_m$
Riskfree asset	β_i	$\beta_i r_f$
	1	$\alpha_i + r_f + \varepsilon_i$

Table 11.1: A market-neutral strategy for positive alpha.

The table shows how you can combine an individual asset having a positive alpha with the market portfolio and the riskfree asset to obtain a position with a positive alpha and no systematic risk.

If we invest a weight w_i in an asset i having zero alpha and a weight $1 - w_i$ in the market portfolio, the excess rate of return is

$$\begin{aligned} w_i r_i + (1 - w_i) r_m - r_f &= w_i [r_f + \beta_i(r_m - r_f) + \varepsilon_i] + (1 - w_i)r_m - r_f \\ &= (1 + w_i[\beta_i - 1])(r_m - r_f) + w_i \varepsilon_i. \end{aligned}$$

Hence, both the expected excess return and the systematic component of the standard deviation (i.e. the square root of the systematic risk/variance) are scaled by the same factor $1 + w_i[\beta_i - 1]$. As the deviation from the market portfolio also increases the non-systematic risk, the overall standard deviation increases by more than the expected excess return so that the Sharpe ratio decreases.

Next, suppose that you believe in the Single-Index Model but, in spite of all warnings, you are confident that you have identified an asset with a non-zero α . How can you profit from it?

If you just invest in an asset with a positive α , you still run the risk of a decline of the entire market or unexpected negative firm-specific surprises that might lead to a negative return on the asset. You can eliminate the market exposure by taking an offsetting position in the market portfolio. For every dollar you invest in the underpriced stock, take a short position worth β_i dollars in the market portfolio, and invest the β_i dollars in the riskfree asset. As shown in Table 11.1, the total rate of return you get is then $\alpha_i + r_f + \varepsilon_i$, which is insensitive to market movements, i.e. it is a market-neutral strategy. The sensitivity to firm-specific risk remains. Alternatively, you can invest only $\beta_i - 1$ dollars in the riskfree asset (if $\beta_i < 1$, this is a loan), which would bring the total investment to zero and the overall return to $\alpha_i + \varepsilon_i$. The attractiveness of either strategy depends on the ratio of α_i to the firm-specific standard deviation $\text{Std}[\varepsilon_i]$. This ratio is called the *information ratio*.

The Treynor-Black model described in Section 13.1 shows how assets with non-zero alphas are optimally combined with the market portfolio when the purpose is to maximize the Sharpe ratio of the overall portfolio.

If you cannot invest directly in the market portfolio, but only in a number of individual assets, then you can still use the Single-Index Model to calculate the necessary inputs to the mean-variance approach, namely

- N estimates of $\alpha_1, \dots, \alpha_N$
- N estimates of β_1, \dots, β_N
- N estimates of firm-specific variances, $\text{Var}[\varepsilon_1], \dots, \text{Var}[\varepsilon_N]$
- one estimate of the market risk premium, $E[r_m] - r_f$
- one estimate of market variance, $\text{Var}[r_m]$

In total $3N+2$ inputs. In the general mean-variance setting without imposing the structure of the Single-Index Model, $2N + \frac{1}{2}N(N - 1)$ inputs are required, cf. the discussion in Section 7.4.2. As long as $N > 4$, the Single-Index Model leads to a reduction in the number of inputs, and the reduction is large when N is large. With $N = 10$, you need 32 inputs instead of 65. With $N = 100$, it is 302 instead of 5150. Moreover, with the Single-Index Model, the inputs can be divided into asset-specific and market-level inputs which might be helpful in organizing the analysis that lead to the input estimates.

11.3 General multi-factor models

The Single-Index Model and other one-factor models are restrictive. You can easily come up with several factors that could influence how asset returns covary. Empirically, the cross-section of stock returns seems to be driven by more than one factor. Multi-factor models date back, at least, to [Chen, Roll, and Ross \(1986\)](#) who suggested a five-factor model with the factors being

1. the growth rate in industrial production,
2. changes in expected inflation quantified as changes in the yields of short-term government bonds,
3. unexpected inflation, i.e., the difference between realized and expected inflation,
4. changes in default risk premiums measured by the difference in yields on corporate bonds and yields on long-term government bonds, and
5. changes in the term premium measured by the slope of the government bond yield curve.

Later sections present more recent specific examples of multi-factor models. In this section we set up and derive basic properties of a general K -factor model, where K is an integer greater than or equal to 1.

11.3.1 Definition and basic properties

In a K -factor model, the return on each risky asset over a certain period is determined by K common factors and an idiosyncratic component:

$$r_i = E[r_i] + \beta_{i1}(F_1 - E[F_1]) + \cdots + \beta_{iK}(F_K - E[F_K]) + \varepsilon_i, \quad i = 1, 2, \dots, N, \quad (11.26)$$

where $E[\varepsilon_i] = 0$ by construction, whereas $\text{Cov}[F_k, \varepsilon_i] = 0$ and $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ (for $i \neq j$) by assumption. Here F_k is the value of factor k , which seen from the beginning of the period is a random variable, but the realized value will become known at the end of the period. The factor beta β_{ik} is a constant representing asset i 's sensitivity to factor k . Other things equal, an unexpected increase of one unit in factor k implies that the rate of return on asset i increases by β_{ik} . Together the K factors capture the systematic risk affecting all assets, whereas the residual ε_i represents the asset-specific return component.

We can write Eq. (11.26) in vector form as

$$r_i = E[r_i] + \boldsymbol{\beta}_i \cdot (\mathbf{F} - E[\mathbf{F}]) + \varepsilon_i, \quad i = 1, 2, \dots, N, \quad (11.27)$$

where

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iK} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_K \end{pmatrix}, \quad E[\mathbf{F}] = \begin{pmatrix} E[F_1] \\ E[F_2] \\ \vdots \\ E[F_K] \end{pmatrix}.$$

Let $\underline{\Sigma}_F = \text{Var}[\mathbf{F}]$ denote the $K \times K$ factor variance-covariance matrix.

Theorems 11.1 summarized some basic properties of a one-factor model. The analogue for a multi-factor model is the following theorem:

Theorem 11.4

Suppose the K -factor model (11.26) holds for assets $i = 1, 2, \dots, N$.

(a) We have

$$\text{Cov}[r_i, \mathbf{F}] = \underline{\Sigma}_F \boldsymbol{\beta}_i, \quad (11.28)$$

$$\text{Var}[r_i] = \boldsymbol{\beta}_i \cdot \underline{\Sigma}_F \boldsymbol{\beta}_i + \text{Var}[\varepsilon_i], \quad (11.29)$$

$$\text{Cov}[r_i, r_j] = \boldsymbol{\beta}_i \cdot \underline{\Sigma}_F \boldsymbol{\beta}_j, \quad (11.30)$$

(b) The return r_p on a portfolio represented by the weight vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ satisfies

$$r_p = \text{E}[r_p] + \boldsymbol{\beta}_p \cdot (\mathbf{F} - \text{E}[\mathbf{F}]) + \varepsilon_p = \text{E}[r_p] + \sum_{k=1}^K \beta_{pk} (F_k - \text{E}[F_k]) + \varepsilon_p, \quad (11.31)$$

where $\boldsymbol{\beta}_p = (\beta_{p1}, \dots, \beta_{pK})^\top$, and

$$\beta_{pk} = \sum_{i=1}^N \pi_i \beta_{ik}, \quad \varepsilon_p = \sum_{i=1}^N \pi_i \varepsilon_i$$

with non-systematic portfolio risk

$$\text{Var}[\varepsilon_p] = \sum_{i=1}^N \pi_i^2 \text{Var}[\varepsilon_i]. \quad (11.32)$$

Proof

(a) It follows from (11.26) and the assumption $\text{Cov}[F_k, \varepsilon_i] = 0$ that

$$\text{Cov}[r_i, F_1] = \beta_{i1} \text{Var}[F_1] + \beta_{i2} \text{Cov}[F_2, F_1] + \cdots + \beta_{iK} \text{Cov}[F_K, F_1] = \sum_{\ell=1}^K \beta_{i\ell} \text{Cov}[F_\ell, F_1]$$

and more generally

$$\text{Cov}[r_i, F_k] = \sum_{\ell=1}^K \beta_{i\ell} \text{Cov}[F_\ell, F_k] = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iK} \end{pmatrix} \cdot \begin{pmatrix} \text{Cov}[F_1, F_k] \\ \text{Cov}[F_2, F_k] \\ \vdots \\ \text{Cov}[F_K, F_k] \end{pmatrix}.$$

Hence

$$\begin{aligned}\text{Cov}[r_i, \mathbf{F}] &= \begin{pmatrix} \text{Cov}[r_i, F_1] \\ \text{Cov}[r_i, F_2] \\ \vdots \\ \text{Cov}[r_i, F_K] \end{pmatrix} \\ &= \begin{pmatrix} \text{Var}[F_1] & \text{Cov}[F_1, F_2] & \dots & \text{Cov}[F_1, F_K] \\ \text{Cov}[F_2, F_1] & \text{Var}[F_2] & \dots & \text{Cov}[F_2, F_K] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[F_K, F_1] & \text{Cov}[F_K, F_2] & \dots & \text{Var}[F_K] \end{pmatrix} \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iK} \end{pmatrix} = \underline{\Sigma}_F \boldsymbol{\beta}_i.\end{aligned}$$

With the factor structure (11.26), the return variance is

$$\begin{aligned}\text{Var}[r_i] &= \text{Var}[\beta_{i1}F_1 + \dots + \beta_{iK}F_K + \varepsilon_i] = \text{Var}[\beta_{i1}F_1 + \dots + \beta_{iK}F_K] + \text{Var}[\varepsilon_i] \\ &= \sum_{k=1}^K \sum_{\ell=1}^K \beta_{ik}\beta_{i\ell} \text{Cov}[F_k, F_\ell] + \text{Var}[\varepsilon_i] = \boldsymbol{\beta}_i \cdot \underline{\Sigma}_F \boldsymbol{\beta}_i + \text{Var}[\varepsilon_i],\end{aligned}$$

where the second equality is due to the assumption $\text{Cov}[F_k, \varepsilon_i] = 0$, the third comes from the standard variance rule (3.44), and the fourth equality follows from how vector and matrix products were defined in Section 4.2.

Finally, it follows from (11.26) and the assumptions $\text{Cov}[F_k, \varepsilon_i] = 0$ and $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ (for $i \neq j$) that

$$\text{Cov}[r_i, r_j] = \text{Cov}\left[\sum_{k=1}^K \beta_{ik}F_k, \sum_{\ell=1}^K \beta_{j\ell}F_\ell\right] = \sum_{k=1}^K \sum_{\ell=1}^K \beta_{ik}\beta_{j\ell} \text{Cov}[F_k, F_\ell] = \boldsymbol{\beta}_i \cdot \underline{\Sigma}_F \boldsymbol{\beta}_j,$$

where we have used the covariance rule (3.46).

(b) The proof is similar to that of Part (b) in Theorem 11.1 and is left for the reader.

From (11.28), we see that the factor beta vector of asset i is

$$\boldsymbol{\beta}_i = (\underline{\Sigma}_F)^{-1} \text{Cov}[r_i, \mathbf{F}], \quad (11.33)$$

which is the multi-dimensional version of (11.8). If the factors are uncorrelated so that $\text{Cov}[F_j, F_k] = 0$ for $j \neq k$, then $\text{Cov}[r_i, F_k] = \beta_{ik} \text{Var}[F_k]$ so that $\beta_{ik} = \text{Cov}[r_i, F_k]/\text{Var}[F_k]$, but this relation is incorrect when the factors are correlated.

In the variance decomposition (11.29), the systematic risk term is $\boldsymbol{\beta}_i \cdot \underline{\Sigma}_F \boldsymbol{\beta}_i$, which captures all the return variance stemming from the K factors. Together, Eqs. (11.29) and (11.30) show that the return variance-covariance matrix depends only on the factor variance-covariance matrix $\underline{\Sigma}_F$ and the assets' factor beta vectors $\boldsymbol{\beta}_i$ and non-systematic variance components $\text{Var}[\varepsilon_i]$. Here, the factor variance-covariance matrix contains K factor variances and $K(K - 1)/2$ factor covariances, while for each asset there are K factor betas and one residual variance. Without imposing the factor structure, the variance-covariance matrix of N assets involve a total of $(N + 1)N/2$ inputs, namely N variances and $N(N - 1)/2$ covariances. With the K -factor structure, we need $[(K + 1)K/2] + (K + 1)N$ inputs,

Number of factors, K	Number of assets, N				
	10	50	100	500	1000
1	21	101	201	1,001	2,001
2	33	153	303	1,503	3,003
3	46	206	406	2,006	4,006
4	60	260	510	2,510	5,010
5	75	315	615	3,015	6,015
Without factor structure	55	1,275	5,050	125,250	500,500

Table 11.2: Inputs to the variance-covariance matrix.

The table shows the number of inputs necessary to generate the full return variance-covariance matrix, when there are N assets and K factors. Without imposing any structure, the $N \times N$ return variance-covariance matrix contains $\frac{(N+1)N}{2}$ quantities. With a K -factor structure, the return variance-covariance matrix is determined by $\frac{(K+1)K}{2} + (K+1)N$ inputs.

which is still considerable smaller as long as K is small compared to N , cf. Table 11.2.

11.3.2 Alternative formulations of the K -factor model

Factor models are often estimated using the excess return on asset i , that is $r_i - r_f$, where r_f is the riskfree rate of return over the period. By subtracting r_f from both sides of Eq. (11.26), we obtain

$$r_i - r_f = E[r_i - r_f] + \beta_{i1}(F_1 - E[F_1]) + \cdots + \beta_{iK}(F_K - E[F_K]) + \varepsilon_i, \quad (11.34)$$

so the excess returns $r_i - r_f$ have exactly the same decomposition as the raw returns r_i .

We can also rewrite (11.26) as

$$r_i = a_i + \beta_{i1}F_1 + \cdots + \beta_{iK}F_K + \varepsilon_i, \quad (11.35)$$

where $a_i = E[r_i] - (\beta_{i1}E[F_1] + \cdots + \beta_{iK}E[F_K])$. Or, using excess returns,

$$r_i - r_f = a'_i + \beta_{i1}F_1 + \cdots + \beta_{iK}F_K + \varepsilon_i, \quad (11.36)$$

where $a'_i = a_i - r_f = E[r_i - r_f] - (\beta_{i1}E[F_1] + \cdots + \beta_{iK}E[F_K])$.

11.3.3 Estimation of inputs

Given a time series of observations of the factors and returns or excess returns on asset i , we can estimate β_i and $\text{Var}[\varepsilon_i]$ from a multi-variate linear regression of r_i or $r_i - r_f$ on F_1, \dots, F_K . Here, we can use any of the above formulations of the factor model. For example, the regression version of (11.36) is

$$r_{it} - r_{ft} = a'_i + \beta_{i1}F_{1t} + \cdots + \beta_{iK}F_{Kt} + \varepsilon_{it}. \quad (11.37)$$

This procedure implicitly assumes that the factor structure holds throughout the sample period with constant coefficients a'_i and β_{ik} . An example is given in Section 11.5.

11.4 The Arbitrage Pricing Theory

11.4.1 Assumptions and main result

In Section 4.5 we defined an arbitrage as a portfolio offering a riskfree profit. We argued that any arbitrage opportunities in financial markets should immediately be exploited and eliminated by some traders with the result that asset prices are generally set so that no arbitrage exists. The arbitrage pricing theory (APT) of Ross (1976) combines this idea with a multi-factor return-generating model.²

Theorem 11.5

In a one-period economy, suppose that

- (i) Returns are described by a K -factor model, i.e.

$$r_i = E[r_i] + \beta_{i1}(F_1 - E[F_1]) + \cdots + \beta_{iK}(F_K - E[F_K]) + \varepsilon_i, \quad i = 1, 2, \dots, N, \quad (11.38)$$

with $E[\varepsilon_i] = \text{Cov}[F_k, \varepsilon_i] = \text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ for all i, j , and k with $i \neq j$.

- (ii) Investors can trade in sufficiently many securities to diversify away all idiosyncratic risk.
- (iii) Prices are set so that no arbitrage opportunities exist.

Then a factor risk premium RP_k exists for each factor so that the expected return on any asset i is given by

$$E[r_i] = r_f + \beta_{i1} \times \text{RP}_1 + \cdots + \beta_{iK} \times \text{RP}_K, \quad (11.39)$$

and the expected return on any portfolio with weight vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^\top$ is similarly given by

$$E[r_p] = r_f + \beta_{p1} \times \text{RP}_1 + \cdots + \beta_{pK} \times \text{RP}_K, \quad (11.40)$$

where $\beta_{pk} = \sum_{i=1}^N \pi_i \beta_{i1}$ for $k = 1, \dots, K$.

In itself, the K -factor model (11.38) says nothing about what expected asset returns should be. But when we combine it with the APT, we get a statement about expected returns. In this sense, the APT gives us an *equilibrium version of the factor model* by stating that the total risk premium $E[r_i] - r_f$ on asset i is due to its exposure to the K priced risks. The conclusion (11.39) of the APT is basically a K -factor version of the CAPM relation between market betas and expected returns. For each factor affecting the returns of all or most assets, there is an associated risk premium. Of course, it is possible that one or more risk premiums are zero so that not all common risks are priced.

If we substitute the APT conclusion (11.39) into the return-generating model (11.38), we obtain

$$r_i - r_f = \sum_{k=1}^K \beta_{ik} (\text{RP}_k + F_k - E[F_k]) + \varepsilon_i. \quad (11.41)$$

²For further theoretical work on the APT, see Huberman (1982), Chen and Ingersoll (1983), Dybvig (1983), Grinblatt and Titman (1983, 1985, 1987), Connor (1984), and Ingersoll (1984).

In other words, in the equation

$$r_i - r_f = \alpha_i + \sum_{k=1}^K \beta_{ik} (\text{RP}_k + F_k - \mathbb{E}[F_k]) + \varepsilon_i, \quad (11.42)$$

the α_i should be zero according to the APT.

The APT says that factor risk premiums $\text{RP}_1, \dots, \text{RP}_K$ exist, but not how big they are. How do we determine the risk premiums? One approach is the following: Suppose we know the riskfree rate and for K specific assets or portfolios we know their factor betas and their expected returns. Applying (11.39) to these K assets or portfolios leaves us with K equations that can (in most cases) be solved for the K unknowns, namely the factor risk premiums. In cases where the factors are specific returns or return differences, we can determine the factor risk premiums more directly, as we illustrate in specific factor models in the following subsections.

Regarding the proof of Theorem 11.5, first note that the portfolio equation (11.40) follows directly from (11.39) and the relation $r_p = \sum_{i=1}^N \pi_i r_i$ as in the proof of Part (b) of Theorem 11.1. We will not present a formal proof of the APT, but here is the intuition. Suppose that (11.39) and, consequently, (11.40) are violated, so that for some portfolio p , the alpha

$$\alpha_p = \mathbb{E}[r_p] - \{r_f + \beta_{p1} \times \text{RP}_1 + \dots + \beta_{pK} \times \text{RP}_K\}$$

is large and the non-systematic risk $\text{Std}[\varepsilon_p]$ is small. You can adjust your exposure to the factors by trading in factor-mimicking portfolios. For each factor k , the associated factor-mimicking portfolio is a well-diversified portfolio with a unit sensitivity to factor k and a zero sensitivity to other factors. If you invest in the mispriced portfolio p above, you can eliminate your exposure to systematic risks by taking off-setting positions in the factor-mimicking portfolios and thus form a factor-neutral or market-neutral strategy (this is a key activity of some hedge funds). If many investors do this, prices and hence expected returns on both the attractive portfolio and the factor-mimicking portfolios adjust until α_p is zero or at least very close to zero.

The second assumption of the APT is critical for the argument that α_p should be equal to zero. Recall Eq. (11.32) which shows the residual or idiosyncratic risk of a portfolio of the available assets. A complete diversification of idiosyncratic risks requires *infinitely many assets*, which we obviously do not have in real life. Based on the above investment argument, we should rather expect the approximate relation

$$\mathbb{E}[r_i] \approx r_f + \beta_{i1} \times \text{RP}_1 + \dots + \beta_{iK} \times \text{RP}_K \quad (11.43)$$

with the approximation being *very* precise for large, well-diversified portfolios, but *less* precise for individual assets. Maybe the residual return component of a particular asset tends to be positive exactly in those states of the world in which the investors appreciate an extra return the most. Then that asset is attractive and will have a higher price and a lower expected return than prescribed by its sensitivity to the factors. Unfortunately, it is difficult to test and apply “approximate theories”, and this is certainly a problem of the APT. Applications therefore assume that the strict, non-approximate version (11.39) holds.

11.4.2 The APT and the Single-Index Model

Suppose the Single-Index Model

$$r_i = E[r_i] + \beta_i(r_m - E[r_m]) + \varepsilon_i$$

holds for all assets and thus for all portfolios. Then the strict version of the APT implies that a factor risk premium RP_m exists so that

$$E[r_i] = r_f + \beta_i \times RP_m$$

for all assets and portfolios, so in particular for the market portfolio itself we have

$$E[r_m] = r_f + \beta_m \times RP_m = r_f + RP_m \Rightarrow RP_m = E[r_m] - r_f.$$

Hence, we get

$$E[r_i] = r_f + \beta_i(E[r_m] - r_f)$$

for all assets and portfolios, as in the CAPM. By super-imposing the APT argument upon the Single-Index Model, we get the same conclusion as in the CAPM.

The derivation of the CAPM in Chapter 10 was based on a model of the optimal investments of all investors, namely the mean-variance model, and then an equilibrium argument was superimposed to conclude that the tangency portfolio is in fact the market portfolio. Among the assumptions were that investors agreed upon the location of the efficient frontier of risky assets, which is questionable. The assumptions about investor behavior underlying the APT are less restrictive, saying simply that investors exploit arbitrage opportunities. On the other hand, the APT really leads to only an approximate equation for expected returns and is thus less precise than the CAPM.

11.4.3 Returns as factors

In the Single-Index Model the factor is the return on the market portfolio. In many multi-factor models, one or more of the factors are also returns on certain portfolios. This is not really a limitation since any non-return factor can be replaced by the return on a factor-mimicking portfolio. Using returns as factors has the benefit that we can easily and frequently observe the value of the factors.

Suppose factor k , i.e F_k , is the return on a specific portfolio. Let us denote this return by r_{Fk} . Since

$$r_{Fk} = E[r_{Fk}] + 1 \times (r_{Fk} - E[r_{Fk}]),$$

this portfolio has a unit beta with respect to itself and a zero beta with respect to all other factors. Hence, according to the Arbitrage Pricing Theory, the expected return is

$$E[r_{Fk}] = r_f + RP_k$$

so that

$$RP_k = E[r_{Fk}] - r_f,$$

i.e., the risk premium associated with a return factor is simply the expected value of that return less the riskfree rate—just as in the CAPM. Note that the k 'th term of the sum in (11.41) and (11.42) is then

$$RP_k + F_k - E[F_k] = E[r_{Fk}] - r_f + r_{Fk} - E[r_{Fk}] = r_{Fk} - r_f.$$

Now consider the case where all K factors are returns. Then the return-generating model can be rewritten as

$$r_i - r_f = \sum_{k=1}^K \beta_{ik} (r_{Fk} - r_f) + \varepsilon_i = \boldsymbol{\beta}_i \cdot (\mathbf{r}_F - r_f \mathbf{1}) + \varepsilon_i, \quad (11.44)$$

where $\boldsymbol{\beta}_i$ is the vector of factor betas $\beta_{i1}, \dots, \beta_{iK}$, where \mathbf{r}_F is the vector of the returns r_{F1}, \dots, r_{FK} of the factor portfolios, and where $\mathbf{1}$ is a K -dimensional vector of ones. If the APT holds exactly, the expected return on each asset becomes

$$\mathbb{E}[r_i] = r_f + \sum_{k=1}^K \beta_{ik} (\mathbb{E}[r_{Fk}] - r_f) = r_f + \boldsymbol{\beta}_i \cdot (\mathbb{E}[\mathbf{r}_F] - r_f \mathbf{1}). \quad (11.45)$$

In such a formulation, we are really pricing the individual assets *relative to* the factor-mimicking portfolios. The factor sensitivities are in this case estimated in the multi-variate regression

$$r_{it} - r_{ft} = \alpha_i + \sum_{k=1}^K \beta_{ik} (r_{Fkt} - r_{ft}) + \varepsilon_{it}, \quad (11.46)$$

where r_{Fkt} is the rate of return in period t on the portfolio corresponding to factor k . Examples are given in the subsequent sections. If the factor structure and the corresponding equilibrium model are both correct, the estimate of α_i should be zero or at least not significantly different from zero. A non-zero α_i indicates a better (positive α_i) or worse (negative α_i) performance than prescribed by the equilibrium model. If you expect this to continue, the asset would be a good buy or sell.

11.4.4 Return differences as factors

As we shall see in the following section, some of the factors in many popular multi-factor models are return differences. To be precise, the value of the factor is the difference between the return on one specific portfolio and the return on another specific portfolio. The factor is basically what you would gain on implementing a long-short strategy taking a long position in the first portfolio and a short position in the second. For example, suppose factor k is

$$F_k = r_\ell - r_s,$$

where the subscripts ℓ and s refer to the long and short leg of the strategy. Since

$$r_\ell - r_s = \mathbb{E}[r_\ell - r_s] + 1 \times (r_\ell - r_s - \mathbb{E}[r_\ell - r_s]) = \mathbb{E}[r_\ell - r_s] + 1 \times (F_k - \mathbb{E}[F_k]),$$

the return difference has—of course—a unit beta with respect to itself and zero betas with respect to other factors. This requires that the long and short portfolios have identical betas with respect to each of the other factors, i.e., $\beta_{\ell j} = \beta_{sj}$, for all $j = 1, \dots, K$ except $j = k$. According to the APT-equation (11.39), we have

$$\begin{aligned} \mathbb{E}[r_\ell] &= r_f + \beta_{\ell 1} \times \text{RP}_1 + \cdots + \beta_{\ell k} \times \text{RP}_k + \cdots + \beta_{\ell K} \times \text{RP}_K, \\ \mathbb{E}[r_s] &= r_f + \beta_{s 1} \times \text{RP}_1 + \cdots + \beta_{s k} \times \text{RP}_k + \cdots + \beta_{s K} \times \text{RP}_K \end{aligned}$$

and therefore

$$\mathbb{E}[r_\ell - r_s] = \underbrace{(\beta_{\ell 1} - \beta_{s1})}_{=0} \times \text{RP}_1 + \cdots + \underbrace{(\beta_{\ell k} - \beta_{sk})}_{=1} \text{RP}_k + \cdots + \underbrace{(\beta_{\ell K} - \beta_{sK})}_{=0} \times \text{RP}_K = \text{RP}_k$$

so that the risk premium related to the return difference factor k is simply the expected return difference,

$$\text{RP}_k = \mathbb{E}[r_\ell - r_s] = \mathbb{E}[F_k].$$

Hence,

$$\text{RP}_k + F_k - \mathbb{E}[F_k] = F_k = r_\ell - r_s.$$

If all factors are return differences, $F_k = r_{\ell k} - r_{sk}$, Equation (11.41) becomes

$$r_i - r_f = \sum_{k=1}^K \beta_{ik} (r_{\ell k} - r_{sk}) + \varepsilon_i.$$

In particular, the expected return on asset i is

$$\mathbb{E}[r_i] = r_f + \sum_{k=1}^K \beta_{ik} (\mathbb{E}[r_{\ell k}] - \mathbb{E}[r_{sk}]). \quad (11.47)$$

Note that if factor k is the return on a certain portfolio, we can replace it by the same portfolio's return in excess of the riskfree rate, which is a return difference. Hence, the above derivation confirms that the associated risk premium is the expected excess return on the portfolio.

11.5 The Fama-French models

The best known multi-factor model is the Fama-French three-factor model developed by [Fama and French \(1992\)](#). The three factors are

1. *market*: the return on a broad stock market index;
2. *small-minus-big* or *SMB*: the return on a portfolio of stocks in small companies (according to the market value of all stocks issued by the firm) minus the return on a portfolio of stocks in large companies;
3. *high-minus-low* or *HML*: the return on a portfolio of stocks issued by firms with a high book-to-market value (value stocks) minus the return on a portfolio of stocks in firms with a low book-to-market value (growth stocks).

The model thus adds two factors, both return differences, to the Single-Index Model.

As discussed in Section 6.5, studies have found that stocks in small companies tend to provide higher returns than stocks in large companies, and value stocks tend to provide higher returns than growth stocks. And, as explained in Section 10.2, the return differences cannot be fully justified by their different exposures to the overall market return. These observations motivate the use of the SMB and HML factors.

Fama and French define the SMB and HML factors from returns on six portfolios of stocks listed either the NYSE, the AMEX, or the NASDAQ exchange. The construction of the six portfolios goes as follows.³ Once a year since 1926, the median market equity value of all NYSE listed stocks is determined. All stocks listed at one of the three exchanges

³For additional details, see the homepage of Professor French at <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/>.

Market cap	Book-to-market ratio		
	Lower than 30th pct	Between 30th and 70th pct	Above 70th pct
Below median	SG (Small Growth)	SN (Small Neutral)	SV (Small Value)
Above median	BG (Big Growth)	BN (Big Neutral)	BV (Big Value)

Table 11.3: Construction of the six size-and-BM sorted portfolios.

The table shows how the stocks are allocated to the six Fama-French portfolios based on the size, i.e. market capitalization, and the book-to-market ratio.

	SG	SN	SV	BG	BN	BV
July 1926	44	69	94	84	102	35
December 2023	451	803	1190	416	341	123
Average	564	638	731	297	252	109
Minimum	26	67	90	84	100	30
Maximum	1807	1688	2165	769	461	215

Table 11.4: Stocks in the six size-and-BM sorted portfolios.

The table shows statistics on the number of stocks in the six Fama-French portfolios sorted on size and book-to-market ratio. The average, minimum, and maximum number is based on the full sample period from July 1926 to December 2023. Data downloaded from the homepage of Professor Kenneth French on May 7, 2024.

are then labelled as small or big in that year depending whether their market equity value is below or above the NYSE median. Similarly, the 30th and 70th percentiles in the distribution of the book-to-market ratios of NYSE listed stocks are determined. All stocks listed at one of the three exchanges are then labelled as growth stocks if they have a lower book-to-market ratio than the 30th percentile, as value stocks if their book-to-market ratio exceeds the 70th percentile, and neutral stocks if their book-to-market ratio fall between these percentiles. Now each stock is allocated to one of six portfolios as shown in Table 11.3.

For each of the six portfolios, a value-weighted return is calculated each month. Let r_{SG} denote the rate of return on the Small Growth portfolio and use similar notation for the other portfolios. Then the SMB and HML factors for that month are calculated as

$$\text{SMB} = \frac{1}{3} (r_{SV} + r_{SN} + r_{SG}) - \frac{1}{3} (r_{BV} + r_{BN} + r_{BG}), \quad (11.48)$$

$$\text{HML} = \frac{1}{2} (r_{SV} + r_{BV}) - \frac{1}{2} (r_{SG} + r_{BG}). \quad (11.49)$$

As the market and book values of companies vary over time, stocks occasionally move to a different portfolio. Note that the ‘breakpoints’ for both size and book-to-market are based only on the NYSE listed stocks. Since the size and book-to-market distributions are different among AMEX and NASDAQ stocks than among NYSE stocks, the number of stocks included in the portfolios is not necessarily evenly split in the size dimension or split 30:40:30 in the book-to-market dimension. In fact, Table 11.4 shows that, typically, there are more value stocks than growth stocks among the small companies, while there are more growth stocks than value stocks among the big companies. In total, there are more small stocks than large stocks. Also, note that the number of stocks in each portfolio has generally grown substantially over the years.

In mathematical terms, the Fama-French three-factor model assumes that returns over

a given period satisfy

$$r_i - r_f = \alpha_i + \beta_{i,m} (r_m - r_f) + \beta_{i,SMB} SMB + \beta_{i,HML} HML + \varepsilon_i, \quad (11.50)$$

where the residual ε_i has zero covariance with r_m , SMB, and HML, and the residuals of any two assets also have zero covariance. Here $\beta_{i,m}$, $\beta_{i,SMB}$, and $\beta_{i,HML}$ are the sensitivities of the return on asset i with respect to the three factors. The resulting APT-equation is

$$E[r_i] = r_f + \beta_{i,m} (E[r_m] - r_f) + \beta_{i,SMB} E[SMB] + \beta_{i,HML} E[HML], \quad (11.51)$$

using the insights about risk premiums from Sections 11.4.3 and 11.4.4. According to the equilibrium version of the model, the α_i in (11.50) should equal zero as explained in the preceding section. Fama and French (1996) show that such a model gives a good fit to U.S. stock market data over the period 1963–1993, and Fama and French (2012) report that the same is true for many other countries.

As acknowledged by Fama and French, their empirical analysis does not explain *why* this three-factor model performs well and what the underlying pricing mechanisms might be. Maybe the small-minus-big and the high-minus-low portfolios are mimicking some underlying state variables capturing relevant information about future consumption and investment opportunities in line with the Consumption-based CAPM and the Intertemporal CAPM, but the link is not clear.

Fama and French (1996) suggest that the success of their three-factor model comes from a premium on financial distress. Small value stocks tend to be the stocks of firms that have performed rather poorly in the recent past and are thus more likely to experience financial distress, see Chan and Chen (1991). In general, small companies may be more sensitive to recessions. If small stocks perform particularly poorly in recessions—where investors would appreciate high returns the most—investors will require higher expected return on small stocks.

There are also various explanations of the value premium and the associated high-minus-low factor. Liew and Vassalou (2000) show that the SMB and HML factors are highly correlated with future growth in GDP and thus future consumption and investment opportunities. Value firms have more tangible capital. In recessions they suffer more from costly excess capacity, whereas growth firms rely more on new investments that can easily be deferred to better times. Value stocks are thus more “pro-cyclical”, which commands a higher expected return. Other justifications of the value premium were suggested by, Carlson, Fisher, and Giammarino (2004), Zhang (2005), Cooper (2006), Gomes, Yaron, and Zhang (2006), and Garlappi and Yan (2011).

To estimate the factor betas we need time series of observations on the three factors. These are available on the homepage of Professor French. Figure 11.1 shows how the three Fama-French factors and the riskfree rate have evolved from January 1950 to December 2023. The monthly returns are in percent and not annualized. Note that the returns on both the SMB and the HML strategies vary substantially and are sometimes very negative. For example, over the year of 1998 the SMB factor was -21.3% (this is simply the sum of the monthly values of the factor), i.e., small stocks significantly underperformed large stocks in that year in which the excess return on the stock market index was roughly 19.8% . In 1999, where the excess market return was 19.1% , the HML factor was -27.4% . The HML strategy has also performed badly in the crisis period from 2007 to 2011. While the strategies seem attractive in the long run, they lead to severe losses in some periods.

Table 11.5 presents some summary statistics for two different starting dates of the sample. Over the full sample 1926–2023, the average monthly excess market return has been

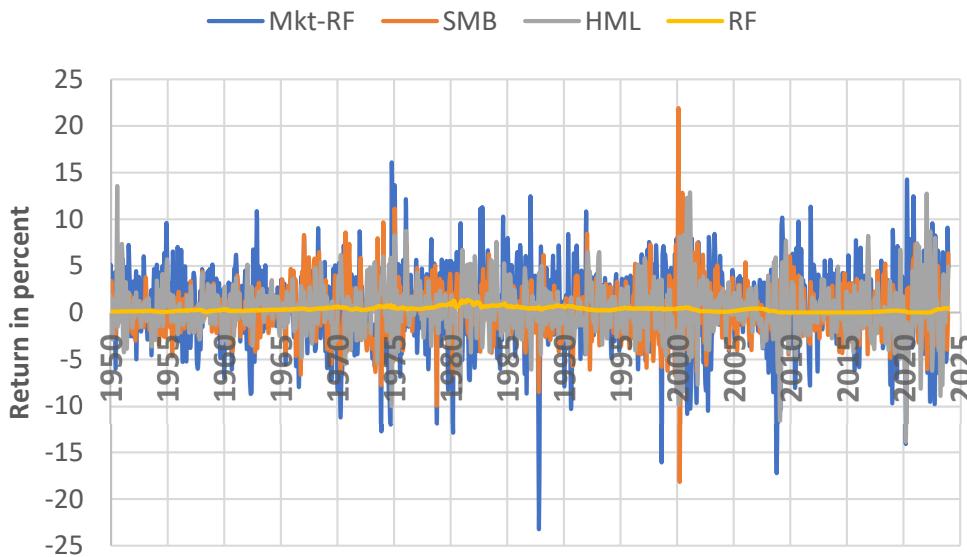


Figure 11.1: A time series of Fama-French factors.

The graph is based on monthly observations of the riskfree rate, the excess market return, the SMB factor, and the HML factor from January 1950 to December 2023. The data were downloaded from the homepage of Professor Kenneth French on May 7, 2024.

0.678%, which corresponds to an annual excess return of $12 \times 0.678\% \approx 8.137\%$ ignoring compounding or $(1.00678)^{12} - 1 \approx 0.08447 = 8.447\%$ with compounding.⁴ The standard deviation is 5.345% per month or $\sqrt{12} \times 5.345\% \approx 18.514\%$ per year if we ignore compounding and 20.098% if we include compounding by using Eq. (3.82). These estimates are in line with the numbers reported in Section 6.6. The 95% confidence interval for the monthly average excess market return goes from 0.372% to 0.985%. For the recent 20-year period, the average excess market return is slightly higher and the standard deviation slightly lower.

The average SMB factor is positive in both periods but not statistically significant in the recent 20-year period as can be seen from the fact that the 95% confidence interval includes zero. In fact, given the estimated positive market-beta of SMB of around 0.19 in both samples, the average return *should* be positive according to the CAPM. The point estimate of the CAPM-alpha is positive in the full sample and negative in the recent sample but insignificant in both periods. In the full sample, the average HML factor is positive and much bigger than the SMB factor and clearly significant. The average return difference exceeds what it should be according to the CAPM as reflected by the clearly significant CAPM-alpha of 0.245%. This indicates a violation of the CAPM. However, in the 2004-23 sample the average HML factor is slightly negative. The market beta of HML is still positive, so the CAPM-alpha of HML is negative although insignificant. Hence, in the recent 20-year period, both SMB and HML show insignificant abnormal returns relative to the CAPM.

We further observe that both the HML and SMB return differences have a large kurtosis in the full sample, indicating a large number of extremely positive or extremely negative realizations. The substantial risk of a highly negative outcome may deter investors away

⁴ As discussed in Section 2.5, excess returns should not really be annualized this way, but the error made by doing so is small.

	July 1926-Dec 2023				Jan 2004-Dec 2023			
	Mkt-RF	SMB	HML	RF	Mkt-RF	SMB	HML	RF
Average	0.678	0.186	0.348	0.268	0.770	0.040	-0.040	0.112
- Lower 95%	0.372	0.004	0.143	0.253	0.204	-0.274	-0.444	0.094
- Upper 95%	0.985	0.368	0.553	0.282	1.335	0.353	0.363	0.130
Std dev	5.345	3.173	3.569	0.251	4.446	2.468	3.175	0.142
Skew	0.158	1.824	2.061	1.122	-0.522	0.264	-0.019	1.118
Kurt	7.412	18.507	18.127	1.428	1.361	-0.127	3.001	-0.082
Correlations	1.000	0.317	0.228	-0.066	1.000	0.358	0.155	-0.063
	0.317	1.000	0.117	-0.050	0.358	1.000	0.091	-0.076
	0.228	0.117	1.000	0.020	0.155	0.091	1.000	-0.037
	-0.066	-0.050	0.020	1.000	-0.063	-0.076	-0.037	1.000
CAPM β		0.188	0.152			0.199	0.111	
- Lower 95%		0.156	0.115			0.133	0.020	
- Upper 95%		0.220	0.190			0.265	0.201	
CAPM α		0.059	0.245			-0.114	-0.125	
- Lower 95%		-0.115	0.044			-0.412	-0.531	
- Upper 95%		0.233	0.446			0.184	0.280	

Table 11.5: Summary statistics for Fama-French factors.

Based on monthly percentage rates of return. The data were downloaded from the homepage of Professor Kenneth French on May 7, 2024.

from these strategies in spite of their (more or less) attractive average returns in the long run. In addition, the implementation of the HML and SMB strategies is likely to involve some relatively illiquid stocks and potentially substantial transaction costs.

The factor betas of a given asset or a given portfolio can be estimated by multivariate linear regression. The following example does this for the returns on Microsoft stocks over the period from January 2019 to December 2023 using monthly observations (see the Excel file `MicrosoftRegressions.xlsx` in the supplementary material to these lecture notes). Recall that Example 11.1 performed the corresponding univariate linear regression relevant for the single-index model with the excess market return as the only factor.

Example 11.2

We regress excess returns on Microsoft stocks on the excess return on the stock market, the SMB factor, and the HML factor using monthly observations from January 2019 to December 2023. This is the regression version of Eq. (11.50). The data on the factors were downloaded from the homepage of Professor Kenneth French, whereas Microsoft returns were derived from adjusted monthly closing prices downloaded from CRSP. The data were downloaded on April 16, 2024.

Figure 11.2 displays the Excel output from this regression. The market beta of Microsoft is 0.9504 in the FF3 regression compared to 0.8223 in the single-index regression. The inclusion of the SMB and HML factors thus affects our estimate of Microsoft's market sensitivity somewhat. Microsoft has a beta of -0.5958 with respect to SMB, which seems reasonable given that Microsoft has a large market capitalization. Note that the SMB beta is significantly different from zero. The HML beta of Microsoft is -0.4883 and is also significantly different from zero.

SUMMARY OUTPUT

Regression Statistics						
Multiple R	0.8516					
R Square	0.7252					
Adjusted R Square	0.7105					
Standard Error	3.4112					
Observations	60					

ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	1719.7541	573.2514	49.2646	1.0156E-15	
Residual	56	651.6255	11.6362			
Total	59	2371.3796				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.1707	0.4523	2.5881	0.0123	0.2646	2.0769
Mkt-RF	0.9504	0.0848	11.2050	0.0000	0.7805	1.1204
SMB	-0.5958	0.1617	-3.6855	0.0005	-0.9197	-0.2720
HML	-0.4883	0.0945	-5.1649	0.0000	-0.6777	-0.2989

Figure 11.2: Microsoft in the Fama-French 3-factor model.

The Excel output from a regression of excess monthly Microsoft returns on the three Fama-French factors over the period from January 2019 to December 2023.

The estimate of α is now 1.1707% compared to 1.3655% in the single-index model. Hence, a part of the apparent extraordinarily high returns on Microsoft in the single-index setting is explained by Microsoft's exposure to the SMB and HML factors. The alpha estimate is still substantial and significantly different from zero.

Not surprisingly, the inclusion of two additional explanatory variables increases the R^2 from 52.0% to 72.5% and reduces the residual standard deviation (shown as standard error in the regression output) from 4.43% to 3.42%.

About 25 years after their original work, Fama and French (2015, 2016) extended their three-factor model to a five-factor model by adding

4. *operating profitability* represented by RMW (robust-minus-weak), the return difference between a portfolio of stocks in the most profitable firms and a portfolio of stocks in the least profitable firms, and
5. *investment* represented by CMA (conservative-minus-aggressive), the return difference between a portfolio of stocks in firms investing conservatively and a portfolio of stocks in firms investing aggressively.

The operating profitability of a company is defined as the annual revenues minus cost of goods sold and expenses on interests, selling, and administration, all divided by the book value of the equity. The investment variable is the growth rate of total assets in the past year. The precise definitions and time series of observations of the two extra factors are also available at Professor French's homepage. Profitability of each company is measured by the ratio of (i) revenues minus cost of goods sold, minus selling, general, administrative, and interest expenses to (ii) the book value of equity. The investments of each company are measured by the percentage change in the company's total assets over the most recent year.

	Low	2	3	4	High
<i>Panel A: Size-B/M portfolios</i>					
Small	0.64	1.14	1.13	1.33	1.47
2	0.88	1.15	1.22	1.26	1.38
3	0.91	1.17	1.11	1.24	1.36
4	1.01	1.01	1.08	1.19	1.23
Big	0.95	0.91	0.94	0.88	1.04
<i>Panel B: Size-profitability portfolios</i>					
Small	0.90	1.31	1.26	1.38	1.20
2	0.96	1.16	1.19	1.18	1.32
3	0.91	1.09	1.13	1.16	1.29
4	0.96	1.05	1.07	1.11	1.18
Big	0.72	0.81	0.91	0.92	1.01
<i>Panel C: Size-investment portfolios</i>					
Small	1.35	1.31	1.32	1.22	0.73
2	1.24	1.25	1.28	1.28	0.86
3	1.28	1.25	1.14	1.16	0.92
4	1.15	1.12	1.12	1.14	0.98
Big	1.10	0.96	0.91	0.93	0.91

Table 11.6: Average returns on double-sorted portfolios.

The table shows the average monthly percent returns of portfolios formed on size and book-to-market (Panel A), size and profitability (Panel B), and size and investments (Panel C). The data are from July 1963 to December 2023 (726 months).

Table 11.6 presents the average monthly percentage return on various portfolios over the period from July 1963 to December 2023. The portfolios are defined by a double-sort on size and either book-to-market, profitability, or investments. At the end of each June, stocks are allocated to five size groups. Stocks are allocated independently to five book-to-market (B/M) groups. The intersections of the two sorts produce 25 value-weighted size-B/M portfolios. The size-profitability and size-investment portfolios are formed in the same way. Panel A confirms that value stocks (high book-to-market ratio) in the long run have outperformed growth stocks (low book-to-market), in particular among the small companies. Panel B shows that stocks in profitable firms provide higher average returns than stocks in unprofitable firms, and this holds in every size quintile. Such findings were first reported by [Novy-Marx \(2013\)](#). Panel C illustrates that stock returns are negatively related to the company's investments, i.e., the stocks of conservative firms seem to outperform stocks of aggressive firms. This is also true in every size quintile, although most predominant among small firms.

11.6 The factor zoo

11.6.1 Hunting factors

Both quantitative analysts in the financial industry and investment researchers at universities and business schools have strong incentives to find factors that are useful in explaining return differences across assets. The financial industry can potentially exploit a new factor to generate extraordinary profits either by investing own money or by setting up factor-driven hedge funds or mutual funds inviting other investors to participate at a fee. Researchers detecting a new factor can often publish a research paper in a renowned academic journal leading to prestige, salary bonuses, and promotions in the academic

system, and maybe lucrative consulting gigs or partnerships in investment funds.

To document a new factor, you basically need to show that the factor is associated with an average return or return-risk tradeoff which is better than could be expected given the well-known standard pricing factors. Just finding a factor for which the average return on the associated trading strategy is positive is not sufficient. You really need to show that the trading strategy gives a significantly positive abnormal return—a significant alpha—relative to the CAPM or the Fama-French models. As plenty of historical data on stock returns across many countries are easily available, the historical performance of a specific trading strategy defined by some new candidate factor is relatively straightforward to investigate. If the strategy provides a statistically significant abnormal return, a new factor is—apparently—discovered.

The intense search for factors driven by these incentives have led to the identification of hundreds of seemingly significant factors driving the cross-section of stock returns, see [Harvey, Liu, and Zhu \(2016\)](#), [Hou, Xue, and Zhang \(2020\)](#), and [Jensen, Kelly, and Pedersen \(2023\)](#), among others. The high number of factors has led to the term *factor zoo* apparently coined by [Cochrane \(2011\)](#).

Several recent papers question the replicability and validity of many of the studies claiming to identify a new factor. By carefully choosing the data period, the stocks included in the sample, or the precise definition of the trading strategy or factor, a researcher can influence the summary statistics of the returns and the *p*-values indicating significance of a performance measure. Given that many different strategies or factors can easily be tested, it may not be so surprising that some factors show up significant.

[Harvey, Liu, and Zhu \(2016\)](#) argue that, due to the extensive data mining and dubious “*p*-hacking” practices, the *t*-ratio hurdle for statistical significance should be much bigger than the usual 2.0, and only few recently suggested factors surpass the higher hurdle. It should also be acknowledged that there are various challenges in the application of proper empirical methods in this field. For example, in order to calculate abnormal returns relative to the CAPM, you need to know the market betas of the assets involved in the trading strategy and the estimation of betas is not straightforward, see [Berk, Green, and Naik \(1999\)](#) and [Gomes, Kogan, and Zhang \(2003\)](#), among others.

Several studies have shown that the significant profitability of a factor reported in the original study often decreases—and sometimes disappears—when the sample is extended to include earlier or later return data or to include data from other countries, see, e.g., [McLean and Pontiff \(2016\)](#), [Linnainmaa and Roberts \(2018\)](#), and [Arnott, Harvey, Kalesnik, and Linnainmaa \(2019\)](#). The poorer performance after publication of a factor may potentially be explained by the market adjusting to the new factor by repricing stocks accordingly, but the inclusion of earlier data or data from other countries should generally not weaken the significance of a factor.

Many of the identified factors are closely related. Intuitively, if one factor is significant, then a variable highly correlated with that factor may also turn out to be significant. [Jensen, Kelly, and Pedersen \(2023\)](#) consolidate the hundreds of identified factors into 13 *themes*, and each of these themes can then be represented by a variety of specific factors.

Like the Fama-French factors, most other factors are specified as return differences, i.e. the return on one portfolio minus the return on another portfolio. Disregarding transaction costs and other practical issues, the factor return is then obtained by a zero net investment strategy taking a long position in the first portfolio and a short position in the second portfolio. Stocks are allocated to the two portfolios based on the value of some variable or characteristic that varies across companies. The two portfolios are rebalanced regularly

at a frequency depending on the specific variable or characteristic. Obviously, a factor return calculated without taking transaction costs into account is exaggerating the returns obtainable by investors in real-life markets. Some factors involve more rebalancing than others and some involve more trading in stocks with high transaction costs (typically stocks of small companies) than others. [Novy-Marx and Velikov \(2016\)](#) show that, for most of the long-short portfolios often used as factors, the large return originally reported is significantly reduced—or completely eliminated—after accounting for transaction costs. In the same vain, [Patton and Weller \(2020\)](#) conclude that, after accounting for costs, mutual funds earn low returns on value strategies and no returns on momentum strategies.

Finally, as an investor thinking about applying a factor for setting up an investment strategy, you would probably be more comfortable if you have some understanding of *why* the factor should work. If the factor is associated with seemingly abnormal returns, why is it that some investors are willing to forego those abnormal returns? Who is on the other side of the trades? Is the factor premium a compensation for some risks not captured by standard models and thus ignored when calculating the abnormal returns? More generally, can the factor be theoretically justified? For example, this could be the case if the factor capture variations in market betas or the market risk premium (see Section 10.3.3), capture information about future investment opportunities (see Section 8.3), or capture variations in investors' marginal utilities of consumption (see Section 10.4). The CAPM and similar models are, of course, simplistic representations of reality, so real-life investors may think of risk in a more nuanced way than captured by the models, and they may face various constraints and frictions excluded from the models. The abnormal returns associated with a factor can also be due to systematic and seemingly irrational behavior of some investor groups. So another useful question to confront a factor with is: Can the factor be related to some well-known and reliable behavioral biases among some other investors? Without a good answer to any of these questions, you should probably be reluctant to bet on the factor delivering abnormal returns also in the future.

11.6.2 The largest animals in the zoo

Amid the confusion about how many factors and which factors contribute to explaining cross-sectional patterns in stock returns, it seems fair to say that many large funds and leading investment professionals seem to agree on 4-5 factors or themes being important in some version and in most periods and most economic scenarios: value, momentum, quality, defensive, and maybe size. See, for example, the discussions in [Ilmanen, Israel, Moskowitz, Thapar, and Lee \(2021\)](#), [Blitz \(2022\)](#), and [Aghassi, Asness, Fattouche, and Moskowitz \(2023\)](#), as well as the description of the MSCI factor indices at <https://www.msci.com/msci-factor-indexes>. Let us briefly describe each of these factors:

Value. A value stock is broadly understood as a “cheap” stock in the sense that the market price is low relative to some fundamental indicator of the value of the company. [Rosenberg, Reid, and Lanstein \(1985\)](#) and [Fama and French \(1992\)](#) used the book-to-market value of the equity, but other price ratios have been applied, including the dividend-price ratio ([Litzenberger and Ramaswamy 1979](#)), the earnings-price ratio ([Basu 1983](#)), the sales-price ratio ([Barbee, Mukherji, and Raines 1996](#)), and the net payout-to-price ratio ([Boudoukh, Michaely, Richardson, and Roberts 2007](#)). The MSCI constructs their value index based on a combination of the book-to-market ratio, the forward earnings to price ratio, and the dividend yield.

Of course, a value stock has a high fundamental-to-price ratio. The contrast to a value stock is an “expensive” stock with a low fundamental-to-price ratio, i.e. a high market

price relative to fundamentals. Expensive stocks are often referred to as growth stocks since a high future growth rate may explain why the current market price is high relative to current fundamentals.

The value factor in a given period is typically defined as the rate of return in that period on a portfolio of cheap stocks minus the rate of return on a portfolio of expensive stocks, which can be seen as the amount of money you will earn over the period from a zero-cost strategy of being one dollar long the portfolio of cheap stocks and one dollar short the portfolio of expensive stocks. Obviously, you also have to decide on how many stocks to include in each portfolio, how the stocks are weighted, and how often you reset the portfolios (i.e. reallocate stocks to the portfolios). For example, you may choose (i) to include the 30% cheapest and 30% most expensive stocks, maybe only considering stocks with a market capitalization above some threshold to avoid investing in illiquid small stocks with relatively high transaction costs, (ii) to weigh the stocks according to market capitalization within each portfolio, and (iii) to reset the portfolios every month based on an updated price ratio for each stock.

Applying different specifications, numerous papers show a positive value premium in the long run, i.e. an abnormal (relative to the CAPM) positive return difference between the cheap portfolio and the expensive portfolio. [Novy-Marx and Velikov \(2016\)](#) find that with a typical implementation of a value-based trading strategy, the long-run value premium is only slightly reduced when taking transaction costs into account. However, it is also clear from the return statistics shown in Sections 6.6 and 10.2 that the value premium has been negative in some shorter samples, including the 2011-2020 decade.

As briefly mentioned in relation to the Fama-French 3-factor model, a value premium can be justified as value stocks seem to be more pro-cyclical than growth stocks. Most investors may prefer counter-cyclical stocks that do relatively well in bad times and, therefore, investors may require a higher average return on value stocks. This is a risk-based explanation. The value premium can also be explained by behavioral mistakes by some investors, e.g. investors who over-extrapolate recent company growth rates when valuing individual stocks and thus consistently undervalue low-growth value stocks and overvalue high-growth stocks. Many household investors may be attracted to glamour stocks such as the stocks of companies often mentioned in the media, companies that produce well-known consumer products, or high-tech companies for which superior future growth rates seem likely. In contrast, value stocks can be stocks of companies that appear old-fashioned and dull to many retail investors.

Momentum. As discussed in Section 6.6.4, [Jegadeesh and Titman \(1993\)](#), [Rouwenhorst \(1998\)](#), and [Asness, Moskowitz, and Pedersen \(2013\)](#) have provided evidence of short-run momentum in individual stock returns so that recent winners (stocks with high returns in recent months) tend to do well also in the near future, whereas recent losers tend to continue to do poorly. This suggests implementing a *winners-minus-losers strategy* in which you take a long position in a portfolio of recent winners and short a portfolio of recent losers. The momentum factor is then represented by WML, the return on the winners-minus-losers strategy, or more precisely the difference between the return on the winners' portfolio and the return on the losers' portfolio. Typically, the recent performance is measured over a 12 month period. Within each portfolio, the stocks can be weighted equally or according to the market capitalization or according to the magnitude of their recent performance. The two portfolios are typically rebalanced every month. Some researchers and analysts suggest limiting the strategy to the more extreme recent winners and losers.

The return statistics presented in Section 10.2 indicate that the WML strategy implemented on the US stock market is profitable in most periods, and the profits cannot be explained by the CAPM or the Fama-French 3-factor model. Similar findings have been reported for other stock markets and even for other asset classes. There are two issues, though:

1. *Transaction costs.* Several studies have shown that implementing the WML strategy in the broad stock market requires a lot of trading, especially in small stocks that are more costly to trade and may turn illiquid when you need to trade them. Empirical results presented by Hvidkjær (2006) indicate that the momentum effect is linked to liquidity issues. After controlling for the trading costs, maybe half of the apparent abnormal vanishes according to Novy-Marx and Velikov (2016).
2. *Crash risk.* Daniel and Moskowitz (2016) find that the winners-minus-losers strategy sometimes crashes—i.e. give very negative returns—following market declines and in volatile periods. For example, in the three-month period from March to May 2009, the portfolio of the 10% best-performing stocks over the recent year offered a return of 8%, whereas the portfolio of 10% worst-performing stocks through the past year gained 163%. The winners-minus-losers strategy thus gave a huge loss in that period.⁵

In spite of these issues, the momentum factor is by now well established and is often added to the Fama-French 3-factor model to produce the so-called Carhart 4-factor model named after Carhart (1997).

Is the momentum premium a compensation for risk? Many empirical studies using decades of data conclude that the momentum strategy provides an abnormally high return (positive alpha) relative to the CAPM and the Fama-French 3-factor model so, if a compensation for risk, the momentum premium must compensate for risks not captured by those models. Part of the premium could be explained by liquidity issues, cf. issue 1 above. Also, Avramov, Chordia, Jostova, and Philipov (2007) report that most of the momentum profits come from trades in stocks of firms with low credit quality, so that bankruptcy risk is relevant and might not be captured by the standard CAPM and Fama-French factors. Furthermore, part of the momentum premium can be a compensation for crash risk, cf. issue 2 above. In fact, several models and empirical studies find that many investors might require a substantial premium to hold portfolios with even a small probability of a large value drop. Finally, Liu and Zhang (2008) show that the momentum factor is closely related to the growth rate of industrial production, a key business cycle indicator. So, a risk-based justification of the momentum premium cannot be ruled out.

If the momentum premium is not compensation for risk, it presents a severe challenge to the efficient market hypothesis that is explained and discussed in Section 12.1 as the implementation of a momentum strategy requires knowledge of past prices only. The momentum premium can be generated by various “behavioral biases” among investors, i.e. patterns in investor behavior that deviate from the prescriptions of traditional economic models. One such bias is that some investors are conservative or not paying attention and therefore initially underreact to companies’ announcements of earnings, dividends, and other relevant news. Such news are only slowly being incorporated into prices and thus generate momentum. Seeing a continued increase in the price of stocks, some investors may

⁵Empirical studies by Gutierrez and Prinsky (2007), Blitz, Huij, and Martens (2011), Blitz, Hanauer, and Vidojevic (2020), and Lin (2020) suggest that momentum strategies perform better when based on *idiosyncratic momentum*, i.e. stocks’ recent residual returns relative to the CAPM or the three-factor Fama-French model.

decide to *jump on the bandwagon* and participate in the success, and this herding behavior can extend a momentum originating from the slow reaction to good news. Eventually the bandwagon investors sell out of their stocks and the stock price might fall, leading to a longer-term reversal. See, e.g., Barberis, Schleifer, and Vishny (1998), Daniel, Hirshleifer, and Subrahmanyam (1998), Hong and Stein (1999), Ottaviani and Sørensen (2015), and Azevedo (2023).

Another potential driver of momentum is the *disposition effect*. This is the tendency of investors to tend to (i) hold on to losing stocks in the hope their price recovers and (ii) sell winning stocks early to lock in the gains, cf. Shefrin and Statman (1985) and Odean (1998), among others. The disposition effect may lead to momentum in stock returns as explained by Grinblatt and Han (2005) and Frazzini (2006). Upon the announcement of bad news about a company, the stock price will initially drop only by little as some disposition investors might be reluctant to sell a falling stock but, if no good news arrive soon after, even some of these investors will sell out and the price will fall further towards the rational value. A losing stock will thus tend to lose further for a while. Conversely, when good news are announced, the stock price begins moving upwards. Some disposition investors have bought the stock at a lower price and will then sell the stocks to realize a gain, even if the stock price should rationally go up even more. Eventually, the relatively low stock price will attract other investors and the stock price continues upwards, generating the momentum pattern. See also Exercise 12.8.

Quality. The quality factor is a more recent animal in the factor zoo and comes in numerous varieties. Loosely speaking, high-quality stocks are stocks of well-managed companies with stable profits, strong balance sheets, consistent asset growth, and strong corporate governance. Obviously, quality can be specified in various ways. Novy-Marx (2011, 2013) showed that low operating leverage and high gross profitability (gross profits to book value of assets) are associated with higher average stock returns, also after controlling for exposure to Fama and French's three factors. Ball, Gerakos, Linnainmaa, and Nikolaev (2016) find that the ratio of operating profits to the lagged book value of assets predict high stock returns. The two newest Fama-French factors—operating profitability (Robust-Minus-Weak) and investment (Conservative-Minus-Aggressive)—are also quality indicators. The MSCI bases its quality index on the return on equity, the debt to equity, and the earnings variability.

Asness, Frazzini, and Pedersen (2019) define a quality measure for each stock as a (sophisticated) average of a profitability measure, a growth measure, and a safety measure. The profitability measure itself is based on several measures such as the gross profits over assets, the return on equity, and the return on assets. The growth measure is constructed from the changes in the individual profitability measures over the past five years. Finally, the safety measure is based on the stock's market beta, the leverage, the bankruptcy risk, and the volatility of the return on equity. High-quality stocks should have higher prices than low-quality stocks, but in the US stock market the difference is small, meaning that high-quality stocks seems underappreciated by investors. A Quality-Minus-Junk (QMJ) factor is defined as the return on a portfolio of high-quality stocks minus the return on a portfolio of low-quality (junk) stocks. Asness, Frazzini, and Pedersen find that the QMJ factor strategy in the US stock market provides a significant alpha relative to the Carhart four-factor model.

Note that the typical quality measures mentioned above are not related to the current stock price, so some high-quality stocks could be (too) expensive and some low-quality stocks could trade at an attractive low price. Some investors work with a price-adjusted

quality measure, sometimes referred to as *Quality at a Reasonable Price* (QARP).

Why might high-quality stocks be undervalued compared to low-quality stocks? Some investors seem to be attracted to stocks of companies that are innovative, apply new technologies, sell trendy consumer products, and are often mentioned in mainstream media. Maybe stocks for which some investors think they can make a quick fortune. High-quality stocks are often the opposite. They are issued by stable, maybe old-fashioned companies that rarely make media headlines and are not associated with new technologies or glamour. Like momentum, a strategy based on operating profitability has a high kurtosis and occasional large negative returns, which might be another reason for why the average profits are not traded away, see, e.g., [Arnott, Harvey, Kalesnik, and Linnainmaa \(2019\)](#).

Defensive. The defensive factor is also known as the low-risk factor. Some investors combine the defensive factor with the quality factor as high-quality stocks tend to be low-risk stocks. Many studies have shown that stocks with low risk provide surprisingly high average returns. Of course, the risk of a stock can be measured in many different ways.

According to the traditional, one-period CAPM, the risk of a stock should be measured by its market beta, and the average stock return should be linearly increasing in the market beta with a slope equal to the average excess return on the market portfolio. However, as explained in Section 10.2, low-beta assets provide higher and high-beta assets lower returns than predicted by the CAPM, a finding that can be due to borrowing constraints. This observation led [Frazzini and Pedersen \(2014\)](#) to suggest a *betting-against-beta* strategy with a long position in a portfolio of low-beta stocks and a short position in a portfolio of high-beta stocks. The corresponding BAB-factor is constructed as

$$\text{BAB}_{t+1} = \frac{1}{\beta_t^L} (r_{t+1}^L - r_f) - \frac{1}{\beta_t^H} (r_{t+1}^H - r_f),$$

where r_{t+1}^L is the return in period $t + 1$ on the portfolio of stocks having low market betas at time t , and r_{t+1}^H is the return in period $t + 1$ on the portfolio of stocks having high market betas at time t . Furthermore, β_t^L and β_t^H are the portfolio betas. The portfolios are constructed so that assets with more extreme betas have higher weights in the portfolios. [Frazzini and Pedersen \(2014\)](#) report high average returns on the betting-against-beta strategy. However, [Novy-Marx and Velikov \(2022\)](#) point out several non-standard procedures in the construction of the strategy which, among other things, imply that the strategy highly overweights very small stocks that are relatively illiquid and costly to trade. Accounting for transaction costs substantially reduces the apparent profitability of the strategy.

A related empirical observation is that *return volatility* matters for average returns as already suggested by the evidence presented in Section 6.6. [Ang, Hodrick, Xing, and Zhang \(2006\)](#) report that aggregate stock market volatility is priced in the cross section of U.S. stocks. Stocks that tend to provide high returns when market volatility is high have lower average returns, indicating that the higher the market volatility, the more the investors appreciate extra payments. If high market volatility represents bad investment opportunities, this is in line with Merton's intertemporal hedging story explained in Section 8.3. They further find that when computing a stock's idiosyncratic risk from the Fama-French three-factor model, stocks with high idiosyncratic risk earn very low returns, suggesting that the Fama-French model is insufficient. [Ang, Hodrick, Xing, and Zhang \(2009\)](#) confirm this pattern in other stock markets. However, the results of [Li, Sullivan, and Garcia-Feijóo \(2014\)](#) show that trying to exploit the apparent mispricing requires a

lot of trading so that profits are substantially reduced by transaction costs.

Other risk measures that have been shown to be associated with abnormal returns include the firm-specific risk component in the CAPM or the Fama-French 3-factor model, the volatility of company cash flows, and also measures of the turnover of the stock.

As explained above, high-risk stocks may be overvalued if risk-seeking investors are constrained from levering up low-risk stocks (or they simply do not want leverage) so that they can only obtain a high-risk position by buying high-risk stocks. Low-risk stocks are then undervalued in relative terms and are therefore generating abnormal high returns compared to the frictionless CAPM equilibrium. More generally, some investors might have lottery-seeking preferences that make them willing to pay more for stocks with a high upside potential even though the stocks also come with a large downside risk.

Size. [Banz \(1981\)](#) found a clear size effect in US stock market returns, where small stocks outperform large stocks, also after controlling for differences in the market beta. More recent studies indicate that the small-minus-big factor and the associated “size effect” have shrunk substantially or even disappeared ([Schwert 2003](#), [Goyal 2012](#)), which is also suggested by the statistics presented in Table 11.5.

In contrast to these observations, [Asness, Frazzini, Israel, Moskowitz, and Pedersen \(2018\)](#) find that, once you control for the quality factor, the size premium is resurrected, and that the premium is sizeable and stable over time and is robust to liquidity risk. On the other hand, [Blitz and Hanauer \(2021\)](#) argue that most of the premium comes from short positions in junk stocks, and that the quality-adjusted size premium is weak in other stock markets than the US market.

While a small-minus-big strategy requires some trading in small stocks with relatively high transactions costs, stocks are only added or removed from the small and big portfolios once a year, basically since the book equity value of each company is only updated at an annual frequency. Hence, the overall turnover and transaction costs associated with the SMB strategy are modest, see also [Novy-Marx and Velikov \(2022\)](#).

11.6.3 Factor diversification and factor timing

An investor should be able to improve the risk-return tradeoff of the overall stock market by tilting the portfolio according to the factors discussed in the previous subsection, for example by overweighing value stocks, recent winners, high-quality stocks, defensive stocks, and maybe small stocks and, of course, underweighing the counterparts. Here, it is important to note that the factors are far from perfectly correlated, so *factor diversification* becomes effective. By simultaneously applying several factors, an investor can reduce the overall portfolio risk, and thus improve the risk-return tradeoff even further. For example, [Asness, Moskowitz, and Pedersen \(2013\)](#) show that value and momentum are negatively correlated so a combination of the corresponding strategies seems very attractive. Moreover, the quality factor is somewhat negatively correlated with value and only slightly positively correlated with momentum, see, e.g., [Aghassi, Asness, Fattouche, and Moskowitz \(2023\)](#).

[Arnott, Harvey, Kalesnik, and Linnainmaa \(2019\)](#) argue that factor diversification is not as powerful as could be expected by looking at the long-run factor correlations. They show in recent US data that large drawdowns of individual factors often happen at the same time. Factor correlations are likely varying over time and may spike in periods of poor performance, which means that diversification is weaker exactly when it would be most valuable. More autocorrelation in returns can exacerbate and prolong periods of poor performance. Of course, diversification does not mean that poor outcomes become

impossible when forming a portfolio. Diversification reduces the risk, it does not eliminate the risk. Simultaneous poor performance of several factors is possible even if the factors are not highly correlated in the long run. Hence, [Aghassi, Asness, Fattouche, and Moskowitz \(2023\)](#) defend the view that factor diversification works for long-run investments.

The typical statistical justification of a factor involves a data set of 50 years or more, which means data collected across various economic scenarios and political and macroeconomic environments. A given factor could be more relevant in some scenarios and environments than in others. If we can identify in what circumstances the individual factors are most efficient, we can engage in *factor timing* strategies. Again, if you select the data carefully and look long enough at the data, you can probably detect some differences in factor premiums across states defined by booms versus recessions or by some other observable indicators of the economic outlook. However, as discussed by [Aghassi, Asness, Fattouche, and Moskowitz \(2023\)](#), the main factors seem to work across different economic conditions. Factor timing seems to be very difficult to implement successfully, see [Asness, Chandra, Ilmanen, and Israel \(2017\)](#) and [Ilmanen, Israel, Moskowitz, Thapar, and Lee \(2021\)](#).

11.7 Portfolio choice with tradeable factors

In most modern factor models, the factors are either returns on specific portfolios (e.g. the stock market index) or return differences (e.g. the return on a value portfolio less the return on a growth portfolio). In principle, investors can then trade the factors. One way to do that is, of course, by the investor herself setting up and rebalancing the relevant portfolios, which would clearly be cumbersome, involve trading in many individual stocks, and cost a lot of time and trading fees. Also note that this may involve shorting specific stocks, which some investors must or prefer to refrain from, and if shorting is done, it is often particularly expensive in terms of costs. Still, some hedge funds and active mutual funds may pursue such an approach.

Nowadays some factor ETFs are available. For example, at the time of writing (March 2023), iShares by BlackRock offers single factor ETFs pertaining to value, quality, momentum, size, and low volatility in the US stock market. Each ETF has an expense ratio of only 0.15%. These ETFs track the “long leg” of the long-short strategy associated with each factor. For instance, the value factor ETF tracks the returns on a portfolio of large- and mid-cap value stocks and thus does not track a high-minus-low return difference such as the Fama-French HML factor. Fidelity also offers factor ETFs across both geographies and asset classes. Both iShares and Fidelity also offer a few multi-factor ETFs. Through the ETFs, even household investors can participate in factor investing at a modest cost.

The next subsection determines the optimal portfolio of a mean-variance investor in the case where all factors are returns of tradeable portfolios, maybe via ETFs as just discussed. Subsequently, Section 11.7.2 considers optimal portfolios in the case where the factors are the market return plus a number of long-short return differences that can be implemented by long and short positions in tradeable portfolios. Of course, you may believe in a factor model where the factors are not directly tradeable. Then the optimal portfolio for an investor can be found using the mean-variance approach based on a large number of individual assets, where the relevant inputs are derived exploiting the implications of the factor model for expected returns, variances, and covariances.

11.7.1 If all factors are returns

Suppose that you believe in the equilibrium version of a factor model where all factors are returns on specific traded portfolios. What is then the best portfolio of risky assets, i.e. which portfolio maximizes the Sharpe ratio? In Section 11.2.3 we showed that within the equilibrium version of the Single-Index Model, the largest Sharpe ratio is obtained by investing only in the market portfolio. Combining the market portfolio with any individual risky assets will only lower the Sharpe ratio due to the additional idiosyncratic risk. For the equilibrium multi-factor model with traded factor portfolios, the corresponding conclusion is that the largest Sharpe ratio is achieved by the tangency portfolio formed from these factor portfolios. Adding individual assets will only lower the Sharpe ratio, again due to the idiosyncratic risk. We summarize this in the next theorem and provide a formal proof. We denote the vector of expected returns on the factor portfolios and the variance-covariance matrix of their returns by

$$\boldsymbol{\mu}_F = \begin{pmatrix} E[r_{F1}] \\ E[r_{F2}] \\ \vdots \\ E[r_{FK}] \end{pmatrix}, \quad \underline{\Sigma}_F = \begin{pmatrix} \text{Var}[r_{F1}] & \text{Cov}[r_{F1}, r_{F2}] & \dots & \text{Cov}[r_{F1}, r_{FK}] \\ \text{Cov}[r_{F2}, r_{F1}] & \text{Var}[r_{F2}] & \dots & \text{Cov}[r_{F2}, r_{FK}] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[r_{FK}, r_{F1}] & \text{Cov}[r_{FK}, r_{F2}] & \dots & \text{Var}[r_{FK}] \end{pmatrix},$$

where r_{Fk} is the rate of return on the k 'th factor portfolio.

Theorem 11.6

Suppose returns on individual assets satisfy, with zero alphas, a K -factor model where all factors are returns on specific traded portfolios, i.e.

$$r_i - r_f = \sum_{k=1}^K \beta_{ik} (r_{Fk} - r_f) + \varepsilon_i \quad (11.52)$$

with $E[\varepsilon_i] = \text{Cov}[\varepsilon_i, \varepsilon_j] = \text{Cov}[\varepsilon_i, F_k] = 0$ for all assets i, j and factors k . Then the portfolio maximizing the Sharpe ratio is the tangency portfolio formed from the factor portfolios,

$$\boldsymbol{\pi}_{\tan}^F = \frac{\underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1})}{\mathbf{1} \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1})}. \quad (11.53)$$

Proof

The expression (11.53) for the tangency portfolio formed from the factor portfolios follows from Theorem 7.6. We have to show that we cannot obtain a larger Sharpe ratio by including any individual assets. Suppose you combine the factor portfolios and a single risky asset i . The rate of return on asset i is

$$r_i = r_f + \sum_{k=1}^K \beta_{ik} (r_{Fk} - r_f) + \varepsilon_i = r_f + \boldsymbol{\beta}_i \cdot (\boldsymbol{r}_F - r_f \mathbf{1}) + \varepsilon_i,$$

so the expected excess return on asset i is

$$\mu_i - r_f = E[r_i] - r_f = \beta_i \cdot (\mu_F - r_f \mathbf{1}).$$

The extended vector of expected excess returns and the extended variance-covariance matrix can be written as

$$\mu - r_f \mathbf{1} = \begin{pmatrix} \mu_F - r_f \mathbf{1} \\ \beta_i^\top (\mu_F - r_f \mathbf{1}) \end{pmatrix}, \quad \underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_F & \underline{\Sigma}_F \beta_i \\ \beta_i^\top \underline{\Sigma}_F & \beta_i^\top \underline{\Sigma}_F \beta_i + \sigma_{\varepsilon,i}^2 \end{pmatrix}. \quad (11.54)$$

Note that the $(K+1) \times (K+1)$ matrix $\underline{\Sigma}$ has a block structure where the upper-left block is the $K \times K$ factor variance-covariance matrix and the lower-right block is just a scalar, namely the variance of asset i . The off-diagonal term is the vector of asset i 's covariances with the K factors, either as the column vector $\underline{\Sigma}_F \beta_i$ or the row vector $\beta_i^\top \underline{\Sigma}_F$. It can be shown that

$$\underline{\Sigma}^{-1} = \begin{pmatrix} \underline{\Sigma}_F^{-1} + \frac{1}{\sigma_{\varepsilon,i}^2} \beta_i \beta_i^\top & -\frac{1}{\sigma_{\varepsilon,i}^2} \beta_i \\ -\frac{1}{\sigma_{\varepsilon,i}^2} \beta_i^\top & \frac{1}{\sigma_{\varepsilon,i}^2} \end{pmatrix}, \quad (11.55)$$

and hence

$$\underline{\Sigma}^{-1} (\mu - r_f \mathbf{1}) = \begin{pmatrix} \underline{\Sigma}_F^{-1} + \frac{1}{\sigma_{\varepsilon,i}^2} \beta_i \beta_i^\top & -\frac{1}{\sigma_{\varepsilon,i}^2} \beta_i \\ -\frac{1}{\sigma_{\varepsilon,i}^2} \beta_i^\top & \frac{1}{\sigma_{\varepsilon,i}^2} \end{pmatrix} \begin{pmatrix} \mu_F - r_f \mathbf{1} \\ \beta_i^\top (\mu_F - r_f \mathbf{1}) \end{pmatrix} = \begin{pmatrix} \underline{\Sigma}_F^{-1} (\mu_F - r_f \mathbf{1}) \\ 0 \end{pmatrix}.$$

The sum of the elements in this vector is $\mathbf{1} \cdot \underline{\Sigma}^{-1} (\mu - r_f \mathbf{1}) = \mathbf{1} \cdot \underline{\Sigma}_F^{-1} (\mu_F - r_f \mathbf{1})$, so the extended tangency portfolio is indeed

$$\pi_{\text{tan}} = \begin{pmatrix} \pi_{\text{tan}}^F \\ 0 \end{pmatrix}.$$

In particular, asset i does not have a role in the tangency portfolio so the maximal Sharpe ratio is still obtained by the best combination of the factor portfolios.

In most implementations, the first factor is the return on the market portfolio. The tangency portfolio is then a combination of the market portfolio and the other factor portfolios, which of course implies that the optimal portfolio differs from the market portfolio due to the assumed risk premium on the other factor portfolios. From Theorem 7.6 we know that the squared Sharpe ratio of the tangency portfolio in this case is $(\mu_F - r_f \mathbf{1}) \cdot \underline{\Sigma}_F^{-1} (\mu_F - r_f \mathbf{1})$, which depends on the risk premiums of the factor portfolios and the factor variance-covariance matrix. In the special case of uncorrelated factors, the squared Sharpe ratio of the tangency portfolio equals the sum of the squared Sharpe ratios of all factor portfolios, but this simple relation is not valid in the typical case with correlated factor portfolios as already indicated by Eq. (4.17). If the non-market factor portfolios have positive Sharpe ratios, they will typically have positive weights in the tangency portfolio so that the weight of the market portfolio is less than one. Assets with positive weights in one or more of the factor portfolios will typically end up with a larger weight in the optimal portfolio than their market weights. Assets with zero or even negative weights in one or more of the factor portfolios will typically end up with a smaller

weight in the optimal portfolio than their market weights.

As the next example illustrates, the combination of factor portfolios generating the largest Sharpe ratio can sometimes involve rather extreme positions.

Example 11.3

Let us consider two of the characteristics underlying the Fama-French 5-factor models, namely the operating profitability (OP) and the equity book-to-market (BM) ratio. We use monthly U.S. data from the homepage of Professor Kenneth French (downloaded December 10, 2021) covering the period from July 1963 (the earliest data available for portfolios sorted on OP) to October 2021. We focus on the overall market portfolio, a robust portfolio of stocks of the companies with the 30% highest OP, a weak portfolio of stocks with the 30% lowest OP, a value portfolio with the 30% highest BM, and a growth portfolio with the 30% lowest BM. To be precise, as in the construction of the portfolios shown in Table 11.3, the breakpoints (i.e. the 30th and 70th percentiles) are based only on the NYSE stocks, but the portfolios also include stocks listed on the AMEX and NASDAQ exchanges. The portfolios are all value-weighted.

Table 11.7 lists average monthly excess returns, sample standard deviations, Sharpe ratios, and pairwise sample correlations for the five portfolios. The robust and value portfolios seem very attractive, whereas the weak portfolio has a poor record. Also note the large correlations between the five portfolios. We use these statistics for portfolio formation.

Under the column heading ‘Unconstr’, Table 11.8 shows the optimal unconstrained combination of the five factor portfolios. This combination involves rather extreme positions including a considerable short position in the market portfolio and a large position in the robust portfolio. As discussed and illustrated in Example 4.5, such extreme positions are common when highly correlated assets or portfolios are combined with the purpose of maximizing the Sharpe ratio. The 0.243 Sharpe ratio of this combination is significantly larger than that of the market portfolio or any of the other four factor portfolios. The next three columns show the combinations of the five factor portfolios that maximize the Sharpe ratio when the weights on the individual factor portfolios are restricted to the interval $[-1, 2]$, $[-1, 1]$, or $[0, 1]$. Obviously, such restrictions reduce the Sharpe ratio, in particular in the last case where short positions are prohibited and the optimal combination is a roughly balanced mix of the robust portfolio and the value portfolio.

The right part of Table 11.8 depicts optimal portfolios when only a subset of the factor portfolios are included in the analysis. These optimal portfolios also involve short and often rather extreme positions. In all these cases, the attainable Sharpe ratio is 0.188 or lower and thus far below the 0.243 Sharpe ratio obtained when all factor portfolios are combined optimally. It pays off to simultaneously tilt your portfolio along the robust-weak dimension and the value-growth dimension.

The above analysis assumes that all individual assets have zero alphas relative to the stated multi-factor model. If some individual assets have a non-zero alpha, they would allow the investor to obtain a larger Sharpe ratio. Intuitively, you would want to put positive weights on the assets with positive alpha and negative weights on the assets with negative alpha. Note that these non-zero weights are relative to the weight of the asset in the market portfolio and any weight of the asset in the factor portfolios. With non-zero weights on individual assets, your portfolio becomes less well diversified due to the

Portfolio	Exp excess	Std dev	Sharpe	Correlations				
				M	R	W	V	G
Market	0.589	4.452	0.132	1.000	0.976	0.952	0.886	0.975
Robust	0.690	4.410	0.156	0.976	1.000	0.877	0.829	0.981
Weak	0.457	5.249	0.087	0.952	0.877	1.000	0.883	0.900
Value	0.820	5.031	0.163	0.886	0.829	0.883	1.000	0.792
Growth	0.591	4.672	0.127	0.975	0.981	0.900	0.792	1.000

Table 11.7: Inputs to example with factor portfolios.

The table shows summary statistics based on monthly excess returns from July 1963 to October 2021. The data were downloaded on December 10, 2021, from the homepage of Professor Kenneth French. See the text for a short description of the construction of the portfolios.

	All factor portfolios				Unbounded, but only selected factors					
	Unconstr	[-1,2]	[-1,1]	[0,1]	MRW	MR	MW	MVG	MV	MG
Market	-2.569	-1.000	-0.049	0.000	0.366	-2.797	3.004	-6.162	-0.432	1.360
Robust	3.216	2.000	1.000	0.422	1.862	3.797	no	no	no	no
Weak	-1.214	-1.000	-1.000	0.000	-1.228	no	-2.004	no	no	no
Value	1.714	1.273	1.000	0.578	no	no	no	2.905	1.432	no
Growth	-0.148	-0.273	0.049	0.000	no	no	no	4.257	no	-0.360
Excess	1.471	1.218	1.054	0.765	0.939	0.974	0.853	1.272	0.920	0.588
Std dev	6.063	5.133	4.675	4.570	4.991	5.348	4.657	6.874	5.571	4.430
Sharpe	0.243	0.237	0.225	0.167	0.188	0.182	0.183	0.185	0.165	0.133

Table 11.8: Optimal combinations of factor portfolios.

Each column of the table shows the combination of factor portfolios that delivers the largest possible Sharpe ratio under given restrictions on the portfolio. In the left part of the table, the investor can invest in all five factor portfolios, either without constraints or with all weights being contained to the interval shown. In the right part of the table, the investor can use only the subset of factor portfolios indicated by the column headings, but always without any bounds on the portfolio weights. The summary statistics from Table 11.7 are used as inputs.

exposure to the idiosyncratic return risk of the mispriced assets. The tradeoff between the asset's alpha and the idiosyncratic risk is therefore important. This tradeoff is captured by the asset's information ratio,

$$\text{IR}_i = \frac{\alpha_i}{\text{Std}[\varepsilon_i]}, \quad (11.56)$$

where ε_i is the non-systematic return component of asset i . The Treynor-Black model presented in Section 13.1 gives an insightful and straightforward way of implementing the mean-variance method when some assets have non-zero alphas relative to a factor model.

11.7.2 Market factor and long-short factors

Suppose you have identified a long-short strategy with an expected return difference that seems significant. How should you exploit that as a risk-averse investor? In the perspective of the mean-variance analysis the first problem is to find the portfolio maximizing the Sharpe ratio of risky assets and then in the second step to find the combination of that portfolio and the riskfree asset which is optimal given the investor's risk aversion. The solution to the second problem is no different than in the standard mean-variance model, so let us focus on the first problem. Note that any position in the long-short strategy is a zero net investment. The strategy consists of a long position in one portfolio and a short position in another portfolio and the two positions are of identical magnitudes.

First consider how to optimally combine the market portfolio with the long-short strategy. This means a portfolio weight of 1 in the market portfolio, some weight w in the long portfolio in the long-short strategy, and the weight $-w$ in the shorted portfolio. If we let r_ℓ and r_s denote the rates of return on the long and the shorted portfolio, then the rate of return on the combined position is

$$r(w) = r_m + w(r_\ell - r_s).$$

Let $\mu_m = \text{E}[r_m]$ denote the expected market return and $\mu_{\ell s} = \text{E}[r_\ell - r_s]$ the risk premium associated with the long-short strategy. Then the expected return on the combined position is

$$\text{E}[r(w)] = \mu_m + w\mu_{\ell s}.$$

If we let $\sigma_m = \text{Std}[r_m]$, $\sigma_{\ell s} = \text{Std}[r_\ell - r_s]$, and $\sigma_{m,\ell s} = \text{Cov}[r_m, r_\ell - r_s]$, then the variance of the return on the combined position is

$$\text{Var}[r(w)] = \sigma_m^2 + w^2\sigma_{\ell s}^2 + 2w\sigma_{m,\ell s}.$$

The Sharpe ratio is thus

$$\text{SR}(w) = \frac{\text{E}[r(w)] - r_f}{\text{Std}[r(w)]} = \frac{\mu_m - r_f + w\mu_{\ell s}}{\sqrt{\sigma_m^2 + w^2\sigma_{\ell s}^2 + 2w\sigma_{m,\ell s}}}.$$

It can be shown that the Sharpe ratio is maximized for $w = w^*$ which is defined by

$$w^* = \frac{\psi\mu_{\ell s} - \sigma_{m,\ell s}}{\sigma_{\ell s}^2}, \quad \psi = \frac{\sigma_{\ell s}^2\sigma_m^2 - \sigma_{m,\ell s}^2}{\sigma_{\ell s}^2(\mu_m - r_f) - \sigma_{m,\ell s}\mu_{\ell s}}. \quad (11.57)$$

Here ψ is the Greek letter 'psi'. Including individual assets with zero alphas relative to the factor model will lead to a lower Sharpe ratio due to the induced non-systematic risk which is not compensated by higher expected returns.

How does the above analysis generalize to the case with multiple return differences as factors? The factor model then assumes that the rate of return on each individual asset i satisfies

$$r_i - r_f = \beta_{im}(r_m - r_f) + \sum_{k=2}^K \beta_{ik}(r_{\ell k} - r_{sk}) + \varepsilon_i. \quad (11.58)$$

When combining only the market portfolio and the long-short strategies corresponding to factors $2, \dots, K$, the market portfolio must have a weight of one as the investments in the long-short strategies have zero net costs. Let $\mathbf{w} = (w_2, \dots, w_K)^\top$ denote the vector of “weights” in the long-short strategies where, for example, w_2 means that for every dollar invested in the market portfolio, w_2 dollars are invested in the long portfolio in factor 2, whereas the short portfolio is shorted for a value of w_2 dollars. Let $\mathbf{r}_{\ell s} = (r_{\ell 2} - r_{s2}, \dots, r_{\ell K} - r_{sK})^\top$ denote the vector of the $K - 1$ long-short return differences, and let $\underline{\Sigma}_{\ell s} = \text{Var}[\mathbf{r}_{\ell s}]$ be its $(K - 1) \times (K - 1)$ variance-covariance matrix and $\boldsymbol{\mu}_{\ell s} = E[\mathbf{r}_{\ell s}]$ its expectation. Furthermore, let $\boldsymbol{\sigma}_{m,\ell s} = (\text{Cov}[r_m, r_{\ell 2} - r_{s2}], \dots, \text{Cov}[r_m, r_{\ell K} - r_{sK}])^\top$ denote the vector of covariances between the market return and the $K - 1$ long-short return differences. The optimal \mathbf{w} can then be stated as in the next theorem.

Theorem 11.7

Suppose returns on individual assets satisfy the K -factor model (11.58) where the first factor is the (excess) return on the market portfolio and the remaining factors are specific return differences. Then the maximum Sharpe ratio is obtained by investing all wealth in the market portfolio together with the fractions \mathbf{w}^* of wealth in the $K - 1$ long-short factors, where

$$\mathbf{w}^* = \underline{\Sigma}_{\ell s}^{-1} (\psi \boldsymbol{\mu}_{\ell s} - \boldsymbol{\sigma}_{m,\ell s}), \quad \psi = \frac{\sigma_m^2 - \boldsymbol{\sigma}_{m,\ell s} \cdot \underline{\Sigma}_{\ell s}^{-1} \boldsymbol{\sigma}_{m,\ell s}}{\mu_m - r_f - \boldsymbol{\sigma}_{m,\ell s} \cdot \underline{\Sigma}_{\ell s}^{-1} \boldsymbol{\mu}_{\ell s}}. \quad (11.59)$$

In the special case of a single long-short factor the expressions for \mathbf{w}^* and ψ in Eq. (11.59) simplify to the expressions in Eq. (11.57).

Proof

A portfolio fully invested in the market portfolio and with long-short positions given by \mathbf{w} has return

$$r(\mathbf{w}) = r_m + \mathbf{w} \cdot \mathbf{r}_{\ell s}$$

and thus a Sharpe ratio of

$$\text{SR}(\mathbf{w}) = \frac{\mu_m - r_f + \mathbf{w} \cdot \boldsymbol{\mu}_{\ell s}}{\sqrt{\sigma_m^2 + \mathbf{w} \cdot \underline{\Sigma}_{\ell s} \mathbf{w} + 2\mathbf{w} \cdot \boldsymbol{\sigma}_{m,\ell s}}}.$$

The first-order condition $\text{SR}'(\mathbf{w}) = \mathbf{0}$ is satisfied if and only if

$$(\sigma_m^2 + \mathbf{w} \cdot \underline{\Sigma}_{\ell s} \mathbf{w} + 2\mathbf{w} \cdot \boldsymbol{\sigma}_{m,\ell s}) \boldsymbol{\mu}_{\ell s} = (\mu_m - r_f + \mathbf{w} \cdot \boldsymbol{\mu}_{\ell s}) (\underline{\Sigma}_{\ell s} \mathbf{w} + \boldsymbol{\sigma}_{m,\ell s}).$$

It can be verified that this equation holds when \mathbf{w} is given by (11.59).

We skip the formal proof of the fact that the Sharpe ratio falls if individual assets are included in the portfolio beyond their appearance in the market portfolio and the long-short portfolios.

Example 11.3 above looked at optimal portfolios when the investor trades in the market portfolio as well as the robust, weak, value, and growth portfolios. The next example assumes that the investor trades in a long-short robust-minus-weak (RmW) strategy instead of separately in a robust and a weak portfolio and, similarly, trades in a long-short value-minus-growth (VmG) strategy instead of separately in a value and a growth portfolio. Note that the definition of the RmW used here differs from the one used by Fama and French (2015). The VmG also differs somewhat from the HML (high-minus-low book-to-market) they define, cf. also Eq. (11.49).

Example 11.4

See Example 11.3 for a description of the data used. We form a robust-minus-weak (RmW) long-short strategy having a long position in the robust portfolio (stocks of the companies with the 30% highest OP) and a short position in the weak portfolio (stocks with the 30% lowest OP). Similarly, we form a value-minus-growth (VmG) strategy with a long position in the value portfolio (the 30% highest BM) and a short position in the growth portfolio (the 30% lowest BM).

The left part of Table 11.9 shows summary statistics for the monthly return differences. Note the negative correlation that the RmW and VmG strategies have with the market return, and the correlation between RmW and VmG is also negative.

The right part of the table shows Sharpe ratio maximizing combinations of the market portfolio and either the RmW or the VmG strategy or both. With access to both RmW and VmG, the optimal weights are 1.874 on RmW and 1.157 on VmG. For every dollar invested in the market portfolio, you should take a long position worth \$1.874 in the robust portfolio and a short position worth \$1.874 in the weak portfolio together with a long position worth \$1.157 in the value portfolio and a short position worth \$1.157 in the growth portfolio. This generates the maximum Sharpe ratio of 0.237. This is slightly smaller than the 0.243 Sharpe ratio attainable when trading separately in the robust, weak, value, and growth portfolios, cf. Example 11.3. By trading in the long-short strategies, you are effectively restricting yourself to have the same weight (with opposite signs, of course) in the robust portfolio as in the weak portfolio and similarly for the value and growth portfolios. If we exclude the VmG strategy, we see that the RmW strategy facilitates a Sharpe ratio of 0.188 which, with the digits shown here, is identical to the Sharpe ratio attainable with separate trading in the robust and the weak portfolios. In contrast, with only the VmG strategy, the maximum Sharpe ratio of 0.153 is notably smaller than the 0.185 attainable when trading value and growth separately.

	Expect	Std dev	Correlations			Portfolio weights		
			Mkt-rf	RmW	HmL	All	w/ RmW	w/ VmG
Mkt-rf	0.589	4.452	1.000	-0.274	-0.032	1.000	1.000	1.000
RmW	0.233	2.529	-0.274	1.000	-0.385	1.874	1.436	no
VmG	0.229	3.145	-0.032	-0.385	1.000	1.157	no	0.810
Excess						1.291	0.924	0.774
Std dev						5.445	4.914	5.060
Sharpe						0.237	0.188	0.153

Table 11.9: Inputs and optimal portfolios with long-short factors.

The summary statistics in the left part of the table are based on monthly returns from July 1963 to October 2021 downloaded on December 10, 2021, from the homepage of Professor Kenneth French. RmW refers to the robust-minus-weak strategy with a long position in the robust portfolio and a corresponding short position in the weak portfolio. Similarly, VmG refers to the value-minus-growth strategy with a long position in the value portfolio and a corresponding short position in the growth portfolio. See the text for details. The right part of the table shows Sharpe ratio maximizing combinations of the market portfolio with the RmW or VmG strategies or both.

11.7.3 More on portfolio choice

Let us take a step back and think about the portfolio decision from a more philosophical point of view. As discussed in Section 10.1.4, the average investor must in any case hold the market portfolio so, for an average investor, the CAPM conclusion of investing in the market portfolio remains correct. In a factor sense, you should have zero exposure to other factors than the market return, after adjusting for the market risk of those factors. But if you are different from the average investor, the weights you should apply to some assets are different from the assets' market weights. Or, in other words, non-zero weights to other factors can be optimal. Of course, markets have to clear, so if you overweight some assets, other investors must underweight the same assets. After all, if you are above average along some dimension, some investors have to be below average.

You might differ from the average investor on several counts. Maybe you have a longer time horizon than the average investor. Then you might want to overweight long-term bonds that provide protection in the long run relative to short-term bonds. Short-horizon investors may want to do the opposite. Maybe long-horizon investors should consider investing more in assets that tend to revert in the long run compared to assets that do not. Long-horizon investors can better accept liquidity risk than short-horizon investors who might have to realize their investments at short notice. Hence, long-horizon investors may want to overweight relatively illiquid assets and gain the associated illiquidity premium.

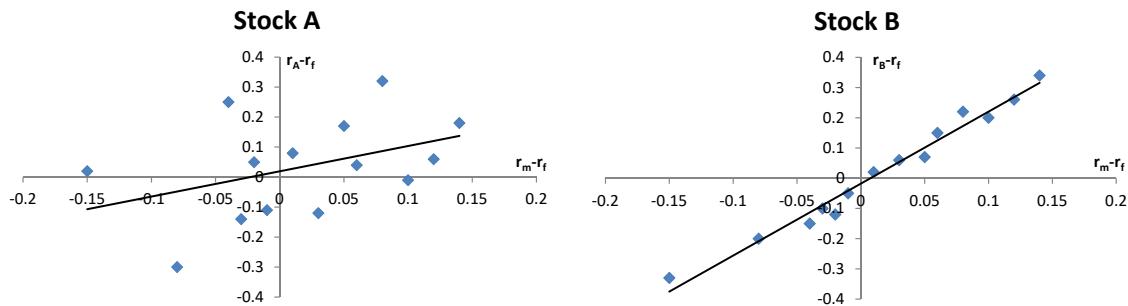
You should also think about your “loss capacity,” i.e., how much you can bear to lose in bad times. As discussed in the following section, many factor-related strategies perform extremely badly once in a while. If you can stand such losses, you can allow yourself to invest in those strategies, which means taking positive exposure to these factors, and thus overweighting, for example, value stocks or small stocks. Other investors who are more risk averse or more capital constrained during bad times may choose to have negative exposure to the same factors and thus underweight such stocks.

Implementing the HML strategy or similar long-short strategies might seem complicated and costly in terms of trading costs. But nowadays you can easily trade ETFs mimicking an index of value stocks or an index of growth stocks, and by doing so you can implement some form of HML strategy. Other ETFs and mutual funds track quantities related to

other factors. For more on “factor investing” the reader is referred to the book by Ang (2014).

11.8 Exercises

Exercise 11.1. The following graphs show the results of the Single-Index regression for two different stocks, stock A and stock B.



Which of the two stocks has the higher ...

- (a) systematic risk?
- (b) asset-specific risk?
- (c) R^2 ?
- (d) α ?
- (e) correlation with the market?

Exercise 11.2. Use the Excel file `Exercise11_2.data.xlsx` from the supplementary material to these lecture notes. The file contains monthly adjusted closing prices from December 2015 to December 2020 for Alphabet (Google) stocks and the SPY ETF tracking the S&P 500 stock index, as well as the 1-month interest rate on U.S. government bonds.

- (a) Compute for each month the excess return on Google stocks and the excess return on the S&P 500 index.
- (b) Construct a scatter diagram in Excel in which you plot all the monthly pairs of the excess stock index return (horizontal axis) and the excess Google return (vertical axis). Add a trendline and the associated equation to the diagram.
- (c) Use Excel to estimate the single-index model for Google stocks.
- (d) What is the estimate of Google’s beta, β_G ? How do you interpret that value? What is the 95% confidence interval for beta? What is the associated p -value and how do you interpret that?
- (e) What is the estimate of Google’s alpha, α_G ? How do you interpret that value? What is the 95% confidence interval for alpha? What is the associated p -value and how do you interpret that?
- (f) What is the R^2 of the regression? Interpret the number.
- (g) What is the standard deviation of the residuals (firm-specific return component)? Interpret the number.
- (h) What is the systematic risk of Google?

Exercise 11.3. Andrea works for a hedge fund and can invest in the following stocks:

Stock	β_i	$E[r_i]$
A	1.4	14%
B	1.0	10%
C	0	5%

Assume the following for the market portfolio: $E[r_m] = 10\%$. Alpha and beta values in this question are based on the Single-Index Model. You can assume that $\alpha_C = 0$.

- Andrea's current portfolio ("portfolio 1") consists entirely of stock A. Show how he could invest in stocks A, B and C to create a market-neutral (i.e., $\beta = 0$) portfolio with the same alpha as the current portfolio. Determine the weights and expected return on this portfolio ("portfolio 2").
- Is it possible to create a market-neutral portfolio (by investing in A, B and C) with a higher alpha than that of portfolio 2? Show how/why not.

Exercise 11.4. The hedge fund *FatReturns* can invest in the following assets:

Asset	β_i	$E[r_i]$	$Std[r_i]$
A	1.4	14%	30%
B	1.0	9%	20%
C	0	5%	0%

The market portfolio is characterized by $E[r_M] = 10\%$ and $Std[r_M] = 15\%$, and the correlation between assets A and B is 0.97. Alpha and beta values are based on the Single-Index Model.

- Find the alpha values for assets A and B.
- In any market-neutral portfolio (that is, $\beta_p = 0$) of the three assets, what must the relation between the weights of assets A and B be?
- FatReturns* wishes to offer a market neutral portfolio to its clients with a standard deviation of 5%. Propose a portfolio with $\beta_p = 0$, $\alpha_p > 0$, and $\sigma_p = 5\%$.
- Compute the Sharpe ratio and the information ratio of the portfolio proposed in question (c).
- The Single-Index Model assumes that the covariance between individual assets can be attributed to the common exposure to the market factor. Show whether this assumption holds for assets A and B.

Exercise 11.5. You are an advisor to a pension fund that considers investing a small portion of its fund with either portfolio manager AB or manager CD. The two managers have provided the following performance information based on their portfolio return over the past 5 years:

Manager	Alpha (% per year)	Beta	Sharpe ratio
AB	2.2	0.5	1.2
CD	4	2	1.2

The average return on the market was 10% per year and the riskfree rate was 5% per year. You can assume that the pension fund can both lend and borrow at the riskfree rate.

The pension fund wants to invest with the CD manager, because this manager has the higher alpha. However, would it be possible to invest in a portfolio consisting of the AB portfolio and the riskfree asset, and achieve a higher alpha but identical beta to investing in the CD portfolio? (hint: yes) Find the weights, alpha, Sharpe ratio and the expected return on this portfolio. Illustrate the different investment opportunities in an SML graph.

Exercise 11.6. You believe that the return r_i on any stock i satisfies the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

You know that the riskfree rate r_f is 2% and the return r_m on the stock market index over the next year has an expectation of 10% and a standard deviation of 20%. You have the following information about the returns over the following year on two stocks, denoted by X and Y :

Stock	β_i	$E[r_i]$	$\sigma[r_i]$
X	2.0	16%	40%
Y	0.5	8%	20%

In addition, you have estimated the correlation between the returns on X and Y to be 0.5.

- Find the alpha values for the two stocks X and Y.
- Suppose you want to set up a market-neutral portfolio consisting of stocks X and Y as well as the riskfree asset. What must be the relation between the weights of the two stocks in such a portfolio? What can you say about the alpha of such a portfolio (the alpha of the riskfree asset is zero)? Under what condition on the portfolio weights will the alpha of the portfolio be positive?
- Propose a portfolio of the two stocks and the riskfree asset so that the portfolio is market-neutral, has a positive alpha, and has a return standard deviation of 20%. What is the alpha of the portfolio? What is the expected return on the portfolio?
- The Single-Index Model assumes that all covariance between individual assets can be attributed to the common exposure to the market factor. Show whether this assumption holds for stocks X and Y.
- Make a sketch of a diagram with betas along the horizontal axis and expected returns along the vertical axis. Draw the security market line in the diagram. Indicate the location of the stocks X and Y, the market portfolio, the riskfree asset, and the portfolio constructed above.

Exercise 11.7. Suppose that returns can be described by the one-factor model

$$r_i = 2\% + \beta_i F + \varepsilon_i,$$

where the expected value of the factor is $E[F] = 10\%$. Consider the three well-diversified portfolios A, B, and C:

Portfolio	Factor beta	Expected return
A	0.7	9%
B	1.3	11%
C	1.5	17%

- Are the expected returns on the three portfolios in line with the factor model?
- If one of the portfolios is mispriced according to the factor model, how can you construct a profitable factor-neutral combination of the three portfolios?
- If many investors follow the strategy suggested by your answer to the previous question, what would happen to the prices and expected returns?
- Discuss the role of idiosyncratic/residual risk in relation to this exercise!

Exercise 11.8. This problem considers the Single-Index Model for the returns on the stocks of the Hewlett-Packard Company (abbreviated HP in the following). Monthly returns on HP stocks were calculated based on adjusted closing prices downloaded from Yahoo Finance. Monthly values for the riskfree rate and for the return on the stock market portfolio were downloaded from the homepage of Professor Kenneth French. The data was downloaded on May 8, 2024. The analysis is based on 60 monthly observations from January 2019 to December 2023. Here are some summary statistics for the excess returns of HP and the market and a graph of the evolution in the price of HP stocks and the value of the entire stock market (indexed at 10 in January 2019) over the period.

	HP-RF	MKT-RF
Mean	0.98	1.18
Median	0.28	2.26
Std dev	9.36	5.56
Variance	87.53	30.91
Kurtosis	0.82	0.04
Skewness	0.29	-0.32
Minimum	-24.21	-13.39
Maximum	27.77	13.65
Count	60	60



Note that excess returns are represented in percent throughout this problem so that, for example, the mean of 0.98 for HP-RF should be interpreted as an average excess return of 0.98% per month. The standard deviation of monthly HP excess return is 9.36%, etc. Below you can see the output from a linear regression in Excel of the excess HP return on the excess market return.

- (a) Find and interpret the beta of HP. Is the beta of HP significantly different from one?
- (b) Find and interpret the alpha of HP. Is the alpha of HP significantly different from zero?
- (c) What is the total variance of the excess returns on HP? How large is the systematic component? How large is the unsystematic (or firm-specific) component?
- (d) What can you conclude from the analysis about the attractiveness of investing in HP stocks?
- (e) A similar regression for IBM stocks leads to the beta-estimate $\beta_{IBM} = 0.6866$. The sample covariance between HP and IBM returns over the 60 months is 36.84. Is this consistent with the assumptions of the Single-Index Model? Why or why not?

<i>Regression Statistics</i>	
Multiple R	0.7244
R Square	0.5247
Adjusted R Square	0.5165
Standard Error	6.5053
Observations	60

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Signif. F</i>
Regression	1	2710.0338	2710.0338	64.0388	0.0000
Residual	58	2454.4810	42.3186		
Total	59	5164.5148			

	<i>Coefficients</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-0.4667	0.8590	-0.5434	0.5890	-2.1862	1.2527
MKT-RF	1.2191	0.1523	8.0024	0.0000	0.9141	1.5240

Exercise 11.9. Suppose that returns can be described by the two-factor model

$$r_i = E[r_i] + \beta_{i1}(F_1 - E[F_1]) + \beta_{i2}(F_2 - E[F_2]) + \varepsilon_i,$$

where F_1 and F_2 denote the two common factors. Because of arbitrage activities as explained in the APT, such a model suggests that expected returns satisfy

$$E[r_i] = r_f + \beta_{i1} RP_1 + \beta_{i2} RP_2, \quad (*)$$

where RP_1 and RP_2 are the risk premiums related to the two factors. The riskfree rate is $r_f = 1\%$.

- (a) Suppose a portfolio A has an expected return of 9% and factor sensitivities $\beta_{A1} = 3$ and $\beta_{A2} = -0.7$. And a portfolio B has an expected return of 17% and factor sensitivities of $\beta_{B1} = 0.8$ and $\beta_{B2} = 1.2$. What are the risk premiums RP_1 and RP_2 ?
- (b) Suppose that the stock X has factor sensitivities $\beta_{X1} = 1.1$ and $\beta_{X2} = 0.5$. What should the expected return of stock X be according to the model? If the expected return on stock X is 13%, can you set up an apparent arbitrage strategy? Is it a true arbitrage (completely riskfree profit)?
- (c) Let r_m denote the return on the market portfolio, and let $\beta_i = \text{Cov}[r_i, r_m]/\text{Var}[r_m]$ denote the usual market beta of asset i . Show that

$$\beta_i = \frac{\text{Cov}[F_1, r_m]}{\text{Var}[r_m]} \beta_{i1} + \frac{\text{Cov}[F_2, r_m]}{\text{Var}[r_m]} \beta_{i2}. \quad (**)$$

- (d) Suppose $\text{Cov}[F_1, r_m] = 0.025$, $\text{Cov}[F_2, r_m] = 0.05$, and $\text{Var}[r_m] = 0.04$. Compute the market betas of portfolios A and B. Can their expected returns be consistent with the CAPM (for *some* value of the expected return on the market portfolio)?
- (e) Suppose instead that $\text{Cov}[F_1, r_m] = \text{Cov}[F_2, r_m] = 0.05$ and $\text{Var}[r_m] = 0.04$. Compute the market betas of portfolios A and B. Can their expected returns now be consistent with the CAPM (for *some* value of the expected return on the market portfolio)?
- (f) Show that if both the CAPM and the two-factor APT relation (*) hold in this economy, then the factor risk premiums are given by

$$\text{RP}_1 = \frac{\text{Cov}[F_1, r_m]}{\text{Var}[r_m]} (\text{E}[r_m] - r_f), \quad \text{RP}_2 = \frac{\text{Cov}[F_2, r_m]}{\text{Var}[r_m]} (\text{E}[r_m] - r_f).$$

Exercise 11.10. Suppose that over the next year the rate of return on the market portfolio has an expectation of 9% and a standard deviation of 20%. The one-year riskfree rate is 1%.

- (a) The correlation between the market return and the return on stocks of the company Generics Inc. over the next year is 0.5. The annual return on Generics stocks has a standard deviation of 50%. Assume that the basic CAPM holds. What is the market beta of Generics' stocks? What is the expected rate of return on Generics' stocks?
- (b) A portfolio consisting of 60% invested in the stocks of Generics and 40% in the stocks of the company Specifics Inc. has an expected return of 12.6% and a standard deviation of 40%. The return standard deviation of Specifics stocks is 50%. Assume that the basic CAPM holds. What is the market beta and expected return on Specifics stocks? What is the return correlation between Generics and Specifics?
- (c) After completing a fundamental analysis of Generics Inc., you firmly believe that the stocks have an expected return which is five percentage points higher than what you found in question (a). How would you invest to exploit the apparent mispricing? Briefly discuss the risks of this investment strategy.
- (d) Now suppose that equilibrium stock prices are in line with a two-factor model, where the first factor is the return on the market portfolio. A portfolio that has a zero market beta and a beta of one with respect to the second factor has an expected return equal to 8%. What is the factor risk premium on the second factor? If Generics stocks have a beta of 0.8 with respect to the second factor, does it still appear mispriced? Why or why not?

Exercise 11.11. One of your friends has owned stocks in Walmart Inc. (just Walmart or WMT in the following) in the last couple of years, and he tells you that it has been an excellent investment and recommends that you invest in it as well. Before deciding whether or not to follow his advice, you want to learn more about the characteristics of the Walmart stock. For that purpose you apply both the Single-Index Model and the Fama-French 3-factor model. You use 60 monthly return observations from January 2019 to December 2023. Returns are in percent throughout this problem. First, using Excel, you compute the summary statistics shown below. Here the columns refer to the excess return on Walmart stocks, the excess return on the stock market index, the return on the small-minus-big strategy, and the return on the high-minus-low strategy, respectively.

	WMT-RF	Mkt-RF	SMB	HML
Mean	1.0066	1.1848	-0.0453	-0.0330
Standard Error	0.6627	0.7177	0.3766	0.6079
Median	1.2627	2.2600	-0.1200	-0.4550
Standard Deviation	5.1334	5.5595	2.9170	4.7084
Sample Variance	26.3513	30.9078	8.5091	22.1693
Kurtosis	0.7429	0.0355	-0.1036	0.8552
Skewness	-0.4274	-0.3151	0.3432	0.0206
Count	60	60	60	60

The table below shows the output from a linear regression in Excel of the excess Walmart return on the excess market return. This is an estimation of the single-index model of returns,

$$r_i - r_f = \alpha_i + \beta_i (r_m - r_f) + \varepsilon_i.$$

Multiple R	0.4849					
R Square	0.2351					
Adjusted R Square	0.2219					
Standard Error	4.5281					
Observations	60					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	365.5149	365.5149	17.8268	0.0001	
Residual	58	1189.2119	20.5037			
Total	59	1554.7268				
	<i>Coefficients</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.4761	0.5979	0.7963	0.4291	-0.7208	1.6730
Mkt-RF	0.4477	0.1060	4.2222	0.0001	0.2355	0.6600

Now, please answer the following questions.

- (a) What is the estimate of the beta of Walmart? What is the interpretation of that value? Is the beta of Walmart significantly different from one?
- (b) What is the estimate of the alpha of Walmart. What is the interpretation of that value? Is the alpha of Walmart significantly different from zero?
- (c) What is the total variance of the excess returns on Walmart? How large is the systematic component? How large is the unsystematic (or firm-specific) component?

Next, you perform the regression

$$r_i - r_f = \alpha_i + \beta_{i,m} (r_m - r_f) + \beta_{i,SMB} SMB + \beta_{i,HML} HML + \varepsilon_i$$

corresponding to the Fama-French three-factor model. The output from Excel for this regression is shown in the table below.

Multiple R	0.5697					
R Square	0.3246					
Adjusted R Square	0.2884					
Standard Error	4.3304					
Observations	60					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	504.5892	168.1964	8.9693	0.0001	
Residual	56	1050.1376	18.7525			
Total	59	1554.7268				
	<i>Coefficients</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.3432	0.5742	0.5977	0.5524	-0.8071	1.4936
Mkt-RF	0.5373	0.1077	4.9901	0.0000	0.3216	0.7530
SMB	-0.4638	0.2052	-2.2599	0.0277	-0.8749	-0.0527
HML	-0.1715	0.1200	-1.4292	0.1585	-0.4120	0.0689

- (d) What are the estimates of the three factor betas of Walmart? How do you interpret those values?
- (e) Given the estimated betas, what should the average monthly excess return of Walmart stocks have been according to (i) the Single-Index Model and (ii) the Fama-French 3-factor model? Compare with the observed average excess return on Walmart stocks over the sample period.
- (f) What can you conclude from the above analysis about how attractive it is to invest in Walmart stocks now?

Exercise 11.12. You firmly believe that stock returns over the next year can be described by the two-factor model

$$r_i - r_f = \beta_{i1} (F_1 - r_f) + \beta_{i2} (F_2 - r_f) + \varepsilon_i.$$

Here F_1 and F_2 are the factors, which are assumed to be uncorrelated and to have expectations and standard deviations of

$$\text{E}[F_1] = 0.2, \quad \text{E}[F_2] = 0.02, \quad \text{Std}[F_1] = 0.2, \quad \text{Std}[F_2] = 0.1.$$

Furthermore, for any stock i , the residual return component ε_i has mean zero and is uncorrelated with both factors, and the residual return components of different stocks are also uncorrelated. The riskfree rate over the next year is $r_f = 0.01$.

You are considering forming a portfolio of four stocks with the following factor betas and residual standard deviations:

Stock number, i	β_{i1}	β_{i2}	$\text{Std}[\varepsilon_i]$
1	0.6	-1.0	0.1
2	-0.4	4.0	0.1
3	1.0	0.6	0.2
4	0.4	2.4	0.1

- (a) Compute the expected returns of the four stocks.
- (b) Show that the variance-covariance matrix of the four stocks is given by

$$\underline{\Sigma} = \begin{pmatrix} 0.0344 & -0.0496 & 0.0180 & -0.0144 \\ -0.0496 & 0.1764 & 0.0080 & 0.0896 \\ 0.0180 & 0.0080 & 0.0836 & 0.0304 \\ -0.0144 & 0.0896 & 0.0304 & 0.0740 \end{pmatrix}.$$

Compute the standard deviation of the return on each stock.

- (c) Determine the minimum-variance portfolio of the four stocks. Compute the expected return and standard deviation of this portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters.
- (d) Determine the tangency portfolio of the four stocks. Compute the expected return and standard deviation of this portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters.
- (e) Construct a diagram in which you draw the efficient frontier of the four risky assets as well as the efficient frontier of all assets (including the riskfree asset). Indicate where each of the stocks is located in the diagram.
- (f) Determine the factor betas of the tangency portfolio and the minimum-variance portfolio.
- (g) Now suppose that you do not want any exposure to the first factor so that you will only choose among portfolios with a zero beta with respect to the first factor. Among all such portfolios, determine the one with the smallest variance and an expected return of 0.05. Do the same for an expected return of 0.10 and 0.15. Mark the location of these portfolios in the diagram constructed above. Are the portfolios efficient? Discuss.

Exercise 11.13. Suppose that the stocks you can invest in can be split into two groups denoted by H and L. All the stocks in group H have the same return standard deviation σ_H , and all the stocks in group L have the same standard deviation σ_L . The portfolios considered in the following are equally-weighted portfolios consisting of N stocks from group H and N stocks from group L.

First suppose that the returns on all stocks are independent of each other.

- (a) Show that the standard deviation of your portfolio return is

$$\text{Std}[r_p] = \frac{\sqrt{\sigma_H^2 + \sigma_L^2}}{2\sqrt{N}}.$$

- (b) Suppose that $\sigma_H = 0.4$ and $\sigma_L = 0.3$. If you have access to 100 stocks from each group, what is the lowest obtainable portfolio standard deviation? How large does N have to be to obtain a portfolio standard deviation below 0.1?

Now suppose that the returns are not independent of each other, but instead that they are described by the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

For all stocks in both groups the standard deviation of the residual ε_i is the same, namely σ_ε . Let σ_m denote the standard deviation of the market return.

- (c) Explain why all stocks in group H must have the same market beta, β_H , and why all stocks in group L must have the same market beta, β_L .
- (d) Show that the standard deviation of your portfolio return is now

$$\text{Std}[r_p] = \sqrt{\frac{\sigma_m^2(\beta_H + \beta_L)^2}{4} + \frac{\sigma_\varepsilon^2}{2N}}.$$

- (e) Suppose that $\sigma_H = 0.4$, $\sigma_L = 0.3$, $\sigma_m = 0.2$, and $\sigma_\varepsilon = 0.25$. Determine β_H and β_L . If you have access to 100 stocks from each group, what is the lowest obtainable portfolio standard deviation?

Exercise 11.14. Suppose that returns can be described by the two-factor model

$$r_i = E[r_i] + \beta_{i1}(F_1 - E[F_1]) + \beta_{i2}(F_2 - E[F_2]) + \varepsilon_i,$$

where F_1 and F_2 denote the two common factors. The two factors are uncorrelated and have standard deviations of $\text{Std}[F_1] = \text{Std}[F_2] = 0.2$. Because of arbitrage activities as explained in the Arbitrage Pricing Theory, such a model suggests that expected returns satisfy

$$E[r_i] = r_f + \beta_{i1} \text{RP}_1 + \beta_{i2} \text{RP}_2, \quad (*)$$

where RP_1 and RP_2 are the risk premiums related to the two factors. The riskfree rate is $r_f = 1\%$.

Suppose the exchange-traded fund ETF-A has factor sensitivities of $\beta_{A1} = \beta_{A2} = 1$ and the exchange-traded fund ETF-B has factor sensitivities of $\beta_{B1} = -1.5$ and $\beta_{B2} = 2$. Both ETFs have expected returns of 15%.

- (a) Assuming that $(*)$ holds, what are the values of the risk premiums RP_1 and RP_2 ?
- (b) Which combination of the two ETFs will have a zero sensitivity with respect to F_2 ?

Suppose that the stocks of the company Hypothetics have factor sensitivities of $\beta_{H1} = -2$ and $\beta_{H2} = 1$ and that the residual return component ε_H for Hypothetics has a standard deviation of $\text{Std}[\varepsilon_H] = 0.4$.

- (c) What is the expected return and the return standard deviation of Hypothetics stocks according to the model?
- (d) If you expect the return on Hypothetics stocks to be 5%, can you set up an attractive portfolio of Hypothetics stocks, ETF-A, and ETF-B that has zero sensitivity to both factors? Is this an arbitrage strategy?

Exercise 11.15. You believe that over the next year the return r_i on any stock i satisfies the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

The one-year riskfree rate r_f is 1%, and the return r_m on the stock market index over the next year has an expectation of 7% and a standard deviation of 20%. You have the following information about the returns over the following year on the stocks of the companies *Hypothetics* and *Illuminati*:

Stock	α_i	β_i	$E[r_i]$	$\sigma[r_i]$
Hypothetics	3%	???	13%	50%
Illuminati	???	0.75	4%	40%

- (a) Find the beta value for *Hypothetics* and the alpha value for *Illuminati*.
- (b) What is the correlation between the returns of the two stocks according to the Single-Index Model?
- (c) You intend to set up a market-neutral portfolio consisting of stocks in *Hypothetics* and *Illuminati* as well as the riskfree asset. How does the weights of the two stocks in such a portfolio have to be related? What can you say about the alpha of such a portfolio? Under what condition on the portfolio weights will the alpha of the portfolio be positive?
- (d) Show that the expected return on such a market-neutral portfolio satisfies $E[r_p] = 1\% + 6\% \times \pi_H$, where π_H is the portfolio weight of *Hypothetics*.
- (e) Which market-neutral portfolio has an expected return of 7%? What is the standard deviation of the return on that portfolio?
- (f) Suppose that you can at most accept a standard deviation of the market-neutral portfolio being 25%. What is then the largest expected return such a portfolio can have?
- (g) Make a sketch of a diagram with betas along the horizontal axis and expected returns along the vertical axis. Draw the security market line in the diagram. Indicate the location of the two stocks, the market portfolio, the riskfree asset, and the two market-neutral portfolios constructed above.

Exercise 11.16. You firmly believe that the rates of return on stocks over the next year can be described by the two-factor model

$$r_i = a_i + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

in which the factors F_1 and F_2 are uncorrelated and have expectations $E[F_1] = E[F_2] = 0$ and variances $\text{Var}[F_1] = 0.03$ and $\text{Var}[F_2] = 0.02$. Furthermore, $E[\varepsilon_i] = 0$ and $\text{Cov}[F_k, \varepsilon_i] = 0$ for all i and for $k = 1, 2$. The riskfree rate over the next year is $r_f = 0.01$ or 1%.

You are considering forming a portfolio of five specific stocks for which you have estimated the following inputs:

Stock i	a_i	β_{i1}	β_{i2}	$\text{Var}[\varepsilon_i]$
Awesome Autos (AA)	0.15	2	2	0.25
Bjuti Bjutiques (BB)	0.12	1	3	0.04
Cyber Cigars (CC)	0.06	-1	5	0.09
Daunting Donuts (DD)	0.09	2	-1	0.36
Eggcellent Eggs (EE)	0.17	4	-2	0.16

- (a) Explain why a_i is the expected rate of return on stock i . For each stock, compute the variance and the standard deviation of the return as well as the Sharpe ratio. Explain your calculations in detail for Awesome Autos.
- (b) Explain in detail how you calculate the covariance between the return on Awesome Autos and the return on Bjuti Bjutiques. Verify that the variance-covariance matrix of the five stocks is given by

$$\underline{\Sigma} = \begin{pmatrix} 0.45 & 0.18 & 0.14 & 0.08 & 0.16 \\ 0.18 & 0.25 & 0.27 & 0.00 & 0.00 \\ 0.14 & 0.27 & 0.62 & -0.16 & -0.32 \\ 0.08 & 0.00 & -0.16 & 0.50 & 0.28 \\ 0.16 & 0.00 & -0.32 & 0.28 & 0.72 \end{pmatrix}$$

and determine the inverse of the variance-covariance matrix.

- (c) Determine the minimum-variance portfolio of the five stocks. Compute the expected return, standard deviation, and the factor betas of this portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters.

- (d) Determine the tangency portfolio of the five stocks. Compute the expected return, standard deviation, and the factor betas of this portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters.
- (e) Construct a diagram in which you draw the efficient frontier of the five risky assets as well as the efficient frontier of all assets (including the riskfree asset). Indicate where each of the stocks is located in the diagram.
- (f) Now suppose that you do not want any exposure to the second factor so that you will only choose among portfolios with a zero beta with respect to the second factor. The portfolio can include the five stocks and the riskfree asset. Among all such portfolios, determine the one with the smallest variance and an expected return of 0.10. Do the same for an expected return of 0.15. Mark the location of these portfolios in the diagram constructed above. Are the portfolios efficient? Discuss.

Exercise 11.17. Suppose that stock returns over the next year satisfy a three-factor model

$$r_i = r_f + \beta_{i1}(r_{F1} - r_f) + \beta_{i2}(r_{F2} - r_f) + \beta_{i3}(r_{F3} - r_f) + \varepsilon_i,$$

where r_i is the return on stock i , and

- $r_f = 0.01$ is the riskfree rate,
- each factor r_{Fk} is the return on a specific portfolio and you estimate that

$$\begin{aligned} E[r_{F1}] &= E[r_{F2}] = E[r_{F3}] = 0.05, \\ \text{Std}[r_{F1}] &= \text{Std}[r_{F2}] = \text{Std}[r_{F3}] = 0.10, \end{aligned}$$

- the three factors are uncorrelated,
- β_{ik} is the sensitivity of stock i 's return to factor k ,
- ε_i is the residual return component for stock i and satisfies

$$E[\varepsilon_i] = \text{Cov}[\varepsilon_i, r_{F1}] = \text{Cov}[\varepsilon_i, r_{F2}] = \text{Cov}[\varepsilon_i, r_{F3}] = 0,$$

- the factors and the residual are all normally distributed.

You believe that the returns on the stocks in the company HyperGigaMega follows the three-factor model with

$$\beta_{i1} = 1.2, \quad \beta_{i2} = 0.8, \quad \beta_{i3} = -0.5, \quad \text{Std}[\varepsilon_i] = 0.12.$$

- (a) What is the expected return, the standard deviation, and the Sharpe ratio of the stocks of HyperGigaMega?
- (b) Your clairvoyant uncle John says that he expects the returns of HyperGigaMega stocks to be 5 percentage points higher than what your three-factor model claims. Suggest a trading strategy which has zero betas with respect to all three factors and gives a positive expected excess return if John's expectations are correct. The strategy can involve positions in HyperGigaMega stocks, the three factor portfolios, and the riskfree asset. Is this strategy an arbitrage? Explain your answers.
- (c) You decide to ignore your uncle John and trust your three-factor model. However, you are worried that the factors are not uncorrelated after all. What would the Sharpe ratio of HyperGigaMega stocks be if the factor correlation was either $\text{Corr}[r_{F1}, r_{F2}] = 0.5$ or $\text{Corr}[r_{F1}, r_{F2}] = -0.5$? Is the Sharpe ratio increasing or decreasing in the correlation $\text{Corr}[r_{F1}, r_{F2}]$? Is the Sharpe ratio increasing or decreasing in the correlation $\text{Corr}[r_{F1}, r_{F3}]$? Explain!
- (d) Suppose you have invested \$1,000,000 in HyperGigaMega stocks. Assuming again uncorrelated factors, what is the 5% value-at-risk of this position in dollar terms? While you are convinced that factors 1 and 3 as well as factors 2 and 3 are pairwise uncorrelated, you are not so sure about the correlation between factors 1 and 2. Suppose you can accept a loss of \$280,000 or more with a probability of 5%. For which values of the correlation $\text{Corr}[r_{F1}, r_{F2}]$ is the maximum loss acceptable?

Exercise 11.18. Alice plans to form a portfolio to hold over the next year. She has identified

three specific stocks X, Y, and Z that she wants to include in her portfolio, together with a riskfree asset. She firmly believes that the annual stock returns satisfy the two-factor model

$$r_i - r_f = \beta_{i,m} (r_m - r_f) + \beta_{i,HML} HML + \varepsilon_i,$$

where m refers to the market portfolio and HML to the HML-factor of Fama and French, i.e. HML is the difference between the return on a portfolio of high book-to-market stocks and the return on a portfolio of low book-to-market stocks.

Based on the past dynamics of the factors, she estimates that

$$\begin{aligned} E[r_m - r_f] &= 0.06, & E[HML] &= 0.04, \\ \text{Std}[r_m - r_f] &= 0.20, & \text{Std}[HML] &= 0.10, & \text{Corr}[r_m - r_f, HML] &= 0.25. \end{aligned}$$

The riskfree rate over the next year is 0.02 (i.e. 2%).

By running regressions, Alice has estimated the betas and the residual risk of the three stocks to be the following:

Stock, i	$\beta_{i,m}$	$\beta_{i,HML}$	$\text{Std}[\varepsilon_i]$
X	1.2	-0.2	0.2
Y	0.9	1.4	0.5
Z	0.5	0.6	0.3

- (a) Verify that the variance-covariance matrix of the three stocks is

$$\underline{\Sigma} = \begin{pmatrix} 0.0956 & 0.0479 & 0.0259 \\ 0.0479 & 0.3146 & 0.0326 \\ 0.0259 & 0.0326 & 0.1066 \end{pmatrix}.$$

Explain in detail how you calculated the variance of X and the covariance between X and Y .

- (b) For each stock, calculate the expected rate of return, the standard deviation, and the Sharpe ratio.

Alice targets an expected rate of return of 0.10 (i.e. 10%).

- (c) Determine the mean-variance efficient portfolio of the three stocks that achieves exactly this expected return. What is the standard deviation of the portfolio return?
- (d) Determine the portfolio weights of the three stocks in the tangency portfolio as well as the portfolio's expected return, standard deviation, and Sharpe ratio. Does the tangency portfolio weights make sense given the inputs?
- (e) Which fraction of her wealth should Alice invest in the riskfree asset and each of the three stocks to obtain a 10% expected return with the lowest possible variance? Calculate the standard deviation of this portfolio and compare with the answer to question (c) above.

Suppose now that Alice can invest directly in the market portfolio through an ETF.

- (f) Which combination of the market portfolio and the riskfree asset has an expected return of 10%? What is the standard deviation of this combination? Compare this standard deviation with those derived in questions (c) and (e).

Suppose now that, in addition to the market portfolio and the riskfree asset, Alice can invest in the HML long-short portfolio. Recall that an investment in the HML long-short portfolio is a zero-cost investment and a weight w on HML means that a fraction w of wealth is invested in a portfolio of high book-to-market stocks and a fraction $-w$ is invested in a portfolio of low book-to-market stocks.

- (g) Which combination of the market portfolio, the riskfree asset, and the long-short HML portfolio has the lowest possible variance among all such combinations having an expected return of 10%? What is the standard deviation of this combination? Compare this standard deviation with those derived in questions (c), (e), and (f).

Exercise 11.19. Anna is thinking about how she should invest her financial wealth of \$50,000. She is aware that she should take her human capital into account when determining her optimal investments, and she has decided to use the extended mean-variance model for this purpose. Anna's relative risk aversion is $\gamma = 4$. Her human capital is \$450,000 and the annual return on the human capital has a standard deviation of 0.12.

First, Anna is considering to invest in a riskfree asset and a market-ETF tracking the stock market index. The riskfree rate is $r_f = 0.02$ per year. The annual return on the market-ETF has an expectation of $\mu_S = 0.07$ and a standard deviation of $\sigma_S = 0.16$.

- (a) Suppose the correlation between the market-ETF and Anna's human capital is 0.1. How should Anna invest her financial wealth if she does not impose constraints on her portfolio? How should Anna invest her financial wealth if she does not want to borrow money?
- (b) How do your answers to the questions in (a) change if the correlation between the market-ETF and Anna's human capital is 0.6? Briefly explain the effect of the correlation on the answers.

In the following, assume the correlation between the index-ETF and Anna's human capital is 0.1. Now Anna is considering to include an ETF called HiRisk in her portfolio. The annual return r_{Hi} on HiRisk is related to the return r_m on the market-ETF as follows:

$$r_{\text{Hi}} = r_f + \alpha_{\text{Hi}} + \beta_{\text{Hi}} (r_m - r_f) + \varepsilon_{\text{Hi}},$$

where ε_{Hi} and r_m are uncorrelated, and $E[\varepsilon_{\text{Hi}}] = 0$. Anna estimates that $\alpha_{\text{Hi}} = -0.01$, $\beta_{\text{Hi}} = 1.5$, and $\text{Std}[\varepsilon_{\text{Hi}}] = 0.2$. The covariance between HiRisk and Anna's human capital is identical to the covariance between the market-ETF and Anna's human capital.

- (c) Determine the expected return on HiRisk as well as the variance-covariance matrix of the market-ETF and HiRisk.
- (d) Without portfolio constraints, what is Anna's optimal portfolio of the market-ETF, the HiRisk ETF, and the riskfree asset? Briefly comment on your results.
- (e) If Anna does not want to borrow money, what is her optimal portfolio of the market-ETF, the HiRisk ETF, and the riskfree asset? Briefly comment on your results.

Exercise 11.20. You consider investing over the next year in four stocks that we refer to as stocks 1, 2, 3, and 4 in the following. The returns on these stocks are given by a two-factor model. More specifically, the rate of return on stock i satisfies

$$r_i = k_i + \beta_{i1} (F_1 - E[F_1]) + \beta_{i2} (F_2 - E[F_2]) + \varepsilon_i,$$

where F_1 and F_2 are random variables representing the factors, ε_i is a random variable, while k_i and the factor betas β_{i1} and β_{i2} are constants. Of course, we have

$$E[\varepsilon_i] = 0, \quad \text{Cov}[F_k, \varepsilon_i] = 0, \quad \text{Cov}[\varepsilon_i, \varepsilon_j] = 0$$

for all $i = 1, 2, 3, 4$, all $k = 1, 2$, and all $j = 1, 2, 3, 4$ with $j \neq i$. Both factors have a standard deviation of $\text{Std}[F_k] = 0.2$ and their correlation is $\text{Corr}[F_1, F_2] = 0.25$. The riskfree rate over the next year is $r_f = 0.01$. You can invest in the four stocks and the riskfree asset without any constraints.

After careful analysis, you have arrived at the following parameter values for the four stocks:

Stock number, i	k_i	β_{i1}	β_{i2}	$\text{Std}[\varepsilon_i]$
1	0.086	1.2	-0.2	0.15
2	0.140	0.5	0.9	0.10
3	0.094	0.8	0.2	0.10
4	0.126	0.2	1.0	0.20

- (a) For each of the four stocks, determine the expected rate of return $E[r_i]$ and the standard deviation $\text{Std}[r_i]$.
- (b) Given the expected returns and factor betas of stocks 1 and 2, determine the factor risk premiums RP_1 and RP_2 so that

$$E[r_i] = r_f + \beta_{i1}RP_1 + \beta_{i2}RP_2 \quad (*)$$

for $i = 1, 2$. With these values of the factor risk premiums, does Eq. (*) above also hold for stocks 3 and 4?

- (c) Show that the variance-covariance matrix of the four stock returns is given by

$$\underline{\Sigma} = \begin{pmatrix} 0.0769 & 0.0266 & 0.0376 & 0.0132 \\ 0.0266 & 0.0614 & 0.0314 & 0.0468 \\ 0.0376 & 0.0314 & 0.0404 & 0.0228 \\ 0.0132 & 0.0468 & 0.0228 & 0.0856 \end{pmatrix}.$$

- (d) Determine the tangency portfolio of the four stocks. Compute the expected return, the standard deviation, the Sharpe ratio, and the factor betas of this portfolio. Discuss whether the portfolio weights seem reasonable given the input parameters.
- (e) Construct a diagram in which you draw the efficient frontier generated by the four stocks and the riskfree asset. Plot standard deviations along the horizontal axis and means (i.e. expected returns) along the vertical axis. Indicate where the tangency portfolio and where each of the stocks are located in the diagram.
- (f) You prefer not to have any exposure to factor 2. Among the portfolios of the four stocks that have zero exposure to factor 2, which portfolio has the maximum Sharpe ratio? Compute the expected return, the standard deviation, the Sharpe ratio, and the factor betas of this portfolio.
- (g) Let us say that a portfolio is “constrained mean-variance efficient” if it has the minimum return variance among all the portfolios that have the same mean *and* a zero exposure to factor 2. The “constrained mean-variance efficient frontier of all assets” is then the set of all the combinations of standard deviation and mean return generated by the constrained mean-variance efficient portfolios of all assets (including the riskfree asset). What shape does the constrained efficient frontier of all assets have in the usual diagram with standard deviations along the horizontal axis and expected returns along the vertical axis? Why? If possible, add the constrained efficient frontier of all assets to the diagram you constructed in Question (e) above.
- (h) Suppose that you want to have a standard deviation of return which is at most 0.20 (i.e. 20%). What is the largest expected return you can obtain if you can invest freely in all five assets? What is the largest expected return you can obtain if you want to have a portfolio which has zero exposure to factor 2? Briefly discuss your results.

Exercise 11.21. The following table shows empirical estimates of the expectation, standard deviation, and correlations of the monthly excess returns on three portfolios. The estimates are based on monthly excess returns from July 1963 to October 2021 from the homepage of Professor Kenneth French. The market portfolio is the portfolio of all stocks listed at the main US stock exchanges. The robust portfolio is a portfolio of the stocks with the 30% highest operating profitability, and the weak portfolio is a portfolio of the stocks with the 30% lowest operating profitability. The portfolios are updated regularly and all the portfolio returns are value-weighted returns.

Portfolio	Exp excess	Std dev	Correlations		
			Market	Robust	Weak
Market	0.589%	4.452%	1.00	0.98	0.95
Robust	0.690%	4.410%	0.98	1.00	0.88
Weak	0.457%	5.249%	0.95	0.88	1.00

Use these empirical estimates in your answers to the questions below. All returns and moments refer to a one-month investment horizon.

- (a) Calculate the Sharpe ratios of the three portfolios and the variance-covariance matrix of their excess returns.
- (b) Which unconstrained combination of the three portfolios is maximizing the Sharpe ratio? What is the maximum value of the Sharpe ratio?

Now define the factor Robust-minus-Weak (or just RmW). The value of RmW in any given month is the return on the robust portfolio minus the return on the weak portfolio. Recall that an investment in RmW is a zero-cost investment with a long position in either the robust or weak portfolio and a corresponding short position in the other portfolio.

- (c) Calculate the expectation and standard deviation of RmW. Calculate the covariance and the correlation between RmW and the return on the market portfolio.
- (d) Which unconstrained combination of the market portfolio and RmW is maximizing the Sharpe ratio? What is the maximum value of the Sharpe ratio? Compare with your answers to Question (b) above.

You think that monthly stock returns satisfy the two-factor model

$$r_i - r_f = \beta_{i,m} (r_m - r_f) + \beta_{i,RmW} RmW + \varepsilon_i.$$

Here, r_m is the return on the market portfolio and RmW is the Robust-minus-Weak factor defined above. The residual ε_i has mean zero and a zero covariance with r_m and RmW, and the residuals of any two assets also have zero covariance.

You are tempted to invest in the stocks of the company Giggle. Based on a regression analysis, you estimate that Giggle's beta-coefficients and residual risk are

$$\beta_{G,m} = 1.4, \quad \beta_{G,RmW} = 0.5, \quad \text{Std}[\varepsilon_G] = 8\%.$$

- (e) What is the expected excess return, the standard deviation, and the Sharpe ratio of Giggle stocks?
- (f) Which combination of Giggle stocks, the market portfolio, and the RmW is maximizing the Sharpe ratio? State the Sharpe ratio of this combination. Briefly discuss your results.

CHAPTER 12

Market (in)efficiency and behavioral finance

NOTE: This chapter is still preliminary and incomplete!

12.1 Efficient markets

A key question in the 1970s was whether financial markets are informationally efficient or not, that is whether the price of a given stock reflects all the available information about its fundamentals (future dividends and appropriate discount rates). Informational efficiency means that prices move because of news. It rules out that risk-adjusted profits can be systematically earned on trading strategies using only the information available to the market participants and, in particular, based on historical price patterns. Competition in the financial markets should lead to informational efficiency, and empirical studies generally confirm that financial markets are to a large extent informationally efficient.

Let us be more specific. The **efficient market hypothesis** formalized by Fama (1970) states that prices in financial markets reflect all the available information. Exactly what “all the available information” means can be specified in different ways. Three increasingly demanding versions are generally considered:

1. **Weak-form** efficiency: current prices reflect all the information that can be derived from *historical market trading data* such as historical prices, trading volume, and short interest. Note that such historical data can be obtained at virtually zero costs.
2. **Semi-strong**-form efficiency: prices reflect *all publicly available information*. In addition to the historical market trading data, this includes e.g. information in accounting reports, earning forecasts, and announcements by corporations.
3. **Strong**-form efficiency: prices reflect *all relevant information* even non-public, that is private or insider, information.

Closely related is the **random walk hypothesis**, which claims that stock price changes are random and unpredictable. Informational efficiency was originally believed to imply the random walk hypothesis, see, for example, Samuelson (1965) and Fama (1970). This is an incorrect conclusion. The strict random walk hypothesis is too strong simply because assets generally have positive expected returns and thus a price increase is more likely than a price decrease. Moreover, returns in efficient markets can be predictable to some extent if there are variations in expected returns and therefore discount rates over time as noted, for example, by Fama and French (1988). Many reasonable economic models lead to time

variations in risk premiums and riskfree rates and can thus explain return predictability without introducing informational asymmetries or inefficiencies. A less restrictive (and less precise) formulation of the random walk hypothesis is that *short-term returns are essentially unpredictable*. If markets are efficient, short-term returns should be essentially unpredictable.

Should we expect markets to be efficient?

- Many professionals analyze stock markets \leadsto prices should reflect all publicly available information
- If an analyst finds a new type of publicly available information useful for predicting the return of a stock, then as investors start trading based on that information, it will be reflected by the price
- Gathering, processing, and analyzing information is costly. Why spend such resources if the price already reflects the information?
- **Grossman and Stiglitz (1980)**: Perfectly efficient markets are impossible!
- The market can only be efficient to an extent where the costs and benefits from information analysis are balanced
- **Pedersen (2015)**: Efficiently inefficient! Financial markets are inefficient enough that money managers can be compensated for their costs through the profits of their trading strategies and efficient enough that the profits after costs do not encourage additional active investing.

How efficient should we expect markets to be?

- Relatively easy to process historical information \leadsto expect near-perfect weak-form efficiency
- A small extra percentage return on a large position can cover lots of costs: 0.1% extra on \$10 billion \sim \$10 million
- Some “fundamental security analysis” of large stocks using all public information should pay off for large investors \leadsto expect near-perfect semi-strong-form efficiency for large stocks
- Expect less efficiency for small stocks, stocks receiving less attention, stocks in countries with less rigorous accounting rules, etc.
- Should not expect strong-form efficiency as this would conflict with insider trading restrictions

Implications of market efficiency:

- Assets/portfolios with identical cash flows should have identical prices. Prices should not allow any systematic pattern of arbitrage opportunities.
- Prices change only when new information arrives
- “Technical analysis” unprofitable
- (Most additional) “fundamental security analysis” unprofitable – only analysis leading to insights other investors don’t have is rewarded
- Impossible to persistently obtain positive abnormal returns.
 - “Abnormal return”: higher than the appropriately risk-adjusted return (requires a perfect CAPM-type model!)
 - Notice the “joint hypothesis problem”: if you detect abnormal returns, you do not know whether it is because markets are inefficient or because your risk adjustment was inadequate
- Investors should primarily follow a passive investment strategy, not an active strategy.

- Still a role for portfolio management: ensure diversification and fine-tune portfolio to investor characteristics.
- Still rebalance portfolio over time if, for example, risk or risk aversion changes.

The return over a given period is determined by the dividend(s) over the period and the price change. As explained in Chapter 6, the price of an asset should equal the sum of the expected future dividends, discounted appropriately. Hence, any price change must be due to changes in expected future dividends or the appropriate discount rates or both. The discount rate can be separated into a riskfree rate and a risk premium. The classic view was to assume constant discount rates so that returns were driven by revisions of expected future dividends. But over the most recent couple of decades, understanding the level and the variations of discount rates have taken most of the attention in the asset pricing literature.

Every now and then a trading strategy is discovered that involves only traded assets and offers an apparently abnormal high return. Is such a discovery evidence of market inefficiency? This is a difficult question to answer. A general principle in finance is that the average return on any asset or trading strategy should depend on the risks involved. Higher risk, higher average return. No pain, no gain. An abnormal return means a return which is higher than it should be, given the risks involved. Here “should be” means according to some theory of which risks are relevant and their quantitative effects on average returns.

Some economists (could be called “fundamentalists” or “rationalists”) claim that any such abnormal return detected is due to an inadequate adjustment either for the risks of the trading strategy or for trading frictions that make it hard or costly to implement the strategy in real-life markets, not that the market is informationally inefficient. Other economists (“behavioralists”) believe that various anomalies are due to behavioral biases in the sense that investors are unable to process the available information correctly or systematically make decisions that are incompatible with rational behavior and typical assumptions about preferences. The reader is referred to Chapter 12 of Bodie, Kane, and Marcus (2021) as well as Hirshleifer (2001), Barberis and Thaler (2003), and Barber and Odean (2013) for an introduction to behavioral finance and to Constantinides (2002), Ross (2005), and Cochrane (2011) for a critique of the behavioral approach.

12.2 Empirical evidence on market efficiency

The clearest evidence against market efficiency would be to document systematic arbitrage opportunities. Here “systematic arbitrage opportunities” mean that the same type of investment strategy frequently result in riskfree profits.

If we focus on stocks, it is generally rare to set up two portfolios that will have identical cash flows in all future. An obvious exception is when the same stock is traded on two exchanges.

Moving beyond arbitrage identification, tests of market efficiency are unlikely to reach a clear conclusion for several reasons:

- An apparent market inefficiency can be due to insufficient adjustment for risk, i.e., an incorrect model for expected returns.
- Maybe active investment strategies that work well are never reported to the public.
- Successful portfolio managers receive far more attention than unsuccessful portfolio managers. If many different investment strategies are implemented, some will appear successful *ex post*.
- The portfolio management industry has a clear interest in opposing market efficiency.

How can we test market efficiency?

- Stock price reaction to information/news (event studies)
- Persistent “anomalies”
- Performance of active investment strategies, investment funds, and stock market analysts

12.2.1 Event studies

Stock price reaction to *takeover announcements*:

- In almost all takeovers, the acquiring firm pays a substantial premium on the listed price of the stocks of target firm.
- Should see a larger price increase when takeover is announced.
- Slight pre-announcement upwards drift in prices, maybe due to leaks of information leading investors to adjust the takeover probability upwards.
- Large increase in prices at the announcement date.
- No post-announcement drift in prices.
- These results are consistent with the semi-strong form of market efficiency.

See, e.g., Keown and Pinkerton (1981).

Stock price reaction to *earnings announcements*:

- Firms divided into 10 deciles based on earnings surprise
- Positive [negative] surprise \rightsquigarrow positive [negative] returns
- Ranking as expected: returns increasing in the earnings surprise
- Large return reaction at the announcement date.
- There is a slight pre-announcement drift: returns are affected before the earnings announcement. Could be due to leak of the earnings numbers or due to the publication of other information indicating whether the upcoming earnings report is positive or negative.
- Some post-announcement drift: returns continue to increase somewhat in the following months for the firms that announced surprisingly high earnings. Conversely for firms that announced surprisingly low earnings.
- These return patterns suggest a profitable strategy, which questions the semi-strong form of market efficiency
- Event studies: always hard to filter out other news and to define “abnormal return”

See, e.g., Reinganum (1981), Rendleman, Jones, and Latané (1982), and Barber, George, Lehavy, and Trueman (2013). A risk-based explanation has been proposed by Savor and Wilson (2016).

Value of insider information:

- Insiders seem to be able to trade profitably in their own stock.
- Tendency for stock prices to rise after insiders intensively bought shares and to fall after intensive insider sales.
- Indication that prices did not already reflect insider information \rightsquigarrow strong-form efficiency doesn't hold.
- For a short period after significant insider buying is published, stock prices typically continue to increase – which questions semi-strong-form efficiency – but not enough to overcome transaction costs (and maybe other positive news is published).

12.2.2 Market anomalies and bubbles

As discussed in previous chapters, researchers have detected many return patterns that are inconsistent with the basic CAPM (see survey by Schwert (2003)). These market anomalies could be seen as an indication of market inefficiency, but they could also reflect that the CAPM provides an insufficient correction for risk. For example, the SMB and HML factors may somehow be measuring some aspect of risk relevant for the average investor.

The short-term momentum, long-term reversal pattern in returns is particularly challenging for the efficient market hypothesis, because it is based purely on past prices. If markets were just weak-form informationally efficient, the winners-minus-losers strategy should not be profitable. Unless, of course, recent winners are somehow conceived by the average investor as being more risky than recent losers. Implementing the strategy does require lots of trading and is thus costly, and most of momentum gains seem to come from short positions in small, illiquid stocks. But the momentum and reversal effects are also consistent with investors having a behavioral bias. If investors initially overreact to news and later modify their expectations, prices could exhibit short-term momentum and a subsequent reversal.

The appearance of *bubbles in asset prices* would question semi-strong market efficiency. Here the term “bubble” refers to a situation in which prices significantly exceed the fundamental, intrinsic value of the asset.

The most extreme bubble was arguably the tulip mania in the Netherlands around the year 1637. Tulips were imported to the Netherlands from Turkey in the mid-1500s. In the 1600s, Amsterdam prospered due to its strong role in growing international trade. Tulips became increasingly popular for decoration and as a symbol of wealth. Prices rose steadily, and bulbs were purchased at higher and higher prices. Some speculators borrowed significant amounts of money to purchase tulip bulbs, being convinced that they could soon resell them at a profit. In February 1637, a single tulip bulb was worth about ten times a craftsman’s annual income. A single rare bulb was allegedly exchanged for:

- Two lasts of wheat
- Four lasts of rye
- Four fat oxen
- Eight fat swine
- Twelve fat sheep
- Two hogsheads of wine
- Four tuns of beer
- Two tons of butter
- 1,000 lb. of cheese
- A complete bed
- A suit of clothes
- A silver drinking cup

But very suddenly, prices started to drop dramatically. The bubble had burst.

Bubbles seem to be quite rare, in particular in modern financial markets. Here are some selected other events often characterized as bubbles:

- South Sea bubble, U.K., 1720.
- Railway mania, U.K., 1840s
- Stock market bubble, U.S., 1922-29
- Stock and real estate bubble, Japan, 1986-91

- Dot-com bubble, mainly U.S., 1996-2000
- Real estate, U.S., Ireland, Spain, U.K., Denmark (?), approximately 2000-2007

Bubbles are closely related to the term Irrational Exuberance coined by Shiller (2000, 2005).

Are bubbles evidence of market inefficiency?

- It is often very difficult to determine the fundamental value of an asset and therefore to detect a bubble.
- A small change in the expected dividend growth of a long-term asset (or in the appropriate risk-adjusted discount rate) can lead to a large change in fundamental stock price.
- Bubbles are typically detected *ex post* from steep fall in prices.
- Sometimes extreme speculative behavior, highly leveraged purchases, and intensive shorting activities are taken as indicators of an on-going bubble.
- Even if you identify a bubble when you are in it, it may be difficult to profit from that knowledge:
 - you do not know when the bubble will burst, so if you invest “against it” you may lose a lot before you will profit;
 - maybe it is impossible to short the asset;
 - it takes courage to be the first to go against the stream.

12.2.3 Investor and analyst performance

Barber, Lehavy, McNichols, and Trueman (2001) investigate empirically whether there is any value in analysts' recommendations. Among their conclusions are:

- There is a strong bias in analysts' recommendations: many more buy recommendations than sell recommendations.
- Without transaction costs:
 - investors will profit from taking a long position in a portfolio of the stocks with the best recommendations and a short position in a portfolio of the stocks with the worst recommendations. Using the Fama-French three-factor model, the abnormal return is 0.989% per month.
 - the return on the long-short strategy decreases with infrequent or delayed rebalancing.
- With transaction costs:
 - the strategy requires lots of trading (maybe this is why analysts frequently change their recommendations, i.e., to generate trade).
 - the risk-adjusted *net* returns are not reliably greater than zero.
- if you want/have to buy or sell, it pays off to follow recommendations.

Many analysts form and publish a target price 12 months into the future for each of the most valuable stocks. These target prices can then be compared to the current market price of the stock and a corresponding predicted return can be calculated. The predicted return implied by the analyst's target price is thus a more granular measure than the buy/sell recommendations. An obvious application is to invest in the stocks with the largest predicted returns and, maybe, sell or short the stocks with the most negative predicted returns. The empirical evidence on the precision of the target prices and the profitability of the associated investment strategies is mixed with Da and Schaumburg

(2011), Bilinski, Lyssimachou, and Walker (2013), and others reporting supporting evidence whereas Dechow and You (2020), Almeida and Gaspar (2021), and others concluding the opposite.

Mutual fund performance: No convincing evidence that the typical mutual fund manager can outperform passive benchmarks \rightsquigarrow support that markets are very efficient in the semi-strong form. See Elton and Gruber (2013). Many studies conclude that most active investors (whether professional fund or portfolio managers or private investors) underperform their benchmarks when costs are taken into account. However, some studies are more supportive of the merits of active investors, cf. the discussion by Cremers, Fulkerson, and Riley (2019).

Hedge fund performance: See Fung and Hsieh (2013), Joenväärä, Kauppila, Kosowski, and Tolonen (2021).

Individual investor performance: See Barber and Odean (2013).

12.3 Investor behavior

Overview of empirical studies of individual investor behavior and performance: Barber and Odean (2013). Evidence from Swedish investors is presented and discussed by Dahlquist, Martinez, and Söderlind (2017).

Empirical evidence on how individuals and households invest can be found in Guiso and Sodini (2013).

Performance evaluation of institutional investors: Ferson (2013) focuses on methodology but provides empirical examples.

Conventional finance theory assumes that investors act rationally by optimizing expected utility and therefore

- know the true probability distribution of all relevant random variables,
- are able to process a lot of information, and
- are able to solve quite difficult optimization problems.

Behavioral finance is based on how investors seem to make decisions in real life, which on some counts appears inconsistent with rational behavior

- errors in information processing
- behavioral biases

Errors in information processing:

- **Forecasting errors:** people give too much weight to recent experience and to personal experience, and tend to make too extreme forecasts
- **Representativeness:** short samples are given too much weight
- **Conservatism:** investors respond slowly to news
- **Overconfidence:** people (particularly males) tend to overestimate their abilities and the precision of their forecasts.

Success due to talent, failure due to bad luck!

- could explain the wide-spread active investment management
- high trading activity of men, despite low returns

... can contribute to an explanation of the momentum-reversal pattern in returns. Using data from a discount brokerate, Barber and Odean (2001) find a link between overconfidence and gender:

- Men trade about 50 percent more than women
- Both men and women reduce their portfolio returns when they trade excessively
- The portfolio return for men is 94 basis points lower than portfolio returns for women
- The portfolio return for single men is 144 basis points lower than the portfolio return for single women
- Men take more risky investments

Overconfidence and underdiversification:

- Investors tend to invest too heavily in shares of the company for which they work.
 - Your earning power (income) also depends on this company.
 - Your retirement nest-egg also depends on this company.
- Invest too heavily in the stocks of local companies.
 - Perhaps you know someone personally who works there.
 - Perhaps you read about them in your local paper.
 - You are confident that you have a high degree of knowledge about local companies.
- In general, overconfidence leads to investor underdiversification.

Behavioral biases.

- **Framing:** decisions can be affected by how choices are framed/presented
- **Sentiments:** decision can be affected by the sentiment of the decision maker, whether he is in a good or a bad mood
- **Mental accounting:** investors fail to see relation between some of their decisions
 - e.g., investing conservatively with one account and aggressively with another account
 - decision depends on past: more likely to sell stocks with gains than those with losses
 - more willing to take risks using newly obtained profits than using initial investment
- **Regret avoidance:** regret bad outcomes more if decision was unconventional or risky \rightsquigarrow reluctant to invest in little-known companies or recent losers or low-valued companies
- **Disposition effect or loss aversion:** do not want to realize losses \rightsquigarrow hold on to loser-stocks

The loss aversion is a key element in the *prospect theory* which was proposed by Kahneman and Tversky (1979) as an alternative to the standard expected utility framework. The theory captures that

- people focus on gains and losses rather than final wealth;
- people are risk-averse over gains and risk-seeking over losses;
- people are more sensitive to losses than to gains (kink at zero);
- people tend to give higher weights to low-probability events such as winning the lottery or experiencing extremely bad returns.

As shown by Frazzini (2006), the disposition effect may explain why investors often underreact to news as indicated by the post-announcement drift in prices following earnings announcements.

Implications for asset prices

Do behavioral biases imply inefficiency? For behavioral biases to affect market prices (and thus make the market inefficient), the following conditions must all hold:

- it must be that many investors make irrational investment decisions, **and**
- the collective irrationality of these investors leads to an overly optimistic or pessimistic market situation, **and**
- this situation cannot be corrected via arbitrage by rational, well-capitalized investors.

Limits to arbitrage.

- **Fundamental risk:** presumed mispricing can get worse.

Keynes: “Markets can remain irrational longer than you can remain solvent.”

- **Implementation costs:** short-selling may be restricted or very costly, in particular when you don’t know the appropriate time horizon
- **Model risk:** Maybe the price is really right and you are wrong?

Models with heterogeneous investors, social networks... see, e.g., Ottaviani and Sørensen (2015), Pedersen (2022)

12.4 Summary

- Financial markets generally seem to be relatively efficient
- Growing recognition that investor irrationality might explain some observed return patterns that are difficult (though not impossible) to reconcile with conventional finance theory
- The behavioral approach to finance is quite unstructured – nice, independent explanations of various pricing anomalies, but no unified theory that can explain a range of anomalies
- Statistical and economic significance of many anomalies is weak, in particular when taking transaction costs into account
- An investor can avoid exposure to any behavioral biases by following a passive, largely indexed, portfolio strategy – hard to beat whether other investors are rational or behavioral.

12.5 Exercises

Exercise 12.1. At a family dinner party, you are seated next to your cousin Carl, whom you have not seen since you were both kids. As soon as you tell him that you are studying finance, he starts bragging about his successful stock investments. He claims to have beaten the market for three consecutive years. How would you try to convince him that financial markets are quite efficient?

Exercise 12.2. Some mutual funds appear to outperform the market several years in a row.

- (a) Does this compromise the Efficient Market Hypothesis?
- (b) Suppose that each year the probability that any given fund outperforms the market is 50% and that the returns of the fund in different years are statistically independent. What is the probability that a given fund will outperform the market in 5 consecutive years? What about 10 consecutive years?
- (c) There are about 80,000 mutual funds worldwide. How many of those funds would you expect to outperform the market in 5 consecutive years? Or in 10 consecutive years?

Exercise 12.3. During the “dot.com boom” in the late 1990’s, the stock prices of companies that announced a name change to an internet-related “dot.com” name increased significantly around the announcement date. This was documented by Cooper, Dimitrov, and Rau (2001). Is this phenomenon consistent with the Efficient Market Hypothesis?

Exercise 12.4. Which of the following events would be the most challenging for the weak-form of market efficiency?

- (a) CEOs earn abnormally high returns on their trades of the company's stocks.
- (b) Stock prices move when the companies announce their earnings over the most recent period.
- (c) In most years, stocks offer higher returns in January than in other months.
- (d) Illiquid stocks tend to offer higher returns than liquid stocks.

Exercise 12.5. Your uncle Tom calls you on the phone and offers a great investment advice for you. He is confident that the demand of wind mills will be increasing over the next couple of decades. Therefore, he has just invested all of his savings in stocks of wind mill producers around the world, and he urges you to do the same. How do you reply?

Exercise 12.6. Illuminati Inc. just announced that their earnings last year were 15% higher than the year before. Yet its stock price dropped by 6% following the announcement. Is that a violation of the Efficient Market Hypothesis?

Exercise 12.7. The equity premium is most likely higher in recessions than in booms. Is this a violation of the Efficient Market Hypothesis? Explain how such variations in the market risk premium may lead to changes in individual stock prices that appear excessive relative to the positive news most companies report in expansions and the negative news most companies report in recessions.

Exercise 12.8. Prospect theory says that investors are risk averse with respect to gains and risk seeking with respect to losses. This could apply to gains and losses on the total portfolio, but in combination with the idea of mental accounting this may even apply to holdings of stocks in any single company. This may impact how the price of a stock responds to news about the issuing company.

Let us say that the stock *trades at a capital gain* if most current owners of a stock have bought the stock at a price lower than the current stock price. Conversely, the stock is said to *trade at a capital loss* if most current owners of a stock have bought the stock at a price exceeding than the current stock price.

- (a) Explain why the stock price response to *negative* news might depend on whether the stock trades at a capital gain or at a capital loss in the way illustrated in the left panel of Figure 12.1.
- (b) Explain why the stock price response to *positive* news might depend on whether the stock trades at a capital gain or at a capital loss in the way illustrated in the right panel of Figure 12.1.
- (c) Explain how the above considerations are related to the so-called post-earnings announcement price drift.

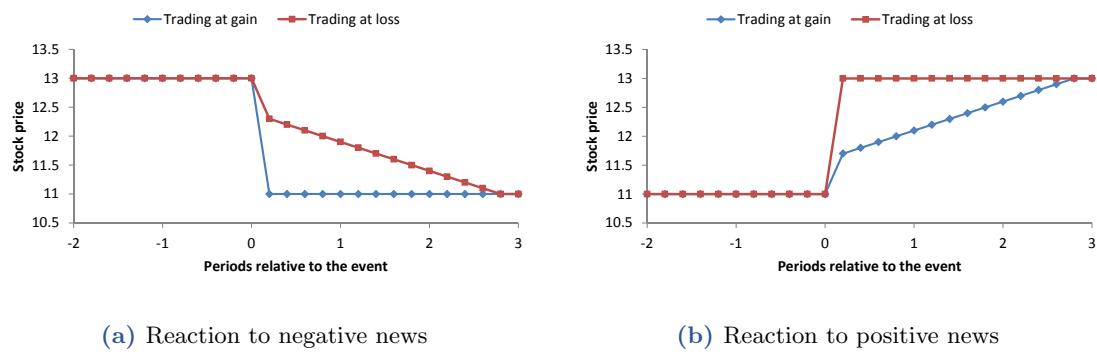


Figure 12.1: Stock price response to news.
At date 0, news are published that changes the fundamental value either from \$13 to \$11 (left panel) or from \$11 to \$13 (right panel). The graphs are reconstructions of Figures 3 and 4 in Frazzini (2006).

CHAPTER 13

Active portfolio management

The basic CAPM of Chapter 10 prescribes a passive investment strategy. Each investor should simply hold a mix of the market portfolio and the riskfree asset, where the exact mix depends only on the investor's risk aversion. This makes some intuitive sense because, by definition, the average investor must hold the market portfolio. Moreover, financial markets are constantly scrutinized by thousands of clever (and maybe some not so clever) traders and analysts looking for favorable investment opportunities, so prices in financial markets are generally believed to be fair and not allowing any easy profits to be made. Hence, a reasonable starting point for a portfolio decision is that you cannot beat the market, i.e., all assets are fairly priced and have zero alphas. Often the market portfolio is thought of as a portfolio of all stocks, and in practice you can obtain a position in the stock market portfolio at low costs through ETFs or passive mutual funds tracking a broad stock market index, cf. the description in Section 1.2.4. Indeed, in the recent decade we have seen a large growth in ETFs and passive funds together with a much smaller growth and most recently a stagnation in actively managed funds. According to Morningstar, the passive U.S. equity funds caught up with the active funds in April 2019, both having about 4.3 trillion USD in assets, whereas ten years earlier the active funds had more than six times as many assets under management as the passive funds.¹

Despite the growth and apparent academic support of passive investment, many investors still pursue an active investment strategy by deviating from the market portfolio in an attempt to beat the market, i.e., to obtain a higher expected return or, more generally, a better risk-return tradeoff than the market portfolio offers. How can an active investment strategy be justified? One of the many critical assumptions of the CAPM is that investors agree on the expectations, variances, and correlations of all assets. Obviously, this is not the case in real life. You may think that the stock of some company is mispriced and thus has a non-zero alpha relative to the CAPM. This means that your view on the expected return on the stock differs from the market view, i.e., the average view of investors. Then you probably want your portfolio weight of that stock to differ from the stock's market weight. But by how much? By overweighting a stock relative to its market share, you expose yourself to the stock's idiosyncratic risk. Hence, you may want to balance the alpha and the idiosyncratic risk. The Treynor-Black model in Section 13.1

¹See <https://www.morningstar.com/blog/2019/06/12/asset-parity.html>, accessed September 5, 2019.

shows how this can be done within a mean-variance setting. Section 13.2 presents the Black-Litterman model which allows the investor to incorporate more complex subjective views on the future returns, for example a view that one stock will outperform another stock by a certain return difference.

Active investors often intend to beat the market or at least a certain benchmark index. Of course, the performance of the active investors should be judged by comparing their realized returns to the returns of the benchmark. Section 13.3 discusses various measures of investment performance that are often used in the evaluation of active fund managers.

The role and merits of active investing in financial markets have been reconsidered in a recent academic literature, see, e.g., Pedersen (2015, 2018) and Garleanu and Pedersen (2018).

13.1 The Treynor-Black model

Suppose that asset returns over a given period follow a factor model but some assets are mispriced in the sense that they have a non-zero alpha relative to the factor model. Intuitively, you will probably want to overweight assets with a positive alpha and underweight asset with a negative alpha. But by doing so, you increase the non-systematic risk of your portfolio. What is the optimal tradeoff between alpha and non-systematic risk? What is the optimal portfolio in the presence of assets with non-zero alpha? The **Treynor-Black model** gives an answer. The approach was introduced by Treynor and Black (1973) in the setting of the Single-Index Model but, as shown below, it is also applicable in multi-factor models. The model builds on Markowitz' mean-variance framework in the sense that the assumed objective of the investor is to maximize the Sharpe ratio of the overall portfolio over the next period.

13.1.1 Treynor-Black in the Single-Index Model

Assume that returns on individual assets follow the Single-Index Model described in Section 11.2, i.e. that the return on any asset i over the next period is

$$r_i = r_f + \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i$$

with the usual assumptions. If the strict version of the Arbitrage Pricing Theory would hold, all assets should have $\alpha_i = 0$, and the market portfolio is the portfolio with the maximum Sharpe ratio so the optimal portfolio is some combination of the market portfolio and the riskfree asset. The CAPM leads to the same conclusions. When an asset has a non-zero alpha, the asset is mispriced relative to the CAPM.

Suppose you firmly believe that J assets have a non-zero alpha relative to the CAPM. We combine these assets into an *active portfolio*, which has an excess return given by

$$r_A - r_f = \alpha_A + \beta_A(r_m - r_f) + \varepsilon_A. \quad (13.1)$$

If we let π_1, \dots, π_J denote the portfolio weights of the J assets in the active portfolio, we have from Eq. (11.20) that

$$\alpha_A = \sum_{i=1}^J \pi_i \alpha_i, \quad \beta_A = \sum_{i=1}^J \pi_i \beta_i, \quad \varepsilon_A = \sum_{i=1}^J \pi_i \varepsilon_i. \quad (13.2)$$

First, we will figure out how a given active portfolio can contribute to the Sharpe ratio of the overall portfolio. The investor can invest in the market portfolio and in the active

portfolio. Let w_A denote the weight of the active portfolio so that the weight of the market portfolio is $1 - w_A$. The return on the combined portfolio is

$$r_p = w_A r_A + (1 - w_A) r_m.$$

With mean-variance preferences, you want to find the combination that maximizes the Sharpe ratio

$$\text{SR}_p = \frac{\text{E}[r_p] - r_f}{\text{Std}[r_p]}.$$

Recall from Section 7.2 that the tangency portfolio is the one maximizing the Sharpe ratio, so we have to find the tangency portfolio in the special case where the only investment objects are the market portfolio and the active portfolio. Then this tangency portfolio can be combined with the riskfree asset as explained in Sections 7.2 and 7.3.

Theorem 13.1

Assume returns follow the Single-Index Model. Then the maximum Sharpe ratio is obtained by the weight

$$w_A^* = \frac{\frac{\alpha_A}{\text{Var}[\varepsilon_A]}}{\frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} + (1 - \beta_A) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}} \quad (13.3)$$

on the active portfolio and the weight $1 - w_A^*$ on the market portfolio. The squared Sharpe ratio of this specific combination satisfies

$$(\text{SR}_p)^2 = (\text{SR}_m)^2 + \left(\frac{\alpha_A}{\text{Std}[\varepsilon_A]} \right)^2, \quad (13.4)$$

where $\text{SR}_m = (\text{E}[r_m] - r_f) / \text{Std}[r_m]$ is the Sharpe ratio of the market portfolio.

The contribution of the active portfolio to the overall Sharpe ratio is determined by its *information ratio* $\alpha_A / \text{Std}[\varepsilon_A]$, which measures the trade-off between the extra expected return and the extra risk caused by the active portfolio. This ratio was defined for an individual asset already in (11.56) and is also known as the *appraisal ratio*. Note that a negative alpha is just as useful as a positive alpha in maximizing the overall Sharpe ratio. Hence, the squared information ratio is what matters.

Proof

We apply the results on the tangency portfolio in Theorem 7.6 to the special case where the investment objects are the market portfolio (“asset 1”) and the active portfolio (“asset 2”). Recall that the tangency portfolio weights are given by the vector $\underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})$ divided by the sum of its elements (which in the general notation is $\mathbf{1} \cdot \underline{\Sigma}^{-1}(\boldsymbol{\mu} - r_f \mathbf{1})$). With the Single-Index specification (13.1), we have

$$\boldsymbol{\mu} - r_f \mathbf{1} = \begin{pmatrix} \text{E}[r_m] - r_f \\ \alpha_A + \beta_A (\text{E}[r_m] - r_f) \end{pmatrix}, \quad \underline{\Sigma} = \begin{pmatrix} \text{Var}[r_m] & \beta_A \text{Var}[r_m] \\ \beta_A \text{Var}[r_m] & \beta_A^2 \text{Var}[r_m] + \text{Var}[\varepsilon_A] \end{pmatrix}.$$

Since the determinant of $\underline{\Sigma}$ is

$$\text{Var}[r_m] \left(\beta_A^2 \text{Var}[r_m] + \text{Var}[\varepsilon_A] \right) - (\beta_A \text{Var}[r_m])^2 = \text{Var}[r_m] \text{Var}[\varepsilon_A],$$

we find that

$$\underline{\Sigma}^{-1} = \frac{1}{\text{Var}[r_m] \text{Var}[\varepsilon_A]} \begin{pmatrix} \beta_A^2 \text{Var}[r_m] + \text{Var}[\varepsilon_A] & -\beta_A \text{Var}[r_m] \\ -\beta_A \text{Var}[r_m] & \text{Var}[r_m] \end{pmatrix}.$$

Straightforward multiplication now leads to

$$\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) = \begin{pmatrix} \frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} - \frac{\alpha_A \beta_A}{\text{Var}[\varepsilon_A]} \\ \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \end{pmatrix},$$

and the sum of the elements are

$$\frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} + (1 - \beta_A) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}.$$

Hence, the weight of the active portfolio becomes

$$w_A = \frac{\frac{\alpha_A}{\text{Var}[\varepsilon_A]}}{\frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} + (1 - \beta_A) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}}$$

which confirms (13.3), and the weight of the market portfolio is

$$w_m = 1 - w_A = \frac{\frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} - \frac{\alpha_A \beta_A}{\text{Var}[\varepsilon_A]}}{\frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} + (1 - \beta_A) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}}.$$

From (7.39), the squared Sharpe ratio of the tangency portfolio is

$$\begin{aligned} (\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) &= \begin{pmatrix} \text{E}[r_m] - r_f \\ \alpha_A + \beta_A (\text{E}[r_m] - r_f) \end{pmatrix} \cdot \begin{pmatrix} \frac{\text{E}[r_m] - r_f}{\text{Var}[r_m]} - \frac{\alpha_A \beta_A}{\text{Var}[\varepsilon_A]} \\ \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \end{pmatrix} \\ &= \frac{(\text{E}[r_m] - r_f)^2}{\text{Var}[r_m]} + \frac{\alpha_A^2}{\text{Var}[\varepsilon_A]}, \end{aligned}$$

which is equivalent to (13.4).

In the above theorem, the active portfolio is assumed to be given. But how do we set up the active portfolio in the best way? Since the overall objective is to maximize the Sharpe ratio—or, equivalently, the squared Sharpe ratio—it follows from (13.4) that the active part should be designed to maximize the squared information ratio $\alpha_A^2 / \text{Var}[\varepsilon_A]$.

The next theorem provides the optimal portfolio weights as well as the contribution of the optimal active portfolio to the squared Sharpe ratio of the total portfolio. As explained above, we assume that the active portfolio consists of K assets with relative weights π_1, \dots, π_J , of course with $\sum_{i=1}^J \pi_i = 1$.

Theorem 13.2

The optimal active portfolio is given by

$$\pi_i = \frac{\frac{\alpha_i}{\text{Var}[\varepsilon_i]}}{\sum_{j=1}^J \frac{\alpha_j}{\text{Var}[\varepsilon_j]}}, \quad i = 1, 2, \dots, J, \quad (13.5)$$

and has the property that

$$\frac{\alpha_A^2}{\text{Var}[\varepsilon_A]} = \sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]} = \sum_{i=1}^J \left(\frac{\alpha_i}{\text{Std}[\varepsilon_i]} \right)^2. \quad (13.6)$$

We see that the relative weight of each asset in the optimal active portfolio is determined by the ratio of its alpha to its residual variance. The contribution of the active portfolio to the squared Sharpe ratio of the overall investment is equal to the sum of the squared information ratios of the mispriced assets.

Proof

Note that

$$\alpha_A = \sum_{i=1}^J \pi_i \alpha_i, \quad \text{Var}[\varepsilon_A] = \text{Var} \left[\sum_{i=1}^J \pi_i \varepsilon_i \right] = \sum_{i=1}^J \pi_i^2 \text{Var}[\varepsilon_i],$$

cf. (11.20), and hence the squared information ratio is

$$\frac{\alpha_A^2}{\text{Var}[\varepsilon_A]} = \frac{\left(\sum_{i=1}^J \pi_i \alpha_i \right)^2}{\sum_{i=1}^J \pi_i^2 \text{Var}[\varepsilon_i]}. \quad (13.7)$$

First, ignore the constraint that weights sum to one. In the maximization of (13.7), the first-order condition with respect to π_i is²

$$2\alpha_i \left(\sum_{j=1}^J \pi_j^2 \text{Var}[\varepsilon_j] \right) = 2\pi_i \text{Var}[\varepsilon_i] \left(\sum_{j=1}^J \pi_j \alpha_j \right),$$

which implies that

$$\pi_i \left(\sum_{j=1}^J \pi_j \alpha_j \right) = \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \left(\sum_{j=1}^J \pi_j^2 \text{Var}[\varepsilon_j] \right), \quad i = 1, 2, \dots, J. \quad (13.8)$$

Obviously, the sum of the left-hand sides must equal the sum of the right-hand sides. The sum of the left-hand sides of (13.8) is

$$\sum_{i=1}^J \pi_i \left(\sum_{j=1}^J \pi_j \alpha_j \right) = \left(\sum_{j=1}^J \pi_j \alpha_j \right) \sum_{i=1}^J \pi_i = \sum_{j=1}^J \pi_j \alpha_j,$$

since $\sum_{i=1}^J \pi_i = 1$. The sum of the right-hand sides of (13.8) is

$$\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \left(\sum_{j=1}^J \pi_j^2 \text{Var}[\varepsilon_j] \right) = \left(\sum_{j=1}^J \pi_j^2 \text{Var}[\varepsilon_j] \right) \sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]}.$$

In (13.8), we divide each side by the corresponding sum to get

$$\pi_i = \frac{\frac{\alpha_i}{\text{Var}[\varepsilon_i]}}{\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]}} = \frac{\frac{\alpha_i}{\text{Var}[\varepsilon_i]}}{\sum_{j=1}^J \frac{\alpha_j}{\text{Var}[\varepsilon_j]}}$$

as was to be shown.

The alpha and residual variance of the optimal active portfolio are

$$\alpha_A = \sum_{i=1}^J \pi_i \alpha_i = \frac{\sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]}}{\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]}},$$

and

$$\text{Var}[\varepsilon_A] = \sum_{i=1}^J \pi_i^2 \text{Var}[\varepsilon_i] = \frac{\sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]^2} \text{Var}[\varepsilon_i]}{\left(\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \right)^2} = \frac{\sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]}}{\left(\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \right)^2},$$

implying that

$$\frac{\alpha_A^2}{\text{Var}[\varepsilon_A]} = \frac{\left(\sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]} \right)^2}{\left(\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \right)^2} \times \frac{\left(\sum_{i=1}^J \frac{\alpha_i}{\text{Var}[\varepsilon_i]} \right)^2}{\sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]}} = \sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]},$$

as claimed in the theorem.

Together Theorems 13.1 and 13.2 define the optimal portfolio. The fraction $1 - w_A^*$ is invested in the market portfolio of all assets, which is often represented by a broad stock market index. The fraction w_A^* is invested in the active portfolio defined by (13.5).

Example 13.1

Suppose the riskfree rate is $r_f = 0.01$, and that the market portfolio has an expected rate of return of $E[r_m] = 0.07$ and a standard deviation of $\text{Std}[r_m] = 0.2$. The market Sharpe ratio is then $\text{SR}_m = 0.3$.

You have identified three underpriced stocks denoted by X, Y, and Z. The left part of

²The expression we want to maximize is of the form $f(x) = g(x)^2/h(x)$. Differentiating this fraction and using the chain rule, we get

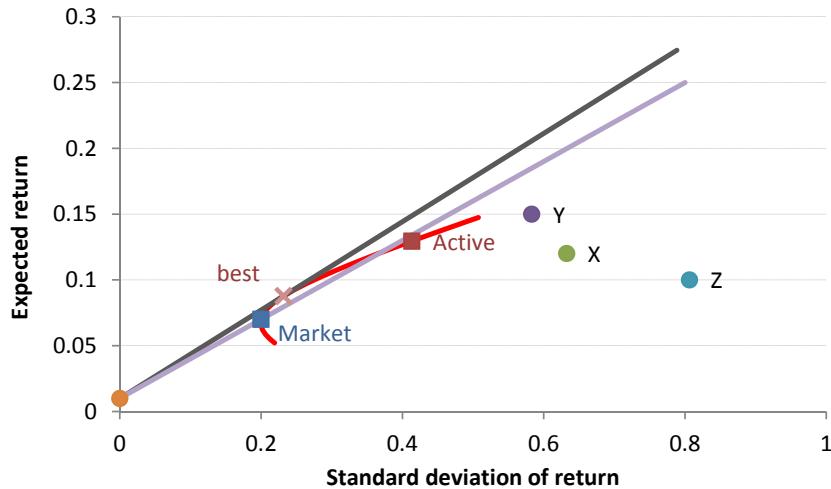
$$f'(x) = \frac{2g(x)g'(x)h(x) - g(x)^2h'(x)}{h(x)^2} = \frac{g(x)}{h(x)^2} (2g'(x)h(x) - g(x)h'(x)),$$

so the first-order condition $f'(x) = 0$ holds if and only if $2g'(x)h(x) = g(x)h'(x)$.

Inputs			Derived values						
	α_i	β_i	$\text{Std}[\varepsilon_i]$	$E[r_i]$	$\alpha_i / \text{Std}[\varepsilon_i]$	$\text{Var}[\varepsilon_i]$	$\text{Var}[r_i]$	$\alpha_i / \text{Var}[r_i]$	π_i
X	0.05	1	0.6	0.12	0.08333	0.36	0.4	0.1389	0.3210
Y	0.05	1.5	0.5	0.15	0.1	0.25	0.34	0.2	0.4623
Z	0.06	0.5	0.8	0.1	0.075	0.64	0.65	0.0938	0.2167

Table 13.1: Information on the mispriced assets in Example 13.1.

The left part of the table lists relevant inputs on the three mispriced assets in the example. The right part shows values derived using the Treynor-Black approach.

**Figure 13.1:** The mean-variance diagram with Treynor-Black.

The diagram is constructed using the inputs explained in Example 13.1.

Table 13.1 presents the relevant information on these stocks. As derived in the right part of the table, the active portfolio has weights 0.3210 in X, 0.4623 in Y, and 0.2167 in Z. Although Z has the highest alpha, its idiosyncratic risk is so high that Z has the lowest weight of the three mispriced stocks.

The active portfolio has $\alpha_A = 0.0522$, $\beta_A = 1.1228$, and $\text{Std}[\varepsilon_A] = 0.3472$. It follows from (13.3) that the active portfolio should have a weight of $w_A^* = 0.2990$ and the market portfolio a weight of $1 - w_A^* = 0.7010$. This combination has an expected rate of return of 0.0878, a standard deviation of 0.2319, and a Sharpe ratio of 0.3355. Compared to a full investment in the market portfolio, the combination with the active portfolio increases both the expected rate of return, the standard deviation, and the Sharpe ratio.

Figure 13.1 illustrates the situation in the usual mean-variance diagram. By combining the mispriced stocks with the market portfolio, you obtain a steeper tangency line. A mean-variance optimizer should therefore combine the riskfree asset and the tangency portfolio which is the optimal combination of the market portfolio and the active portfolio.

When all the mispriced assets have positive alphas, it is clear from (13.5) that they will have positive weights in the active portfolio. Unless the beta of the active portfolio is very large, we see from (13.3) that the active portfolio then has a positive weight in the optimal

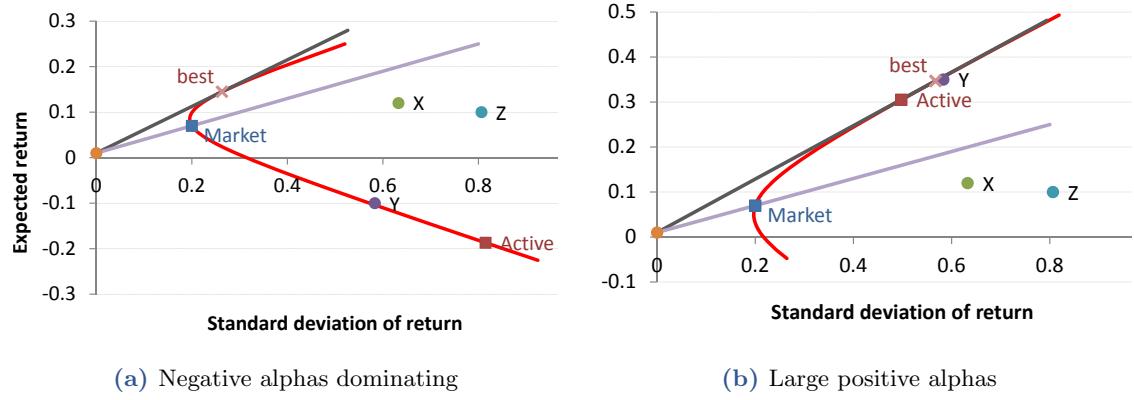


Figure 13.2: Further mean-variance diagrams with Treynor-Black.

The left diagram is constructed using the inputs explained in Example 13.2, whereas the right diagram comes from Example 13.3.

combination with the market portfolio.

If some of the mispriced assets have negative alphas, these results might change. In particular, if $\sum_j (\alpha_j / \text{Var}[\varepsilon_j])$ is negative, then Eq. (13.5) implies that assets with negative alpha have a positive weight and assets with positive alpha have a negative weight in the optimal active portfolio. At first, this seems counter intuitive. But in this case the alpha of the active portfolio

$$\alpha_A = \sum_{i=1}^J \pi_i \alpha_i = \sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]} \sum_{j=1}^J \frac{\alpha_j}{\text{Var}[\varepsilon_j]}$$

is negative as the numerator is positive and the denominator is negative. Hence, you want to short the active portfolio—which is consistent with (13.3)—and therefore you end up shorting the assets with negative alphas and taking long positions in the assets with positive alphas. This situation is illustrated in the following example.

Example 13.2

Suppose the situation is exactly as described in Example 13.1, except that stock Y has an alpha of -20% instead of 5% . Then the total expected return on Y is -10% .

In this case the active portfolio has weights -0.2448 in X, 1.4100 in Y, and -0.1652 in Z. The active portfolio has $\alpha_A = -0.3042$, $\beta_A = 1.7876$, and $\text{Std}[\varepsilon_A] = 0.7322$. It follows from (13.3) that the active portfolio should have a weight of $w_A^* = -0.2914$ and the market portfolio a weight of $1 - w_A^* = 1.2914$. More precisely, the Treynor-Black approach tells you that the tangency portfolio consists of investing 7.1% of the wealth in stock X, -41.1% in stock Y, 4.8% in stock Z, and 129.1% in the market portfolio. This combination has an expected rate of return of 0.1449 , a standard deviation of 0.2632 , and a Sharpe ratio of 0.5124 . The left panel of Figure 13.2 shows the mean-variance diagram for this case.

If the mispriced assets have high alphas relative to their residual risks, the optimal weight of the active portfolio stated in (13.3) might exceed one, in which case the weight of the market portfolio is negative. In words, the optimal portfolio would involve shorting

the market portfolio. Here is an example:

Example 13.3

Suppose the situation is exactly as described in Example 13.1, except that stock Y has an alpha of 25% instead of 5%. It is not unusual that stock analysts think that some stock is undervalued by 25%. For example, an analyst may conclude from a fundamental analysis of a company that its stock is worth \$7.50 per share, but the current market price is only \$6.00, so the analyst must expect the price to increase by 25% of its current value.

Now the optimal weight in the active portfolio is 1.1774 and, consequently, the optimal weight in the market portfolio is -0.1774 . More precisely, the Treynor-Black approach tells you that the tangency portfolio consists of 13.3% in stock X, 95.5% in stock Y, 9.0% in stock Z, and -17.7% in the market portfolio. The expected return on the tangency portfolio is a lustrous 34.7%, but the standard deviation is as high as 56.7%. Obviously, this is a very extreme and risky position. The right panel of Figure 13.2 shows the mean-variance diagram for this example.

Example 13.3 illustrates a general challenge with the Treynor-Black method. It often results in a very risky portfolio with extreme positions in some assets. You could try to impose some constraints on portfolio weights and in particular require that the weight on the market portfolio must stay non-negative, implying that the weight on the active portfolio is bounded by 100%.

Many investors prefer not to deviate too much from the market portfolio, i.e., they care about the tracking error, the difference between the return on their portfolio and the return on the market portfolio (or some other benchmark portfolio). With the Treynor-Black procedure, the tracking error is

$$\begin{aligned} \text{TE} &= r_p - r_m = w_A r_A + (1 - w_A) r_m - r_m = w_A(r_A - r_m) \\ &= w_A ([r_A - r_f] - [r_m - r_f]) = w_A (\alpha_A + (\beta_A - 1)(r_m - r_f) + \varepsilon_A), \end{aligned}$$

where the final equality is due to (13.1). Since the residual return component is independent from the market return, the standard deviation of the tracking error is

$$\text{Std[TE]} = |w_A| \sqrt{(\beta_A - 1)^2 \text{Var}[r_m] + \text{Var}[\varepsilon_A]}. \quad (13.9)$$

This quantifies the active investor's benchmark risk. If the investor has a maximum benchmark risk, this translates into a maximum value of the active weight w_A . Possibly, a different active portfolio is now optimal. Note that the term tracking error is often used directly for the standard deviation Std[TE] .

As discussed in Section 11.2, it is very difficult to detect statistically significant alphas from time-series regressions of individual stock returns on market returns. We will explore this further in Section 13.3 below. It is also difficult for analysts to forecast non-zero alphas from other information, so it may very well be that their predicted alphas turn out to be wrong. Then you should be cautious taking extreme positions based on those predictions. If you have a series of alpha estimates from the same analyst, you can check his prediction record, and use that to adjust his current alpha predictions. [Treynor and Black \(1973\)](#) also suggested a procedure for how to adjust the alphas. For an analyst who is not consistently too pessimistic or optimistic, and thus right on average, you should shrink all

his alpha estimates towards zero. Since alphas are difficult to predict, you should shrink the predictions substantially. More specifically, you should calculate the R^2 (typically very low) in the regression of the realized alphas on his predicted alphas, and then multiply his future predictions of alpha by that R^2 . This generally reduces the optimal weight on the active portfolio and therefore the problem of too extreme positions. If the analyst tends to be too optimistic by predicting too high alphas, of course you should generally lower all his predicted alphas. Conversely, if the analyst tends to be pessimistic.

13.1.2 Treynor-Black in multi-factor models

We can generalize the Treynor-Black approach to multi-factor models in which all the factors are returns on specific portfolios so that the excess returns on individual assets can be decomposed as

$$r_i - r_f = \alpha_i + \sum_{k=1}^K \beta_{ik} (r_{Fk} - r_f) + \varepsilon_i = \alpha_i + \boldsymbol{\beta}_i \cdot (\mathbf{r}_F - r_f \mathbf{1}) + \varepsilon_i, \quad (13.10)$$

cf. Eq. (11.44). Here $\boldsymbol{\beta}_i$ is a column vector of the factor-betas $\beta_{i1}, \dots, \beta_{iK}$ of asset i and \mathbf{r}_F is a column vector of the returns on the factor portfolios. If the exact version of the Arbitrage Pricing Theory holds, all assets should have $\alpha_i = 0$. If this is the case, then the portfolio maximizing the Sharpe ratio is the tangency portfolio constructed from the factor portfolios as explained in Section 11.7.

As in the previous subsection, suppose we have identified J assets with a non-zero alpha relative to the K -factor model. We form an active portfolio of these assets with weights π_1, \dots, π_J so that the excess return of the active portfolio satisfies

$$r_A - r_f = \alpha_A + \boldsymbol{\beta}_A \cdot (\mathbf{r}_F - r_f \mathbf{1}) + \varepsilon_A,$$

where

$$\alpha_A = \sum_{i=1}^J \pi_i \alpha_i, \quad \boldsymbol{\beta}_A = \sum_{i=1}^J \pi_i \boldsymbol{\beta}_i, \quad \varepsilon_A = \sum_{i=1}^J \pi_i \varepsilon_i.$$

We combine the active portfolio and the factor portfolios and wants to maximize the (squared) Sharpe ratio of the combined portfolio. Let w_A denote the weight of the active portfolio and let \mathbf{w}_F denote the vector of weights in the factor portfolios. Of course, we need $w_A + \mathbf{w}_F \cdot \mathbf{1} = 1$. The next theorem shows both the optimal weights in this combination and the optimal composition of the active portfolio.

Theorem 13.3

Assume returns follow a multi-factor model where all factors are returns on specific portfolios. Then the maximum Sharpe ratio is obtained by the weights

$$w_A^* = \frac{\frac{\alpha_A}{\text{Var}[\varepsilon_A]}}{\mathbf{1} \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) + (1 - \boldsymbol{\beta}_A \cdot \mathbf{1}) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}}, \quad (13.11)$$

$$\mathbf{w}_F^* = \frac{\underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) - \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \boldsymbol{\beta}_A}{\mathbf{1} \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) + (1 - \boldsymbol{\beta}_A \cdot \mathbf{1}) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}}. \quad (13.12)$$

The squared Sharpe ratio of this specific combination satisfies

$$(\text{SR}_p)^2 = (\text{SR}_F)^2 + \left(\frac{\alpha_A}{\text{Std}[\varepsilon_A]} \right)^2, \quad (13.13)$$

where $(\text{SR}_F)^2 = (\boldsymbol{\mu}_F - r_f \mathbf{1}) \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1})$ is the maximal squared Sharpe ratio when trading only in the factor portfolios.

The optimal active portfolio is given by

$$\pi_i = \frac{\frac{\alpha_i}{\text{Var}[\varepsilon_i]}}{\sum_{j=1}^J \frac{\alpha_j}{\text{Var}[\varepsilon_j]}}, \quad i = 1, 2, \dots, J, \quad (13.14)$$

and has the property that

$$\frac{\alpha_A^2}{\text{Var}[\varepsilon_A]} = \sum_{i=1}^J \frac{\alpha_i^2}{\text{Var}[\varepsilon_i]} = \sum_{i=1}^J \left(\frac{\alpha_i}{\text{Std}[\varepsilon_i]} \right)^2. \quad (13.15)$$

Note, in particular, that the optimal composition of the active portfolio stated in the latter part of Theorem 13.3 is identical to that given in Theorem 13.2, but of course the α 's and residual variances in the two theorems refer to different models.

Proof

We calculate the tangency portfolio constructed from the set of factor portfolios and the active portfolio. Recall that the tangency portfolio weights are given by the vector $\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})$ divided by the sum of its elements which is $\mathbf{1} \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})$. In the multi-factor model, we have

$$\boldsymbol{\mu} - r_f \mathbf{1} = \begin{pmatrix} \boldsymbol{\mu}_F - r_f \mathbf{1} \\ \alpha_A + \boldsymbol{\beta}_A \cdot (\boldsymbol{\mu}_F - r_f \mathbf{1}) \end{pmatrix}$$

and

$$\underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_F & \underline{\Sigma}_F \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_A^\top \underline{\Sigma}_F & \boldsymbol{\beta}_A^\top \underline{\Sigma}_F \boldsymbol{\beta}_A + \sigma_{\varepsilon, A}^2 \end{pmatrix}, \quad \underline{\Sigma}^{-1} = \begin{pmatrix} \underline{\Sigma}_F^{-1} + \frac{1}{\sigma_{\varepsilon, A}^2} \boldsymbol{\beta}_A \boldsymbol{\beta}_A^\top & -\frac{1}{\sigma_{\varepsilon, A}^2} \boldsymbol{\beta}_A \\ -\frac{1}{\sigma_{\varepsilon, A}^2} \boldsymbol{\beta}_A^\top & \frac{1}{\sigma_{\varepsilon, A}^2} \end{pmatrix}$$

similarly to Eqs. (11.54) and (11.55) except for the α_A in the expectation.

Multiplication now leads to

$$\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) = \begin{pmatrix} \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) - \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \boldsymbol{\beta}_A \\ \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \end{pmatrix},$$

and the sum of the elements are

$$\mathbf{1} \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) + (1 - \boldsymbol{\beta}_A \cdot \mathbf{1}) \frac{\alpha_A}{\text{Var}[\varepsilon_A]}.$$

Dividing the elements of $\underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1})$ with the sum of all elements, we get the weights

w_F^* and w_A^* stated in the theorem. The squared Sharpe ratio of the tangency portfolio is

$$\begin{aligned} (\boldsymbol{\mu} - r_f \mathbf{1}) \cdot \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}) &= \left(\frac{\boldsymbol{\mu}_F - r_f \mathbf{1}}{\alpha_A + \boldsymbol{\beta}_A \cdot (\boldsymbol{\mu}_F - r_f \mathbf{1})} \right) \cdot \left(\frac{\underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) - \frac{\alpha_A}{\text{Var}[\varepsilon_A]} \boldsymbol{\beta}_A}{\frac{\alpha_A}{\text{Var}[\varepsilon_A]}} \right) \\ &= (\boldsymbol{\mu}_F - r_f \mathbf{1}) \cdot \underline{\Sigma}_F^{-1} (\boldsymbol{\mu}_F - r_f \mathbf{1}) + \frac{\alpha_A^2}{\text{Var}[\varepsilon_A]} \\ &= (\text{SR}_F)^2 + \left(\frac{\alpha_A}{\text{Std}[\varepsilon_A]} \right)^2 \end{aligned}$$

which shows (13.13).

The proof of the rest of the theorem is identical to the proof of Theorem 13.2.

13.2 The Black-Litterman model

The key contribution of the Black-Litterman model is to allow the investor to build her own views about asset returns into Markowitz' mean-variance portfolio choice model of Chapter 7. The model was introduced in 1990 by Fischer Black and Robert Litterman in a Goldman Sachs internal research note, and the model was later explained (barring some important calculations) in a published paper, cf. [Black and Karasinski \(1992\)](#), and more rigorously in the Goldman Sachs note by [He and Litterman \(1999\)](#). First, we outline the basic idea and procedure. Then we put the method into perspective and provide a comprehensive example.

13.2.1 The basic procedure

The output from the model is an optimal portfolio which as in previous chapters can be characterized by a vector $\boldsymbol{\pi}$ of portfolio weights in the risky assets with the understanding that the fraction $1 - \boldsymbol{\pi}^\top \mathbf{1}$ of wealth is then invested in the riskfree asset. In Markowitz' mean-variance model, the optimal portfolio vector is given by

$$\boldsymbol{\pi}^* = \frac{1}{\gamma} \underline{\Sigma}^{-1} (\boldsymbol{\mu} - r_f \mathbf{1}), \quad (13.16)$$

cf. Equation (7.56). Here γ is the investor's relative risk aversion, $\underline{\Sigma}$ is the variance-covariance matrix of asset returns, and $\boldsymbol{\mu} - r_f \mathbf{1}$ is the vector of excess expected returns on the risky assets. This expression for $\boldsymbol{\pi}$ is based on a linear mean-variance tradeoff in the investor's preferences. The entire mean-variance approach relies on returns being normally distributed.

In line with the mean-variance approach, the Black-Litterman model assumes that the *excess* rates of return on the N risky assets considered are normally distributed, $\mathbf{r}_x \sim N(\boldsymbol{\mu}_x, \underline{\Sigma})$ for some mean vector $\boldsymbol{\mu}_x$ and some variance-covariance matrix $\underline{\Sigma}$. Of course, with a riskfree rate of r_f , the expected return vector is then $\boldsymbol{\mu} = \boldsymbol{\mu}_x + r_f \mathbf{1}$, but the model focuses directly on excess returns as they ultimately determine the optimal portfolio. Since the riskfree rate is certain, there is no difference between the variance-covariance matrix of the excess returns and of the returns themselves.

A key feature of the Black-Litterman model is to recognize that the true values of $\boldsymbol{\mu}_x$ and $\underline{\Sigma}$ are generally unknown. The variances and covariances in $\underline{\Sigma}$ do typically not seem

to vary much over time and can be estimated quite precisely from return time series, so let us just assume that we know the true $\underline{\Sigma}$. In contrast, the expected excess return vector μ_x is much more difficult to estimate precisely, cf. the discussion in Section 3.7, and expected returns are more likely to change over time due to variations in the market-wide risk premiums and the market sensitivity (betas) of the individual stocks. We have some *prior* point estimate \mathbf{m} for μ_x but realize that this is only an estimate.

Prior. We can think of the true unknown mean as being equal to our point estimate plus a deviation:

$$\mu_x = \mathbf{m} + \boldsymbol{\varepsilon}.$$

If we assume the deviation is normally distributed, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \underline{V})$, then $\mu_x \sim N(\mathbf{m}, \underline{V})$. Here the matrix \underline{V} captures the uncertainty about the precision of our point estimate. Taking this uncertainty about μ_x into account, the overall return variance-covariance matrix is³

$$\text{Var}[\mathbf{r}] = \text{Var}[\mathbf{r}_x] = \underline{\Sigma} + \underline{V}.$$

With a linear mean-variance tradeoff and a relative risk aversion of γ , the vector of optimal unconstrained portfolio weights is then

$$\boldsymbol{\pi}^* = \frac{1}{\gamma} (\underline{\Sigma} + \underline{V})^{-1} \mathbf{m}.$$

cf. Equation (13.16). A common assumption is that $\underline{V} = \tau \underline{\Sigma}$ for some constant scalar $\tau \geq 0$, and then

$$\text{Var}[\mathbf{r}] = (1 + \tau) \underline{\Sigma}, \quad \boldsymbol{\pi}^* = \frac{1}{\gamma(1 + \tau)} \underline{\Sigma}^{-1} \mathbf{m}.$$

The prior mean estimate \mathbf{m} is assumed to be based on some objective procedure. One obvious approach is to use a time series of past returns. Then \mathbf{m} is simply the vector of arithmetic average excess returns on the assets. With a single risky asset, the standard deviation of the average excess return is equal to the sample standard deviation of the asset's excess returns divided by the square root of the number of return observations, cf. the discussion and the confidence interval for μ shown in Section 3.7.2. In other words, the variance of the mean estimate is the sample variance divided by the number of observations. Similarly, with multiple assets the variance-covariance matrix of the mean vector equals the sample variance-covariance matrix of the returns divided by the number of observations. These considerations can justify setting $\underline{V} = \tau \underline{\Sigma}$ with $\tau = 1/T$ where T is the number of periods in the data sample. For example, with $T = 100$ observations, τ equals 0.01. In fact, 0.01 appears to be a value commonly used in practice, whether or not the prior mean estimate is based on 100 time periods.

Due to the difficulties in generating precise and reliable estimates for expected returns from past average returns, Black and Litterman suggested to use CAPM-implied estimates for expected excess returns. Section 10.1.6 already explained how to derive these estimates and concluded that

$$\mathbf{m} = \bar{\gamma} \underline{\Sigma} \boldsymbol{\pi}_{\text{mkt}} \tag{13.17}$$

is the vector of expected excess returns that the average investor must assume in order to find the current market portfolio weights optimal, cf. Equation (10.26). Here, $\boldsymbol{\pi}_{\text{mkt}}$ is the vector of market portfolio weights and $\bar{\gamma}$ is the average risk aversion in the market, which

³We can write the excess return vector as $\mathbf{r}_x = \mu_x + \boldsymbol{\varepsilon}_r$, where $\boldsymbol{\varepsilon}_r \sim N(\mathbf{0}, \underline{\Sigma})$. Then $\mathbf{r}_x = \mathbf{m} + \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}_r$ gives $\text{Var}[\mathbf{r}_x] = \text{Var}[\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}_r] = \underline{\Sigma} + \underline{V}$, provided that $\text{Cov}[\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_r] = \underline{0}$ which we assume holds.

can be determined as

$$\bar{\gamma} = \frac{E[r_m] - r_f}{\text{Var}[r_m]}.$$

This procedure does not provide information about the matrix \underline{V} that captures the precision of the mean vector estimate, but often the assumption $\underline{V} = \tau \underline{\Sigma}$ is applied also in this setting. With this approach, the optimal portfolio based solely on the prior is

$$\boldsymbol{\pi}^* = \frac{1}{\gamma(1+\tau)} \underline{\Sigma}^{-1} \mathbf{m} = \frac{\bar{\gamma}}{\gamma(1+\tau)} \boldsymbol{\pi}_{\text{mkt}}. \quad (13.18)$$

This optimal portfolio is a simple scaling of the market portfolio and is sometimes referred to as the *neutral portfolio* in the context of the Black-Litterman model. For $\tau = 0$, the scaling is determined by the ratio of the average investor's risk aversion to your own risk aversion. If you are more risk averse than the average investor, you should reduce the weights of the market portfolio accordingly and invest part of your wealth in the riskfree asset. If you are less risk averse than the average investor, you should lever up the market portfolio. These conclusions follow from the CAPM as already explained in Section 10.1.4. With $\tau > 0$, the portfolio weights are dampened to account for the uncertainty about the expected returns.

Of course, the true market portfolio includes all risky assets. The Black-Litterman approach considers only a limited set of risky assets. The market portfolio weights in the above expressions are the weights relative to the market value of only the included risky assets, just as in Example 10.3. As also shown in that example, the exclusion or addition of one asset may change the implied expected returns for the other assets considered due to correlation structure embedded in $\underline{\Sigma}$.

Views. The Black-Litterman model allows you to combine the objective prior for the expected excess returns with a set of subjective *views* on the returns. The combination leads to a posterior estimate $\hat{\mathbf{m}}$ of the vector of expected excess returns and an associated precision from which the optimal portfolio $\hat{\boldsymbol{\pi}}_{\text{BL}}^*$ is calculated using the standard formula. Each view is a prediction for the excess return on one of the risky assets or for a linear combination of excess returns on the risky assets. Of course, perfect prediction is impossible and we should account for the prediction error.

Suppose you have K views. For each $k = 1, 2, \dots, K$, view number k is represented by an equation of the form

$$P_{k1}\mu_{x1} + \dots + P_{kN}\mu_{xN} = Q_k + \varepsilon_{vk}.$$

Here μ_{xi} is the expected excess rate of return on asset i , and P_{k1}, \dots, P_{kN} are constants. Hence, the left-hand side is a linear combination of the expected excess returns on the assets. On the right-hand side, Q_k is the view you have on this combination, whereas the mean-zero variable ε_{vk} is the error related to this view.

Together, the views can be written as

$$\underline{P} \boldsymbol{\mu}_x = \mathbf{Q} + \boldsymbol{\varepsilon}_v, \quad (13.19)$$

where \underline{P} is a $K \times N$ matrix, \mathbf{Q} is a K -vector, and $\boldsymbol{\varepsilon}_v \sim N(\mathbf{0}, \underline{\Omega})$ is the prediction error with some $\bar{K} \times K$ variance-covariance matrix $\underline{\Omega}$. Note that according to the objective prior, the expected excess returns on the portfolios represented by \underline{P} are given by $\underline{P} \mathbf{m}$. The subjective views twist these expectations to \mathbf{Q} and thus modify expectations by $\mathbf{Q} - \underline{P} \mathbf{m}$.

The matrix $\underline{\Omega}$ is assumed to be diagonal, i.e. of the form

$$\underline{\Omega} = \begin{pmatrix} \omega_1 & 0 & 0 & \dots & 0 \\ 0 & \omega_2 & 0 & \dots & 0 \\ 0 & 0 & \omega_3 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & & \dots & \omega_K \end{pmatrix}$$

and we assume that all $\omega_k > 0$. The zeros off the diagonal mean that the views are assumed to be independent of each other. This is not necessarily realistic, in particular when some assets enter several views. Another implicit assumption is that the error terms ε and ε_v are independent.

For example, one view might be that the excess return on the first asset is going to be 0.05, and another view that the return of asset 2 is expected to be 4 percentage points larger than the return of asset 3. The first view is an *absolute view*, the second a *relative view*. These views can be written as

$$1 \times \mu_{x1} = 0.05, \quad 1 \times \mu_{x2} - 1 \times \mu_{x3} = 0.04.$$

If the three assets are the only assets considered, the two views correspond to

$$\underline{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \quad \underline{Q} = \begin{pmatrix} 0.05 \\ 0.04 \end{pmatrix}.$$

As in this simple example, each row in \underline{P} typically has elements summing either to one so that the row represents a portfolio or to zero in which case the row represents a zero-investment or long-short portfolio.

Obviously, $\underline{\Omega}$ is reflecting the investor's confidence in the views, but exactly how should $\underline{\Omega}$ be specified? One approach is to think in terms of confidence intervals for each view. The first view above is that the excess return on the first asset is expected to be $\mu_{x1} = 0.05$. Maybe the investor can formulate her confidence in this view by saying that with a probability of $p \in (0,1)$ the excess return is in a symmetric interval around μ_{x1} , i.e. $[\mu_{x1} - \Delta, \mu_{x1} + \Delta]$ for some $\Delta > 0$. Then we can back out a corresponding variance from the normal distribution, which turns out to be

$$\omega_1 = \left(\frac{\Delta}{N^{-1} \left(\frac{1+p}{2} \right)} \right)^2. \quad (13.20)$$

This procedure would thus generate the diagonal elements of $\underline{\Omega}$. A second approach is to use the track record of the investor: how precise have her views been in the past? Yet another approach is simply to assume that the variance of the views is proportional to the variance of the returns and for example let $\underline{\Omega}$ be equal to $\underline{P} \underline{V} \underline{P}^\top$ and overwrite off-diagonal elements by zero.

Posterior. Because of your views, you revise your prior estimate of expected excess returns. It follows from the rules of Bayesian statistics that the *posterior* distribution for the vector μ_x of expected excess returns is a normal distribution with mean

$$\hat{\mathbf{m}} = \mathbf{m} + \underline{V} \underline{P}^\top (\underline{\Omega}^{-1} + \underline{P} \underline{V} \underline{P}^\top)^{-1} (\underline{Q} - \underline{P} \mathbf{m}) \quad (13.21)$$

and variance

$$\hat{V} = \left(\underline{\underline{V}}^{-1} + \underline{\underline{P}}^T \underline{\underline{\Omega}}^{-1} \underline{\underline{P}} \right)^{-1}. \quad (13.22)$$

Note that the updated mean $\hat{\mathbf{m}}$ equals the old mean plus a term involving the subjective twist $\mathbf{Q} - \underline{\underline{P}}\mathbf{m}$ to the expected excess returns on the view portfolios. It is instructive to think of a couple of extreme cases. If the prior is already infinitely precise, i.e. $\underline{\underline{V}} = \underline{\underline{0}}$, then $\hat{\mathbf{m}} = \mathbf{m}$ so that the prior is not updated. Conversely, if the posterior is infinitely precise, i.e. $\underline{\underline{\Omega}} = \underline{\underline{0}}$, then $\underline{\underline{P}}\hat{\mathbf{m}} = \mathbf{Q}$, i.e. the posterior mean fully respects the views. Given the posterior distribution for $\boldsymbol{\mu}_x$, the overall variance of the return is revised to

$$\text{Var}[\mathbf{r}] = \underline{\underline{\Sigma}} + \hat{V}. \quad (13.23)$$

Optimal portfolio. Based on the posterior distribution for the expected excess returns, the vector of optimal unconstrained portfolio weights is

$$\hat{\boldsymbol{\pi}}_{\text{BL}}^* = \frac{1}{\gamma} (\underline{\underline{\Sigma}} + \hat{V})^{-1} \hat{\mathbf{m}}. \quad (13.24)$$

In the presence of portfolio constraints, a numerical optimization can be implemented in Excel exactly as explained in Chapter 7 just using the posterior estimates of the return moments.

It is not clear from Eq. (13.24) exactly how the views affect the optimal portfolio. Assuming that $\underline{\underline{V}} = \tau \underline{\underline{\Sigma}}$ and that the CAPM-implied prior is used, He and Litterman (1999) show that the portfolio can be restated as

$$\hat{\boldsymbol{\pi}}_{\text{BL}}^* = \frac{1}{1+\tau} \left(\boldsymbol{\pi}^* + \sum_{k=1}^K \lambda_k \mathbf{p}_k \right), \quad (13.25)$$

where $\boldsymbol{\pi}^*$ is given by (13.18), \mathbf{p}_k is the vector corresponding to the k 'th row in $\underline{\underline{P}}$, and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)^\top$ is given by

$$\boldsymbol{\lambda} = \frac{\tau}{\gamma} \underline{\underline{\Omega}}^{-1} \mathbf{Q} - \frac{1}{1+\tau} \underline{\underline{A}}^{-1} \underline{\underline{P}} \underline{\underline{\Sigma}} \boldsymbol{\pi}^* - \frac{\tau}{(1+\tau)\gamma} \underline{\underline{A}}^{-1} \underline{\underline{P}} \underline{\underline{\Sigma}} \underline{\underline{P}}^T \underline{\underline{\Omega}}^{-1} \mathbf{Q}. \quad (13.26)$$

Here, $\underline{\underline{A}} = \underline{\underline{\Omega}}/\tau + \underline{\underline{P}} \underline{\underline{\Sigma}} \underline{\underline{P}}^T/(1+\tau)$. The effect of each view portfolio \mathbf{p}_k on the final portfolio is determined by the coefficient λ_k . He and Litterman (1999) interpret each of the three terms in $\boldsymbol{\lambda}$. Concerning the first term, note that $\underline{\underline{\Omega}}^{-1} \mathbf{Q} = (q_1/\omega_1, \dots, q_K/\omega_K)^\top$, where q_k is the k 'th element of \mathbf{Q} . Hence, we see from the first term in $\boldsymbol{\lambda}$ that λ_k is larger, the stronger the view is in terms of either a higher q_k or a lower uncertainty ω_k . The second term in $\boldsymbol{\lambda}$ reflects a covariance between the view portfolio and the market portfolio. If this covariance is large, the view carries less new information and thus has a smaller effect on the final portfolio, as indicated by the minus in front of the term. The third term in $\boldsymbol{\lambda}$ reflects the covariance of each view with the other view portfolios. If, for example, one view portfolio is highly correlated with other view portfolios, we should be careful not to “double count” the view, and thus reduce the separate impact of each view.

13.2.2 Discussion

The Bayesian updating approach is typically applied in an experimental setting. A researcher has some best guess about a given variable or parameter based on objective facts, previous research, or respected theories, but knows that there is some uncertainty about

whether the guess is correct. Hence, he specifies a prior distribution for the variable with a mean equal to the best guess and a variance whose reciprocal indicates the precision of the best guess. Subsequently, the researcher makes additional observations, maybe based on an experiment, that induce an update of the prior distribution to a posterior distribution for the variable in question. In the Black-Litterman model, the “experiment” is just a formulation of the investor’s views, but the updating procedure is applied in the same way. In a more standard experimental setting, the precision of the experiment (the inverse of $\underline{\Omega}$) often follows from the number of additional observations made. In the Black-Litterman model, it is less obvious how to specify $\underline{\Omega}$. It is also not obvious how to fix the precision of the prior (the inverse of \underline{V}) if the prior estimate of expected returns is implied by the CAPM. These are some weak points of the approach.

Let us briefly compare the Black-Litterman model and the Treynor-Black model. The latter includes subjective views on individual assets in terms of non-zero alphas and such views can also be captured by that Black-Litterman model. If, for example, we think that only a single asset i has a non-zero alpha, our investment universe consists of the market portfolio and asset i . The view can then be represented by

$$\mathbf{m} = \begin{pmatrix} \mathbb{E}[r_m] - r_f \\ \beta_i (\mathbb{E}[r_m] - r_f) \end{pmatrix}, \quad \underline{P} = (0 \ 1), \quad Q = \alpha_i + \beta_i (\mathbb{E}[r_m] - r_f), \quad Q - \underline{P}\mathbf{m} = \alpha_i.$$

The Black-Litterman model allows more complex views than the Treynor-Black model and it allows to account for uncertainty about both the prior and the view.

13.2.3 An example

Consider an investor wanting to invest in only the riskfree asset and five stocks over the next month. The five stocks are Tesla (ticker symbol TSLA), Amazon (AMZN), Netflix (NFLX), Walmart (WMT), and Exxon Mobile (XOM). Table 13.2 shows the arithmetic average and the variance-covariance matrix of the excess returns on the five stocks based on 60 monthly observations from July 2015 to June 2020. Note that both Tesla, Amazon, and Netflix have offered large returns in the period with a monthly average of more than 3%, but also with substantial volatilities especially for Tesla with a monthly standard deviation of 0.1570 corresponding to an annualized volatility of $\sqrt{12} \times 0.1570 \approx 54\%$ (ignoring compounding). The table also shows the market capitalization of each company as of July 3, 2020, from which we calculate the market weights of the five stocks relative to the total value of only these companies, not the total market. Then, applying Equation (13.17), we back out which expected excess returns these stocks must have for the average investor to optimally hold this portfolio. Here we use an average risk aversion of $\bar{\gamma} = 2$ which is consistent with an expected excess market return of 0.005 (6% annualized without compounding) and a market volatility of 0.05 (annualized $\sqrt{12} \times 0.05 \approx 17.3\%$), in line with historical levels. We can see that the CAPM-implied expected excess return is larger than the historical average for Exxon Mobile but smaller for the other four stocks, in particular for Tesla, Amazon, and Netflix.

Table 13.3 shows optimal portfolios based on various inputs for an investor with a risk aversion of 2. The column labeled ‘Historical’ gives the optimal portfolio if the empirical estimates are used without modifications. The weights on Amazon and Walmart both exceed 2 primarily due to their large realized Sharpe ratios (Walmart’s modest average return is accompanied by a rather low volatility). Also Tesla and Netflix have positive weights, whereas a significant short position is taken in Exxon Mobile. Overall the weights sum to much more than 1, indicating that the portfolio involves substantial leverage through

Stock	Average	Variance-covariance matrix						Market info		
		TSLA	AMZN	NFLX	WMT	XOM	Cap	π_{mkt}	\mathbf{m}	
TSLA	0.0336	0.02465	0.00327	0.00294	0.00049	0.00215	224.2	0.0934	0.0095	
AMZN	0.0340	0.00327	0.00765	0.00567	0.00059	0.00244	1442.0	0.6008	0.0113	
NFLX	0.0315	0.00294	0.00567	0.01236	0.00016	0.00113	209.7	0.0874	0.0098	
WMT	0.0114	0.00049	0.00059	0.00016	0.00276	0.00089	337.6	0.1407	0.0017	
XOM	-0.0050	0.00215	0.00244	0.00113	0.00089	0.00527	186.5	0.0777	0.0046	

Table 13.2: Objective inputs to the Black-Litterman model.

The averages, variances, and covariances are based on observed excess returns in each of the 60 months from July 2015 to June 2020. The excess returns are calculated from adjusted closing stock prices from Yahoo Finance and one-month riskfree rates from Professor French's homepage, all accessed on July 3, 2020. (The June 2020 riskfree rate was not yet included in French's data series and is assumed to be identical to the May 2020 value.) The market capitalization of each company (listed in billions of USD) was also taken from Yahoo Finance on July 3, 2020. The market weight of each company is simply its market capitalization divided by the total market capitalization of the five companies. The estimate \mathbf{m} for the expected excess returns are implied from the CAPM assuming an average risk aversion of $\bar{\gamma} = 2$ which is consistent with an expected excess market return of 0.005 and a market volatility of 0.05.

a short position in the riskfree asset. The column labeled ‘Implied’ shows the optimal portfolio based on the CAPM-implied expected returns. Since our investor is assumed to have the same risk aversion as the average investor, her optimal portfolio is given by the relative market weights and the riskfree asset is not used. If we take the uncertainty about the mean estimate into account by using $\underline{V} = \tau \underline{\Sigma}$ and $\tau = 1/60$, we obtain the optimal portfolio in the column ‘BL-neutral’. This is the optimal portfolio according to the Black-Litterman approach in the absence of any subjective views, and is simply a slightly scaled-down version of the market portfolio accompanied by a small position in the riskfree asset.

Suppose now that our investor has two subjective views on the five stocks. The first view is that Tesla is over-valued and the investor expects the monthly return to be one percentage point lower than the current market weight implies. As the implied expected excess return on Tesla is 0.0095, the view is that the expected excess return is really -0.0005 . The second view is that Amazon is under-valued relative to Netflix. More specifically, the difference between the return on Amazon and the return on Netflix next year is expected to be two percentage points larger than what the market weights currently imply. According to the market-implied estimates, the expected return difference is $0.0113 - 0.0098 = 0.0016$ (all numbers rounded to four decimals), the view is that the return difference is expected to be 0.0216. The views thus correspond to

$$\underline{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \end{pmatrix}, \quad \underline{Q} = \begin{pmatrix} -0.0005 \\ 0.0216 \end{pmatrix}.$$

To translate these views into posterior estimates and optimal portfolios, we have to specify the matrix $\underline{\Omega}$ of uncertainty about the views. We consider two ways to do that. The first is based on confidence intervals on the views, the second is simply to assume that $\underline{\Omega}$ equals $\underline{P} \underline{V} \underline{P}^\top$ with zeros imposed off the diagonal. For the first approach, suppose that the investor thinks that there is a 90% probability that the underperformance of Tesla is between zero and two percentage points so that the expected excess return is between

	Historical	Implied	BL-neutral	$\underline{\Omega}$ choice 1		$\underline{\Omega}$ choice 2	
				\hat{m}	BL-tilted	\hat{m}	BL-tilted
TSLA	0.4870	0.0934	0.0919	0.0004	-0.1021	0.0047	-0.0109
AMZN	2.3656	0.6008	0.5910	0.0137	1.5142	0.0129	1.1663
NFLX	0.2437	0.0874	0.0859	-0.0038	-0.8373	0.0014	-0.4894
WMT	2.1657	0.1407	0.1384	0.0023	0.1384	0.0021	0.1384
XOM	-2.1890	0.0777	0.0764	0.0062	0.0764	0.0057	0.0764
Sum	3.0730	1.0000	0.9836		0.7896		0.8809

Table 13.3: Outputs from the Black-Litterman model.

The left part of the table lists the optimal portfolio for an investor with a risk aversion of $\gamma = 2$, when inputs are determined from either historical return moments, CAPM-implied expected returns, or the Black-Litterman model without any subjective views. For each of two choices of the view uncertainty matrix $\underline{\Omega}$, the right part of the table shows the resulting posterior estimate of expected excess returns and the optimal portfolio. The first choice of $\underline{\Omega}$ is based on confidence intervals for the views as explained in the text, the second choice is the diagonal version of the matrix $\underline{P} \underline{V} \underline{P}^\top$.

–0.0105 and 0.0005. Applying (13.20) with $p = 0.9$ and $\Delta = 0.01$, this corresponds to a variance of 3.7×10^{-5} which is then the upper-left element of $\underline{\Omega}$. Suppose that the confidence regarding the second view is similar, i.e. that the investor is 90% sure that the extraordinary return difference between Amazon and Netflix is in the range from one to three percentage points. This leads to the lower-right element in $\underline{\Omega}$ also being equal to 3.7×10^{-5} . For the second approach, the diagonal elements of $\underline{\Omega}$ are an order of magnitude larger: 4.1×10^{-4} in the upper left element and 1.4×10^{-4} in the lower right.

For both choices of $\underline{\Omega}$, the optimal portfolio is clearly tilted as you would expect given the subjective views. The weight of Tesla is reduced from roughly 9% to either –10% or –1%; the change is relatively small due to the large variance of Tesla returns. The weight of Amazon is increased from 59% to either 151% or 117%. For Netflix, the neutral weight of about 9% is replaced by a significant short position of either –84% or –49%. Walmart and Exxon Mobile are not involved in the subjective views and their weights remain the same. The views have a smaller impact with the second choice of $\underline{\Omega}$ as this has larger elements and thus corresponds to a lower precision of the views.

The example includes both an absolute and a relative view. In general, relative views affect only the relative portfolio weights, whereas absolute views change the sum of the weights. For example, if only the absolute view on Tesla is included, the optimal weight on Tesla changes from 9.19% to 8.58%, while the weights of the other assets remain as they are in the neutral portfolio. In contrast, if only the relative view on Amazon and Netflix is included, the optimal weight of Amazon increases by 0.99 and the weight of Netflix decreases by 0.99 relative to the neutral portfolio, while the weights of the three other stocks remain unchanged. In these calculations, we specify $\underline{\Omega}$ from the investor's confidence interval on the view (choice 1 above).

13.3 Performance evaluation

13.3.1 Common performance measures

Suppose a portfolio manager over a sequence of T periods of equal length has obtained rates of return of r_{p1}, \dots, r_{pT} . How can we evaluate the performance of the portfolio

manager from these returns? Did he perform better or worse than one could expect? Obviously we need to adjust for the risk he took.

Various frequently used measures of investment performance have already been introduced in earlier chapters. The **Sharpe ratio** of an asset or portfolio was defined in (3.19) in terms of the expectation and standard deviation of the future return. The empirical counterpart is

$$\text{SR} = \frac{\bar{r}_p - \bar{r}_f}{\sigma_p}, \quad (13.27)$$

where \bar{r}_p is the (arithmetic) average and σ_p is the sample standard deviation,

$$\bar{r}_p = \frac{1}{T} \sum_{t=1}^T r_{pt}, \quad \sigma_p = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (r_{pt} - \bar{r}_p)^2}.$$

Likewise, \bar{r}_f is the average riskfree rate over the period considered. Sometimes the Sharpe ratio is calculated as

$$\text{SR} = \frac{\overline{r_p - r_f}}{\sigma[r_p - r_f]},$$

where $\overline{r_p - r_f}$ and $\sigma[r_p - r_f]$ denote the average and sample standard deviation of the time series $r_{p1} - r_{f1}, r_{p2} - r_{f2}, \dots, r_{pT} - r_{fT}$ of portfolio returns in excess of the periodic riskfree rate. Obviously, the average of the return differences $r_{pt} - r_{ft}$ is identical to the difference between the average portfolio returns \bar{r}_p and the average riskfree returns \bar{r}_f , i.e. $\overline{r_p - r_f} = \bar{r}_p - \bar{r}_f$, but the standard deviation $\sigma[r_p - r_f]$ of the return differences typically deviates slightly from $\sigma_p = \sigma[r_p]$. The Sharpe ratio compares excess returns to the total risk of the investment as measured by the standard deviation or volatility.

The numerical value of the Sharpe ratio might be difficult to interpret. How much better is a Sharpe ratio of 0.4 than a Sharpe ratio of 0.3? Modigliani and Modigliani (1997) advocated the use of a related performance measure which is easier to interpret. While they dubbed it RAP for risk-adjusted performance, it is now often referred to as **M2** or **M-squared** due to the authors' names. The measure is calculated as

$$M_p^2 = \bar{r}_p \times \frac{\sigma_m}{\sigma_p} + \bar{r}_f \left(1 - \frac{\sigma_m}{\sigma_p}\right), \quad (13.28)$$

where σ_m is the sample standard deviation of the market return. The idea behind this measure is to adjust the portfolio return for the risk taken. The portfolio p is combined with the riskfree asset in such a way that the combined portfolio has the same standard deviation as the market portfolio. This is obtained by taking a weight of $w = \sigma_m/\sigma_p$ on portfolio p and a weight of $1-w$ on the riskfree asset, cf. (4.21). Being the average return on this combined portfolio, M^2 can be directly compared to the return on the market portfolio as they share the same standard deviation. Sometimes the terms M2 and M-squared are used for the return difference $M_p^2 - \bar{r}_m$ relative to the market portfolio or the excess return $M_p^2 - \bar{r}_f$ relative to the riskfree asset. If the Sharpe ratio of your portfolio and the market portfolio are calculated by using (13.27), it can be shown that

$$M_p^2 - \bar{r}_m = \sigma_m (\text{SR}_p - \text{SR}_m), \quad (13.29)$$

which shows the link between the M-squared and the Sharpe ratio (Exercise 13.2 asks for a proof of this result).

If the portfolio considered represents the entire investment of the investor, then the

standard deviation of the portfolio return is an appropriate risk measure and, hence, both the Sharpe ratio and the M2 of the portfolio are suitable performance measures. This is not the case if the portfolio you want to evaluate is one out of many portfolios you have invested in. Then only the systematic risk of each portfolio is relevant, since most of the unsystematic risk is diversified away across portfolios. In this case, the **Treynor ratio** defined in (10.16) is a frequently used portfolio performance measure. The empirical version of the measure is

$$\text{TR}_p = \frac{\bar{r}_p - \bar{r}_f}{\beta_p},$$

where β_p is the sample beta of the portfolio over the investment horizon considered, which is estimated as the slope in a regression of excess portfolio returns $r_{pt} - r_{ft}$ on excess market returns $r_{mt} - r_{ft}$. Again, the numerator can be replaced by the average excess return, $\bar{r}_p - \bar{r}_f$. The Treynor ratio compares excess returns to the systematic risk taken.

The most frequently used portfolio performance measure is **Jensen's alpha**, which is often just called the alpha. It measures the part of the return that is unexplained by the systematic risk taken. If we use the CAPM to capture the systematic risk, the alpha is calculated as

$$\alpha_p = \bar{r}_p - (\bar{r}_f + \beta_p [\bar{r}_m - \bar{r}_f]) = \bar{r}_p - \bar{r}_f - \beta_p \times \bar{r}_m + \beta_p \bar{r}_f, \quad (13.30)$$

which is the empirical (or ex-post) version of the theoretical (or ex-ante) alpha defined in (10.17). This coincides with the intercept estimate in a linear regression of the excess portfolio returns on the excess market returns. The alpha can also be calculated with reference to a different equilibrium model than the CAPM, for example one of the Fama-French multi-factor models. Note that we can rewrite the Sharpe ratio of the portfolio as

$$\text{SR}_p = \frac{\alpha_p + \beta_p(\bar{r}_m - \bar{r}_f)}{\sigma_p} = \frac{\alpha_p}{\sigma_p} + \frac{\beta_p \sigma_m}{\sigma_p} \text{SR}_m = \frac{\alpha_p}{\sigma_p} + \rho_{pm} \text{SR}_m, \quad (13.31)$$

where ρ_{pm} is the correlation between portfolio returns and market returns and we have used (10.5). Since the correlation is less than one, we can see that just because the portfolio has a positive alpha, it does not necessarily have a higher Sharpe ratio than the market. By deviating from the market portfolio, your diversification is reduced, and your alpha has to more than compensate for that before the overall Sharpe ratio improves.

The **information ratio** was defined in (11.56) as the ratio of the expected alpha to the standard deviation of the non-systematic return component. The empirical counterpart is

$$\text{IR} = \frac{\alpha_p}{\sigma[\varepsilon_p]}, \quad (13.32)$$

where α_p is calculated as above. Furthermore, $\sigma[\varepsilon_p]$ is the sample standard deviation of the residual returns

$$\varepsilon_{pt} = r_{pt} - r_{ft} - \alpha_p - \beta_p [r_{mt} - r_{ft}],$$

where α_p and β_p are the estimates from the regression of the excess portfolio returns $r_{pt} - r_{ft}$ on the excess market returns $r_{mt} - r_{ft}$ (with an intercept in the regression).

Example 13.4

You are considering investing some money in the investment fund FatReturns, but first you want to evaluate the fund's performance record. Its realized returns over the most recent years average 22% per year with a standard deviation of 40%. The average riskfree return was 2% per year. By regressing the fund returns on the market returns, you estimate the beta of the fund to be 1.5. The annual market return has been 12% on average with a standard deviation of 18%. Has FatReturns performed well?

The Sharpe ratio of FatReturns is $\frac{22\%-2\%}{40\%} = 0.5$, which is somewhat lower than that of the market portfolio, which is $\frac{12\%-2\%}{18\%} \approx 0.56$. The M^2 of the portfolio is $22\% \times \frac{0.18}{0.4} + 2\% \times (1 - \frac{0.18}{0.4}) = 11\%$ is lower than the 12% return on the market portfolio. Both these performance measures indicate that investing all of your savings in the FatReturns fund is rather unattractive.

With a beta of 1.5, the average fund return should have been $2\% + 1.5 \times (12\% - 2\%) = 17\%$ according to the CAPM. Hence, FatReturns have delivered an alpha of 5% per year, which is quite impressive. However, this alpha comes with a substantial unsystematic risk. If we decompose the total fund return as in the Single-Index Model, cf. (11.17), the standard deviation σ_e of the unsystematic return component can be computed as

$$(40\%)^2 = (1.5)^2 \times (18\%)^2 + \sigma_e^2 \Rightarrow \sigma_e = \sqrt{(40\%)^2 - (1.5)^2 \times (18\%)^2} \approx 29.51\%.$$

The information ratio of the fund is $5\%/29.51\% \approx 0.169 = 16.9\%$. Seen as your only investment, FatReturns is unattractive because of the high unsystematic risk compared to the alpha. But seen as one out of many funds you invest in, FatReturns is quite attractive as most of that unsystematic risk will be diversified away. The Treynor ratio is $\frac{22\%-2\%}{1.5} \approx 13.33\%$, which is bigger than that of the market portfolio (10%), indicating that you are more than compensated for the systematic risk.

13.3.2 Performance measures and benchmarks

Several of the portfolio performance measures defined above are in some way comparing the return on the portfolio to that of the market portfolio. For some investors or portfolio managers, a different portfolio or index may be the relevant benchmark.

Some performance measures can be adapted to this case. For example, we can compute the M-squared as

$$M^2 = \bar{r}_p \times \frac{\sigma_b}{\sigma_p} + \bar{r}_f \left(1 - \frac{\sigma_b}{\sigma_p}\right), \quad (13.33)$$

where the b -subscript indicates the benchmark. This is identical to our earlier definition (13.28) except that we scale the portfolio's standard deviation to the standard deviation σ_b of the benchmark instead of the standard deviation σ_m of the market. We can then compare this M^2 to the average return \bar{r}_b on the benchmark portfolio and use the difference $M^2 - \bar{r}_b$ as a measure of the investor's performance.

13.3.3 Statistical significance

No matter which of these performance measures we apply, we face a problem of limited statistical significance. As explained in Section 3.7, the mean of a return series is very difficult to estimate precisely due to the large variation in returns, even when we have return observations for many periods. If we want to evaluate the performance of a certain

portfolio manager following a specific investment strategy, we often have relatively few observations, which makes the average return very little informative. Even if a portfolio manager achieves a high average return, it may not be significantly better than the average return of the appropriate benchmark. In other words, we cannot tell whether the portfolio manager was lucky or skillful. This problem carries over to the other performance measures. For example, while you almost always find a non-zero alpha if you regress a return series on the market returns, the alpha is rarely statistically significantly different from zero.

Example 13.5

Suppose that over the past 60 months you have followed a certain investment strategy. By regressing your monthly excess returns relative to the one-month riskfree rate on the monthly excess returns on the market portfolio, you estimate your alpha to be $\alpha_p = 0.0025$ or 0.25% per month, corresponding to an annual excess return of 3% without compounding. This is quite impressive, but can you decide whether your performance is due to skill or luck? Suppose the residuals in the regression have a modest standard deviation of 0.04 or 4%, i.e., $\sigma[\varepsilon_p] = 0.04$. Then the standard error of the alpha-estimate is

$$\sigma_\alpha = \frac{\sigma[\varepsilon_p]}{\sqrt{T}} = \frac{0.04}{\sqrt{60}} \approx 0.005164.$$

Consequently, the t -statistic for the alpha-estimate is

$$t(\alpha) = \frac{\alpha}{\sigma_\alpha} = \frac{0.0025}{0.005164} \approx 0.4841.$$

Since this is clearly less than 2.0017, which is the 97.5% percentile of the t -distribution with $T - 2 = 58$ degrees of freedom, your alpha-estimate is not significantly different from zero in a statistical sense.

With the 60 observations and the assumed residual standard deviation, the alpha-estimate would have to exceed 1% per month to be statistically significant.

On the other hand, suppose that we fix the estimates of alpha and the residual risk. How many observations do we need for the alpha to be statistically significant? When we vary T , the 97.5% percentile of the t -distribution varies somewhat, but it stays very close to 2. Hence, we see that

$$t(\alpha) = \frac{0.0025}{0.04/\sqrt{T}} > 2 \Leftrightarrow \sqrt{T} > \frac{2 \times 0.04}{0.0025} = 32 \Leftrightarrow T > (32)^2 = 1024,$$

that is, we need 1024 monthly observations and thus more than 85 years of observations before the alpha-estimate would be statistically significant.

The way in which performance measures are typically interpreted and how their statistical significance is evaluated implicitly assumes that all return observations are independent draws from the same probability distribution, in particular a distribution with a constant mean and variance. For a rather stable or passive investment strategy, this may be a not-too-bad assumption although earlier chapters have discussed some evidence on auto-correlation in returns. However, if you have been varying your investment strategy substantially during the data period, it makes little sense to calculate performance

measures based on the full period.

13.4 ESG investing

13.4.1 What is ESG and how is it measured?

What does ESG stand for? (Picture source: Anevis Solutions)



ESG reporting by companies. Mandatory reporting?

- EU: starting 2024, the Corporate Sustainability Reporting Directive (CSRD) requires large companies to publish detailed ESG information in a standardized way
- US: currently no mandatory ESG reporting but discussing a new climate disclosure rule focused on greenhouse gas emissions and environmental impact

Both institutional investors and the companies they invest in may engage in *greenwashing*, e.g. by overstating their environmental credentials, or create the impression of an environmental benefit where one doesn't exist.

Close cousins: Social washing (appearing socially conscious), rainbow washing (appearing diverse), pinkwashing (appearing LGBT friendly), bluewashing (singing up for UN principles without acting upon them), ...

Avoiding greenwashing? Common reporting standards. Third-party assessments (ratings).

ESG ratings:

- Many mutual funds invest according to ESG ratings
- Often separate rating on E, S, and G, plus combined ESG rating
- ESG ratings are offered by numerous providers (e.g. Sustainalytics, Moody's ESG, S&P Global, Refinitiv, MSCI) who are paid by investors using the ratings, not the companies being rated, unlike credit ratings
- Ratings from different providers disagree substantially, see, e.g., Berg, Kölbel, and Rigobon (2022)
 - Scope: different attributes included in ratings (explains 38% of rating divergence)
 - Measurement: same attributes, different indicators (56%)
 - Weight: same indicators, different weights (6%)

- Need more harmonization and transparency in ESG reporting?
- Often used alternative to E-rating: CO₂ emissions, either total or relative to sales ('emission intensity')
 - Backward-looking reports vs. forward-looking improvement potential

13.4.2 What is ESG investing?

- Investor takes ESG considerations into account when making investment decisions
- 'ESG' in ESG investing can be replaced by *sustainable, responsible, impact, green, SRI, CSR, ...*
- Started in 1960s with some investors excluding "sin stocks" (i.e., stocks of companies profiting from tobacco, alcohol, weapons, or gambling) or stocks of companies active in certain countries (South Africa's apartheid regime)
- The term ESG was first mentioned in the 2006 UN Principles for Responsible Investment (UNPRI)
- Big increase in ESG investing in recent years

How to make an impact through investing:

- Exit or voice
 - Exit: divest from "badly behaving" companies
 - Voice: stay invested and use voting rights to change bad behavior
- Most research finds "voice" to be more effective than "exit," see, e.g., Berk and van Binsbergen (2024)

Approaches to ESG investing:

- Active ownership (\approx voice)
- Bottom-up ESG integration: systematic and explicit inclusion of ESG risks and opportunities in investment analysis
- Top-down ESG integration: systematic and explicit inclusion of ESG factors in portfolio construction
- Best-in-class selection: Prefer companies with good or improving ESG profiles relative to sector peers
- Positive/negative screening: include/exclude companies with good/bad ESG profile

Why engage in ESG investing?

- Do good for society
 - Reduced investments in bad-ESG companies should increase their cost of capital and thus make their bad activities more costly to finance
 - Conversely, increased investments in good-ESG companies should lower their cost of capital and thus lead to more "good" investments
 - Obtain better returns or risk-return tradeoffs
 - Maybe returns are just higher for good-ESG companies?
 - Maybe companies' ESG scores predict future earnings and dividends, i.e. good scores signal quality?
 - Maybe good-ESG companies have lower long-term risks?
- See, e.g., Hoepner, Oikonomou, Sautner, Starks, and Zhou (2024).

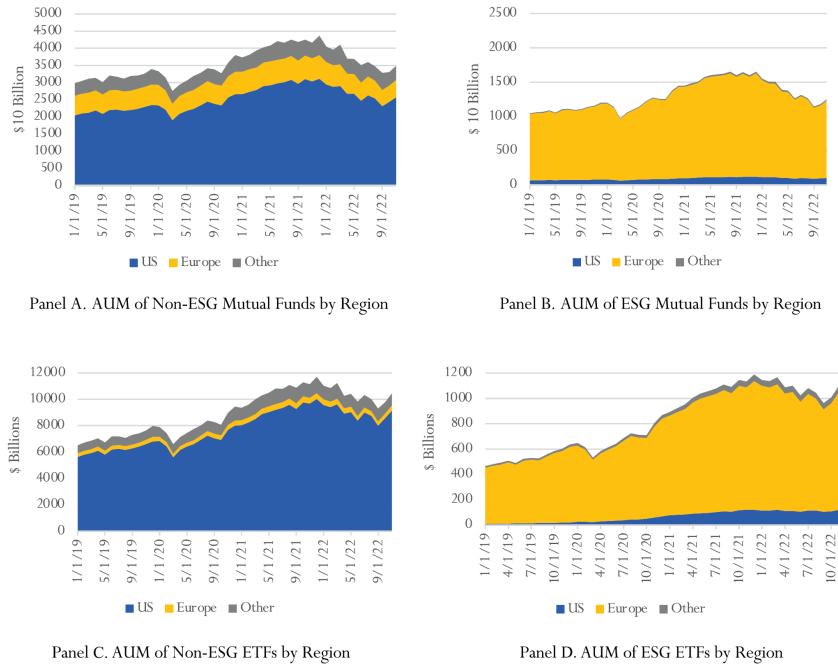


Figure 13.3: The extent of ESG investing.

This is Figure 4 from Starks (2023).

13.4.3 How big is ESG investing?

Figure 13.3 shows how the assets under management (AUM) in mutual funds and ETFs have evolved from 2019 to 2022 divided into non-ESG funds (Panels A and C) and ESG funds (Panels B and D). The blue colored areas represent the US, the yellow areas represent Europe, and the gray areas the rest of the world. First of all, we see that ESG funds have a much larger share of the market in Europe. Measured by AUM, European non-ESG funds are far smaller than US non-ESG funds, but European ESG funds are far bigger than US ESG funds. The AUM of ESG funds (especially for the ETFs) has been growing more than the AUM of the non-ESG funds.

More than 5000 investment companies worldwide have signed the UNPRI by Quarter 1 of 2023, representing \$121tn of AUM. These signatories must report on their responsible investment activities.

ESG investing is expected to continue to grow. According to Bloomberg, ESG investing reached \$35 tn in global AUM in 2020 and is expected to reach \$50 tn in 2025, more than 1/3 of total global AUM.

The European Union continues to push for investors and investment companies to consider ESG issues. For example, EU's MiFID rules (Markets in Financial Instruments Directive) require investment companies to consider sustainability when developing products and to assess retail clients' ESG preferences (in addition to the clients' risk tolerance and investment objectives).

Note that the ESG agenda is experiencing considerable pushback in the US. As of late 2023, 15 Republican-governed states have introduced anti-ESG legislation, e.g., banning ESG investing in public retirement funds. The Republican-dominated House of Rep's passed a bill banning retirement fund managers to consider ESG, but the bill was vetoed by President Biden.

Figure 13.4 shows, for a range of countries, the average environmental score of the companies in each country.

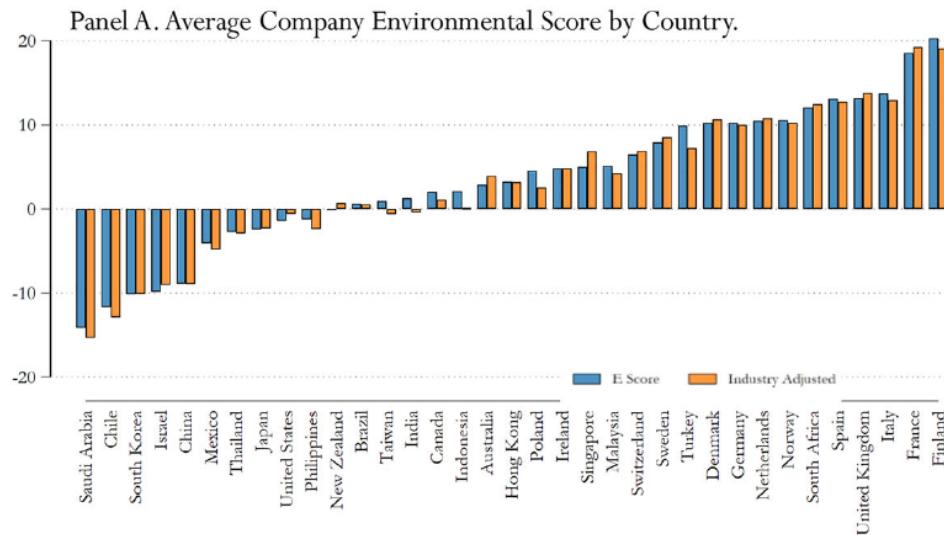


Figure 13.4: Companies' environmental score across countries.

This is Figure 5 from Starks (2023).

While these AUM numbers are huge, they also exaggerate the actual extent of ESG investing. For example, BlackRock is the world's largest asset management company and signed the UNPRI in 2008. However, only a fraction of its \$9trn AUM is invested based on ESG criteria. It is more relevant to look how much BlackRock changes its portfolio due to ESG considerations.

Pastor, Stambaugh, and Taylor (2024) investigate the actual ESG-tilt or green tilt of large US investment companies. For 2021, for example, their data set covers 3,086 companies with a total AUM of \$31.3trn. 76% of the AUM in their sample belongs to institutions having signed UNPRI. The data includes the individual stock holdings of the investment companies from 2012 to 2021. The authors represent the greenness of each stock by the MSCI score, and they compare the weight of each stock in their portfolio with the stock's market weight. By combining this information, the authors can measure the green tilt of the investment company's portfolio at a given point in time. Obviously, green investors should generally over-weight stocks with good ESG scores and under-weight stocks with bad ESG scores, which would generate a positive green tilt by the authors' measure. Note that, by construction, if some investors have a green tilt, other investors must have a brown (i.e., non-green) tilt as the tilt comes from deviating from market weights. For each stock, MSCI provides separate scores for E, S, and G, as well as a combined ESG, and a portfolio tilt can thus be calculated based on either of the four scores. Among the findings of the study are:

- The total dollar ESG-related tilt is only 6% of industry's AUM in equity investments in 2021
- Most investors with a green tilt do not exclude brown stocks from their portfolios, they just reduce the weights of these stocks in the portfolio
- E, S, G contribute almost equally to ESG-tilts
- Most ESG tilting is done by the largest 1/3 of the investment companies
- Investment companies may deviate from the market portfolio for reasons not related to ESG. The ESG tilts comprise around 1/4 of the total portfolio tilts away from market portfolio in 2021
- After signing UNPRI, investors tilt more towards green stocks

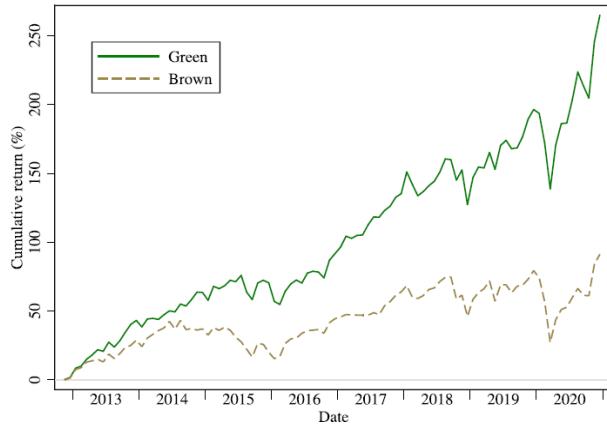


Figure 13.5: Cumulative returns on a green and a brown portfolio.

This is Figure 3 from Pastor, Stambaugh, and Taylor (2022).

13.4.4 The returns on ESG investments

Theory

Good-ESG assets should have lower expected returns than bad-ESG assets because of

- 1 Direct preference: Some investors have direct preference for good-ESG assets and pay higher prices for such assets which leads to lower expected return
- 2 Hedging: Good-ESG assets hedge against ESG-risks so, if ESG-risks are important, all risk-averse investors should tilt towards good-ESG assets. This will drive up the prices of such stocks and thus lower the future expected returns if expected dividends stay unchanged. Hoepner, Oikonomou, Sautner, Starks, and Zhou (2024) report empirical evidence supporting the view that institutional investors' engagement on ESG issues reduce firms' downside risk, in particular risk related to climate change.

Good-ESG assets can have higher realized returns in a transition phase with increasing ESG awareness either among investors (directly increasing asset prices) or among consumers (driving up profits of good-ESG companies), see, e.g., van der Beck (2021), Pastor, Stambaugh, and Taylor (2021), and the discussion in Section 6.1.3.

Empirical evidence: Realized returns

Pastor, Stambaugh, and Taylor (2022) compute a greenness score for individual US stocks based on MSCI E-ratings and E-weights. They consider the period from November 2012 to December 2020. They form a green-stock [brown-stock] portfolio of the 1/3 of stocks with the highest [lowest] greenness. The portfolios are regularly rebalanced. They find that over the roughly 8-year period the green portfolio outperforms the brown portfolio by $265 - 91 = 174$ percentage points, cf. Figure 13.5, which is a large and significant outperformance. They show that the Green-Minus-Brown strategy has a significant, positive α relative to CAPM, FF3, FF5.

Note that growth stocks are generally greener than value stocks. Hence, the increasing investor E-awareness may explain why growth stocks have outperformed value stocks in the recent decade.

Empirical evidence: Expected returns

In the dividend discount model presented in Section 6.1, the price of a stock equals the present value of the discounted future dividends, and the discount rate equals the

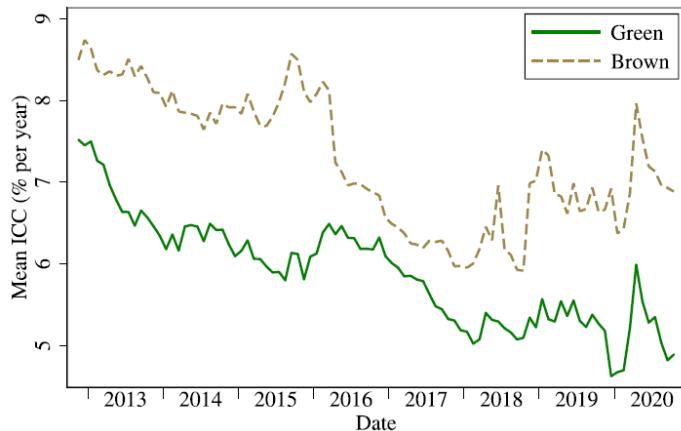


Figure 13.6: Implied cost of capital on green stocks and brown stocks.
This is Figure 4 from Pastor, Stambaugh, and Taylor (2022).

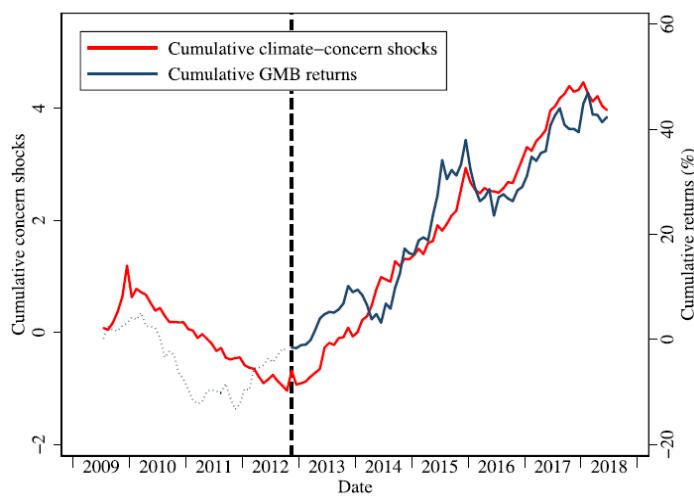


Figure 13.7: Climate concerns and Green-Minus-Brown returns.
This is Figure 6 from Pastor, Stambaugh, and Taylor (2022).

expected return of the stock. Pastor, Stambaugh, and Taylor (2022) thus estimate each stock's expected return as its implied cost of capital (ICC), the discount rate that equates the stock's current price to the present value of expected future cash flows. Figure 13.6 shows that the average ICC is lower among green stocks than brown stocks. The difference varies between 0.4% and 2.4% per year, with an average of 1.4% across all years.

Empirical evidence: Expected vs. realized returns

Pastor, Stambaugh, and Taylor (2022) also find that positive realized returns on the Green-Minus-Brown strategy are highly correlated with increasing climate change concerns. This is illustrated in Figure 13.7. Here, the climate change concerns are measured by the Media Climate Change Concerns index developed by Ardia, Bluteau, Boudt, and Inghelbrecht (2023) which is based on climate concerns raised in news articles in 8 major US newspapers.

Further empirical evidence

Brøgger and Kronies (2022) compare returns to two investor types:

1. Strict investors: banks, insurance companies, pension funds with strict ESG mandates focus on current, backward-looking ESG scores
2. Flexible investors: investment companies, hedge funds are more flexible and can focus on expected future ESG score increases

They find that

- Strict investors earn up to 3.1% per year less than flexible investors
- Return gap is higher in times of rising climate sentiment (based on Google searches for ‘Climate change’)
- Flexible investors buy stocks before ESG improvements are realized and then strict investors enter and drive up prices
- ESG investor mandates should focus on forward-looking ESG scores

13.4.5 A model for ESG investing

This section presents the model of Pedersen, Fitzgibbons, and Pomorski (2021) which can be seen as an ESG-extension of Markowitz’ mean-variance model of Chapter 7 and the CAPM equilibrium model of Chapter 10. Hence, we consider a one-period setting with the following assumptions about the assets available to investors:

- a riskfree asset with rate of return r_f
- n risky assets with excess rates of return $r = (r_1, \dots, r_n)^\top$ and ESG scores $s = (s_1, \dots, s_n)^\top$
- Excess returns have unconditional expectation $E[r]$ and variance-covariance matrix $\text{Var}[r]$
- Conditional on s , excess returns have expectation $\mu = E[r|s]$ and variance-covariance matrix $\Sigma = \text{Var}[r|s]$

We denote by $x = (x_1, \dots, x_n)^\top$ a vector of portfolio weights in the risky assets, so that the weight of the riskfree asset is $1 - x^\top \mathbf{1}$, and the rate of return on wealth given a portfolio x is $r_f + x^\top r$.

The model features three types of investors:

- Unaware investor unaware of (or uninterested in) ESG; solves

$$\max_{x \in X} \left\{ x^\top E[r] - \frac{\gamma}{2} x^\top \text{Var}[r] x \right\}$$

where X denotes the set of possible portfolios. This is the standard mean-variance problem from Section 7.3, so we know that the solution is to combine the tangency portfolio and the riskfree asset.

- Aware investor: applies ESG scores for estimation; solves

$$\max_{x \in X} \left\{ x^\top \mu - \frac{\gamma}{2} x^\top \Sigma x \right\}$$

This is again the standard mean-variance problem from Section 7.3, so the solution is to combine the tangency portfolio and the riskfree asset. However, in this case the tangency portfolio is based on the ESG-conditional moments.

- Motivated investor: applies ESG scores for estimation and has preference for high ESG scores; solves

$$\max_{x \in X} \left\{ x^\top \mu - \frac{\gamma}{2} x^\top \Sigma x + f(\bar{s}) \right\}$$

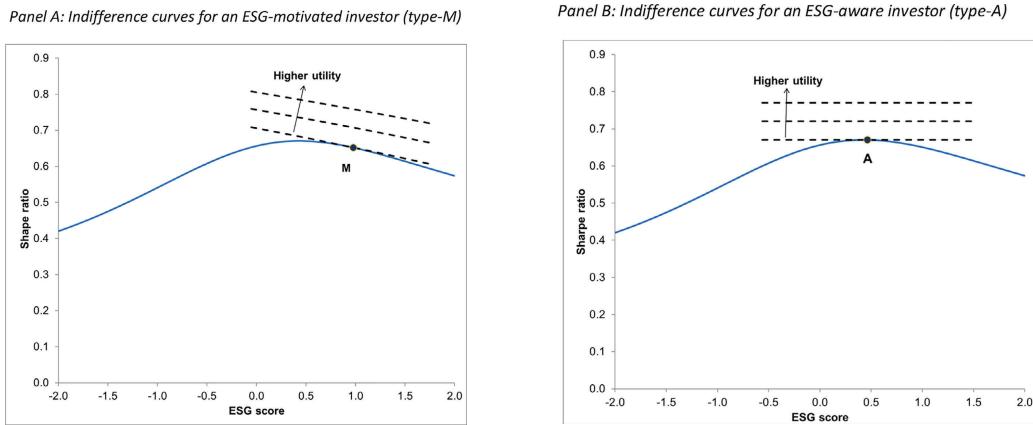


Figure 13.8: The ESG-efficient frontier and investor indifference curves.

This is Figure 3 from Pedersen, Fitzgibbons, and Pomorski (2021).

where $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty\}$ is the ESG preference function, and $\bar{s} = x^\top s / x^\top \mathbf{1}$ is the portfolio's weighted ESG score scaled by total position in risky assets (no ESG-utility from riskfree investments).

Portfolio choice of motivated investors

Recall from standard mean-variance analysis that the tangency portfolio has the maximum Sharpe ratio among all portfolios, and any combination of the tangency portfolio and the riskfree asset has the same Sharpe ratio as the tangency portfolio. Any mean-variance investor picks some combination of the tangency portfolio and the riskfree asset. We can thus think of the investor first identifying the SR-maximizing (tangency) portfolio of risky assets, and then maximizing the mean-variance tradeoff by picking the weight w_{tan} on that portfolio or, equivalently, the standard deviation $\sigma = w_{\text{tan}}\sigma_{\text{tan}}$ of the position:

$$\max_{w_{\tan}} \left\{ w_{\tan} \underbrace{(\mu_{\tan} - r_f)}_{= \sigma_{\tan} \text{SR}_{\tan}} - \frac{\gamma}{2} w_{\tan}^2 \sigma_{\tan}^2 \right\} \quad \sim \quad \max_{\sigma} \left\{ \sigma \times \text{SR}_{\tan} - \frac{\gamma}{2} \sigma^2 \right\}$$

The ESG-motivated investor is willing to accept a lower Sharpe ratio if the ESG score of the portfolio is improved sufficiently. The preferences are illustrated in Figure 13.8. The blue curve in the figure is the ESG-SR frontier. This frontier is generated in the following way: For every given ESG score \bar{s} , we find the portfolio that maximizes the Sharpe ratio among all portfolios with the ESG score \bar{s} . In mathematical terms, the problem to be solved is

$$\max_{x \in X} \left\{ \frac{x^\top \mu}{\sqrt{x^\top \Sigma x}} \right\} \quad \text{s.t.} \quad x^\top \mathbf{1} = 1 \quad \text{and} \quad x^\top s = \bar{s}$$

Then we can solve the motivated investor's problem as

$$\max_{\bar{s}} \left\{ \max_{\sigma} \left\{ \sigma \times \text{SR}(\bar{s}) - \frac{\gamma}{2} \sigma^2 + f(\bar{s}) \right\} \right\} \quad (13.34)$$

The first-order condition with respect to σ is $\text{SR}(\bar{s}) - \gamma\sigma = 0$, which implies $\sigma = \text{SR}(\bar{s})/\gamma$,

and then the inner maximization in (13.34) becomes

$$\sigma \times \text{SR}(\bar{s}) - \frac{\gamma}{2}\sigma^2 + f(\bar{s}) = \frac{\text{SR}(\bar{s})}{\gamma} \times \text{SR}(\bar{s}) - \frac{\gamma}{2} \left(\frac{\text{SR}(\bar{s})}{\gamma} \right)^2 + f(\bar{s}) = \frac{1}{2\gamma} \text{SR}(\bar{s})^2 + f(\bar{s}).$$

Note that the value of \bar{s} that maximizes $\frac{1}{2\gamma} \text{SR}(\bar{s})^2 + f(\bar{s})$ will also maximize $\text{SR}(\bar{s})^2 + 2\gamma f(\bar{s})$. This leads to part (a) of the following theorem. We skip the proof of part (b).

Theorem 13.4

(a) The motivated investor should pick the \bar{s} solving

$$\max_{\bar{s}} \left\{ (\text{SR}(\bar{s}))^2 + 2\gamma f(\bar{s}) \right\} \quad (13.35)$$

(b) The maximum Sharpe Ratio, $\text{SR}(\bar{s})$, attainable with an ESG score of \bar{s} is

$$\text{SR}(\bar{s}) = \sqrt{c_{\mu\mu} - \frac{(c_{s\mu} - \bar{s}c_{1\mu})^2}{c_{ss} - 2\bar{s}c_{1s} + \bar{s}^2 c_{11}}} \quad (13.36)$$

where $c_{ab} = a^\top \Sigma^{-1} b \in \mathbb{R}$ for any two vectors a, b .

Across all portfolios, the maximum SR is

$$\text{SR}(s^*) = \sqrt{c_{\mu\mu}} \quad \text{where} \quad s^* = \frac{c_{s\mu}}{c_{1\mu}} = \frac{s^\top \Sigma^{-1} \mu}{1^\top \Sigma^{-1} \mu}. \quad (13.37)$$

Note that the first term in (13.35) depends only on assets, the second term depends only on preferences. Hence, the ESG-SR frontier is determined independently of investor preferences.

The next theorem characterizes the portfolio with the largest Sharpe ratio among all portfolios with an ESG score of \bar{s} .

Theorem 13.5

Given the ESG score \bar{s} , the Sharpe ratio is maximized by the portfolio

$$x = \frac{1}{\gamma} \Sigma^{-1} (\mu + \pi (s - \mathbf{1} \bar{s})), \quad (13.38)$$

where π is the constant scalar

$$\pi = \frac{c_{1\mu} \bar{s} - c_{s\mu}}{c_{ss} - 2c_{1s} \bar{s} + c_{11} \bar{s}^2}. \quad (13.39)$$

Note that the portfolio x combines the tangency portfolio $\Sigma^{-1} \mu / 1^\top \Sigma^{-1} \mu$, the “ESG-tangency portfolio” $\Sigma^{-1} s / 1^\top \Sigma^{-1} s$, and the minimum-variance portfolio $\Sigma^{-1} \mathbf{1} / \mathbf{1}^\top \Sigma^{-1} \mathbf{1}$. Since x is combined with the riskfree asset, we have a four-fund separation result.

To find the optimal portfolio for a given ESG-motivated investor, we can apply Theorem 13.5 with \bar{s}^* , the investor’s optimal ESG score from Theorem 13.4. We can think of

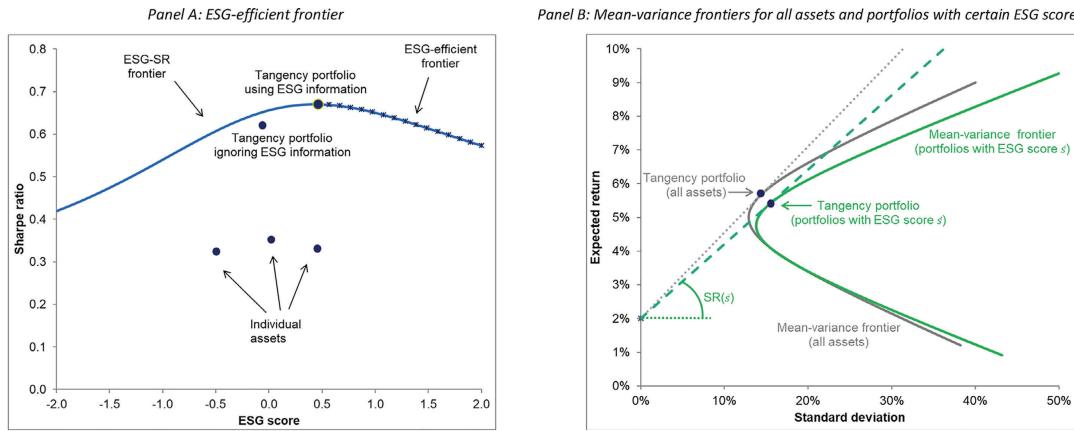


Figure 13.9: The ESG-efficient frontier and mean-variance frontiers.

This is Figure 1 from Pedersen, Fitzgibbons, and Pomorski (2021).

the constant π as a scaling factor indicating the strength of the investor's ESG preferences.

Figure 13.9 illustrates the procedure. Panel B is the standard mean-variance diagram with the standard deviation and the expectation of the return along the axes. The riskfree rate is assumed to be 2%. The green curve is the mean-variance frontier for a given ESG score \bar{s} where the moments of all the stocks take any information in ESG scores into account (i.e., μ and Σ are used to generate the frontier). The dashed green line identifies the tangency portfolio conditional on the ESG score \bar{s} , and the slope of the line is the maximum Sharpe ratio conditional on the score \bar{s} , i.e., $SR(\bar{s})$. The black curve and black-dotted line are for the case without the extra condition of a given score \bar{s} , and are thus located to the left of the green equivalents. The ESG-SR frontier is the blue curve in Panel A that depicts the maximum Sharpe ratio $SR(\bar{s})$ as a function of the ESG score \bar{s} . The maximum of this curve is given by the maximum Sharpe ratio across all ESG scores. The part of the ESG-SR frontier that lies to the right of the maximum point is the efficient part of the ESG-SR frontier; any ESG-motivated investor will pick a point on this part of the frontier.

Since we assume that ESG scores can improve the estimates of expected returns, variances, and covariances, the tangency portfolio generated from asset moments ignoring ESG information has a lower Sharpe ratio and is thus below the blue curve. Individual assets are further below.

Equilibrium

Next, we think about the equilibrium implications of investors' portfolio decisions. Recall from Chapter 10 that the standard CAPM follows from the mean-variance model if all investors are unconstrained, have mean-variance preferences, and homogeneous beliefs. In the ESG setting above, it is then clear that if all investors are ESG-unaware, unconstrained, have mean-variance preferences, and homogeneous beliefs, then we obtain a CAPM-style result. The same is true if all investors are ESG-aware, unconstrained mean-variance optimizers with homogeneous beliefs, but here the moments entering the CAPM-equation for expected returns are all conditional on ESG scores. If all investors are ESG-motivated, on the other hand, the expected return on a given stock will deviate from the CAPM-result, and the deviation depends on the stock's ESG score relative to the market portfolio's ESG score as well as the constant π that indicates the strength of investors' ESG preferences. The precise results stated in the theorem below assumes that

the ESG score is informative about expected dividends in the following manner:

$$E[v_i|s_i] = \hat{\mu}_i + \lambda(s_i - s_m)$$

where $s_m = \sum_{i=1}^n m_i s_i$ is the ESG score of market portfolio.

Theorem 13.6

(a) If all investors are ESG-unaware, then

$$E[r_i] = \beta_i E[r_m] \quad \text{and} \quad E[r_i|s] = \beta_i E[r_m] + \lambda \frac{s_i - s_m}{p_i}$$

(b) If all investors are ESG-aware, then

$$E[r_i|s] = \bar{\beta}_i E[r_m|s] \quad \text{where } \bar{\beta}_i = \frac{\text{Cov}[r_i, r_m|s]}{\text{Var}[r_m|s]}$$

(c) If all investors are ESG-motivated, then

$$E[r_i|s] = \bar{\beta}_i E[r_m|s] - \pi(s_i - s_m)$$

where π is given by (13.39).

Note that with ESG-motivated investors the expected return is lower than predicted by CAPM for companies with above-average ESG scores. Hence, such companies can issue stocks at higher prices, supporting investments by good-ESG companies. Conversely, the expected return is higher than predicted by CAPM for companies with below-average ESG scores, which reduces the revenue that such companies can obtain when issuing stocks.

The above considerations are for the extreme cases where all investors are of the same type. The more interesting case is where the market has a mix of U, A, and M investors. While Pedersen, Fitzgibbons, and Pomorski (2021) present no formal results for this case, they formulate some natural conjectures:

- If most investors are ESG-unaware, the ESG-aware investors can generate alpha relative to CAPM (if $\lambda > 0$: overweight stocks with higher-than-average ESG score...)
- Intuitively, good ESG stocks can have lower or higher expected returns
- If good ESG predicts high dividends and the market contains many U-investors (who ignore this), then good ESG-stocks may have higher expected returns
- With many M-investors, good ESG-stocks have lower expected returns

Selected empirical results

Pedersen, Fitzgibbons, and Pomorski (2021) also perform some empirical analyses in which they apply the following ESG measures:

E: High E ~ Low carbon intensity, i.e., carbon emissions relative to sales

S: High S ~ non-sinful stock, i.e., not in tobacco, gambling, alcohol...

G: High G ~ Low accruals, i.e., only little income or expenses recorded in accounting reports before payment is made

ESG: MSCI ESG score ranging from 0 (worst ESG) to 10 (best)

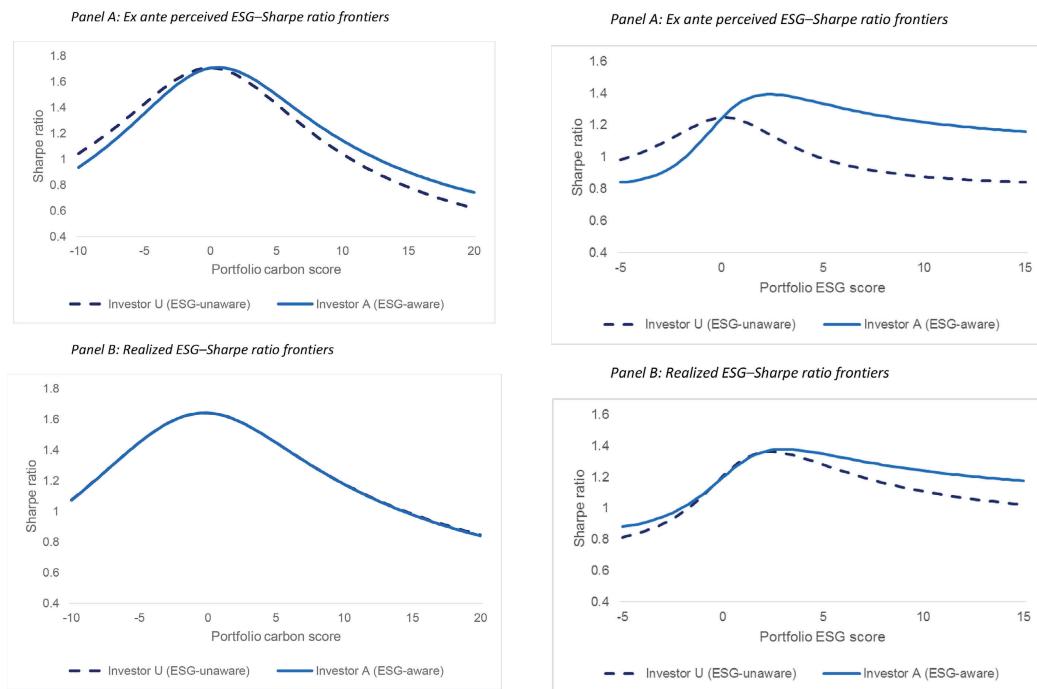


Figure 13.10: The ESG-SR frontier based on E or G.

This is Figures 4 and 5 from Pedersen, Fitzgibbons, and Pomorski (2021).

In each case, the scores are normalized so that zero represent the average score across stocks, and a value of x represents a score x standard deviations away from the average.

Figure 13.10 shows estimated ESG-SR frontiers when the ESG score is based either on the E measure (left panels) or the G score (right panels). The picture based on the overall ESG scores is similar to the one using only the E score and is therefore not included. The dashed curves are based on moments estimated without using ESG scores, whereas the solid curves include the ESG scores in the estimation. Each month, the ESG-SR frontier is determined, and the curves shown are the time series average of these frontiers. The realized or ex post frontier in the lower panel show the realized Sharpe ratios of the frontier portfolios. In the left panels, the curves are quite close, which indicates that the E measure (carbon intensity) has little influence on estimated return moments. The overall maximum Sharpe ratio is obtained with an E score near zero. If the ESG-motivated investor wants an E score two standard deviations higher, she would have to live with a Sharpe ratio about 3% lower than the overall maximum, a relatively modest reduction. In the right-hand panels, the solid and dashed curves are further apart, which suggests that G scores are more informative about asset return moments than E scores. The blue curve reaches its maximum for a normalized G score of around 2.25. A motivated investor can obtain an even higher G score with only a small reduction in the Sharpe ratio.

The S measure used in the study is binary in the sense that each stock is either a sin stock or a non-sin stock. In this case, a preference for non-sin stocks corresponds to screening, i.e., excluding the sin stocks. Screening is frequently used by ESG investors. Figure 13.11 shows the effects of screening based on the G score. Each curve represents an ESG-SR frontier where the G scores is used in the estimation of asset moments. The blue curve applies all stocks, whereas the green-dashed and the red-dotted curves ignore the 10% and the 20% of stocks with the lowest G scores. Of course, the curve moves down

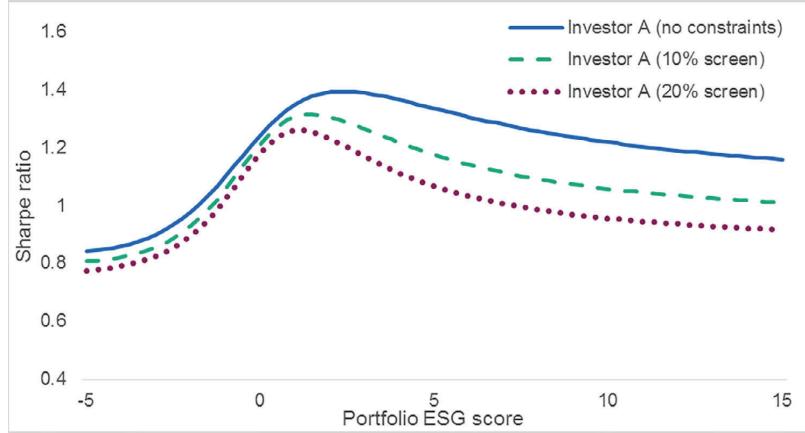


Figure 13.11: The impact of screening (based on G) on the ESG-SR frontier.
This is Figure 6 from Pedersen, Fitzgibbons, and Pomorski (2021).

when more restrictions are imposed. We observe that, for portfolios with large positive G scores, the reduction in Sharpe ratio caused by screening is substantial. Also note that the portfolio with the highest Sharpe ratio has a lower ESG score when the worst ESG stocks are removed! An intuitive explanation of this finding is that low-ESG assets can be valuable funding sources in the sense that an unconstrained investor might short them in order to build larger long positions in high-ESG securities.

Among the other findings reported in the paper are:

- High G-measure (low accruals) predicts strong future fundamentals and, apparently, receives little investor attention \leadsto low valuations and significantly positive abnormal returns
- Mixed results on whether E-measure (carbon emissions) predicts fundamentals, but some evidence of abnormal returns for high E-stocks in sample period due to increased investor attention
- Some evidence that sin stocks have stronger future fundamentals and some evidence of abnormal returns for sin stocks

13.5 Exercises

Exercise 13.1. Suppose returns follow the Single-Index Model. You believe that all assets have zero alphas, except for three stocks whose relevant characteristics are stated in the following table.

	α_i	β_i	$\text{Std}[\varepsilon_i]$
Hypothetics	0.08	2	0.5
Illuminati	0.06	1	0.4
Jinx Constructions	0.04	1.2	0.6

You intend to apply the Treynor-Black model to find the optimal portfolio. The riskfree rate is 0.02, and the stock market index has an expected rate of return of 0.08 and a standard deviation of 0.16.

- What is the Sharpe ratio of the market index?
- For each of three mispriced stocks, determine the expected rate of return, the return variance, the information ratio, and the Sharpe ratio.

- (c) What is the optimal active portfolio of the three stocks? What is the optimal active portfolio's alpha, beta, expected rate of return, return standard deviation, and Sharpe ratio?
- (d) How do you optimally combine the market portfolio and the active portfolio? Verify that Eq. (13.4) holds. Discuss your results!
- (e) How does your results change if the alpha of Hypothetics is 0.24 instead of 0.08? Discuss!

Exercise 13.2. Give a proof of Equation (13.29).

Exercise 13.3. Suppose that the data on the market portfolio and the riskfree rate are exactly as in Example 13.4. However, the fund FatReturns has just closed down. In your search for another fund to invest in, you are currently focusing on the funds SuperDuper and RockSolid. Your estimates of their key parameters are the following:

Fund	Average return	Standard deviation	Market beta
SuperDuper	16%	25%	1.2
RockSolid	20%	42%	1.4

- (a) Calculate for each fund the Sharpe ratio, the M-squared, the Treynor ratio, the alpha, and the information ratio.
- (b) If you want to invest all of your savings in a single fund, which one is better?
- (c) If you want to select one of the two funds and combine that with investments in numerous other funds and assets, which fund is better?
- (d) If you want to invest only in a combination of the market portfolio and one of these funds, which fund is better?

Exercise 13.4. Suppose returns follow the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

You believe that all assets have zero alphas, except for four stocks whose relevant return characteristics over the next year are stated in the following table.

	E[r _i]	α _i	β _i	Var[r _i]	Std[ε _i]
Awesome Airlines (AA)	???	0.05	2.00	???	0.5
Bad Bars (BB)	-0.01	-0.06	???	0.1300	???
Classy Casinos (CC)	0.14	0.08	1.25	0.2225	0.4
Delicious Dairy (DD)	0.11	0.04	1.50	0.2500	0.4

You intend to apply the Treynor-Black model to find the optimal portfolio. The riskfree rate is 0.01, and the stock market index has an expected rate of return of 0.05 and a standard deviation of 0.2.

- (a) Determine the missing information in the table above, i.e., find the expectation and variance of the return on AA, as well as the beta and residual standard deviation of BB.
- (b) What is the optimal active portfolio of the four stocks? What is the optimal active portfolio's alpha, beta, expected rate of return, return standard deviation, and Sharpe ratio?
- (c) How do you optimally combine the market portfolio and the active portfolio? What is the Sharpe ratio of this optimally combined portfolio?
- (d) Suppose your objective is to maximize E[r] - $\frac{1}{2}\gamma \text{Var}[r]$ with a relative risk aversion of $\gamma = 2$. If you only invest in the market portfolio and the riskfree asset, what is your optimal combination of the two? If you include the four mispriced stocks studied above, what is then your optimal portfolio, i.e., what are the optimal weights in the market portfolio, in the riskfree asset, and in each of the four mispriced stocks?

Exercise 13.5. Suppose returns follow the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

You believe that all assets have zero alphas, except for four stocks whose relevant return characteristics over the next year are stated in the following table.

	α_i	β_i	$\text{Std}[\varepsilon_i]$
Great Garbage	0.05	2.00	0.5
Hunky Homes	0.04	1.00	0.3
Improper Insurance	-0.05	1.25	0.4
Jinx Jets	-0.02	1.50	0.4

You intend to apply the Treynor-Black model to find the optimal portfolio. The riskfree rate is 0.01, and the stock market index has an expected rate of return of 0.05 and a standard deviation of 0.2.

- (a) For each of the four mispriced stocks, determine the expectation and the standard deviation of the rate of return.
- (b) What is the optimal active portfolio of the four stocks? What is the optimal active portfolio's alpha, beta, expected rate of return, return standard deviation, and Sharpe ratio?
- (c) How do you optimally combine the market portfolio and the active portfolio? What is the expected return, the standard deviation, and the Sharpe ratio of this optimally combined portfolio?
- (d) Construct a graph with the standard deviation of the return along the horizontal axis and the expected return along the vertical axis. In this graph, show the line/curve corresponding to
 - i) combinations of the riskfree asset and the market portfolio,
 - ii) combinations of the market portfolio and the optimal active portfolio, and
 - iii) combinations of the riskfree asset and the optimally combined market-active portfolio identified in the previous question.

Also indicate the location of the four mispriced stocks in the diagram.

- (e) Suppose your objective is to maximize $E[r] - \frac{1}{2}\gamma \text{Var}[r]$ with a relative risk aversion of $\gamma = 1$. If you only invest in the market portfolio and the riskfree asset, what is your optimal combination of the two and what is the standard deviation and expected return of that portfolio? If you include the four mispriced stocks studied above, what is then your optimal portfolio and its standard deviation and expected return? Answer the same questions for $\gamma = 2$ and compare the results.

Exercise 13.6. Suppose that the riskfree rate is 1% per year. You have the following information on the annual returns on the market portfolio and three investment funds that you consider investing some money in:

Fund	Expected return	Standard deviation	Market beta
Market portfolio	7%	16%	1.0
FineFunds	11%	28%	0.9
GroovyGains	11%	25%	1.2
IncredibleInvest	16%	40%	1.5

- (a) Calculate for each fund the Sharpe ratio, the M-squared, the Treynor ratio, the alpha, and the information ratio. Explain your calculations for FineFunds in detail.
- (b) If you want to invest all of your savings in a single fund, which one is better? Why?
- (c) If you want to invest only in a combination of the market portfolio and one of the three funds, which fund is better? Why?
- (d) If you want to select one of the three funds and combine that with investments in numerous other funds and assets, which fund is better? Why?

Exercise 13.7. You intend to apply the Black-Litterman model for determining how much to invest in the stocks of the three companies Daydreams, Hypothetics, and Illuminati as well as the riskfree asset over the next year. The following table contains relevant information about the three stocks. Column 2 lists the standard deviation of the annual return of each stock, columns 3-5 the pairwise correlations, and column 6 the market capitalization of the company in millions of USD. The overall stock market is assumed to have an expected excess return of 0.06 and a standard deviation of 0.20. The model parameter τ is fixed at 0.02, and your relative risk aversion is assumed to be 2.

Stock	Standard deviation	Correlations			Market cap (mUSD)
		Daydreams	Hypothetics	Illuminati	
Daydreams	0.30	1.0	0.4	0.6	500
Hypothetics	0.20	0.4	1.0	0.2	300
Illuminati	0.25	0.6	0.2	1.0	200

- (a) Calculate the CAPM-implied expected excess returns of the three stocks.
- (b) If you don't have any subjective views on the three stocks, what is your optimal portfolio?

In fact, you have two subjective views on the returns of the three stocks. The first view is that the expected excess return on Daydreams is really 0.02. You are 90% certain that the true expected excess return is between 0 and 0.04. The second view is that you expect the return on Hypothetics to be 0.05 larger than the return on Illuminati. You are 90% certain that the expected return difference is between 0.01 and 0.09.

- (c) How does your subjective view differ from the objective expectations captured by the CAPM-implied expected excess returns? (With model notation, this means compare \mathbf{Q} to $\mathbb{P}\mathbf{m}$.)
- (d) What is your optimal portfolio if you incorporate only your first view? How does it differ from the optimal portfolio in the absence of the view (see question b)? Discuss!
- (e) What is your optimal portfolio if you incorporate only your second view? How does it differ from the optimal portfolio in the absence of the view (see question b)? Discuss!
- (f) What is your optimal portfolio if you incorporate both views? How does it differ from the optimal portfolio in the absence of the view (see question b)? Discuss!

Exercise 13.8. Your cousin Carl has inherited a large sum of money. Since he knows that you are a clever finance student, he has asked for your advice on how to invest the money. He has identified three stocks that he would like to invest in over the next year. We refer to the stocks as X, Y, and Z in the following.

You assume that the return on each stock is given by the Single-Index Model

$$r_i - r_f = \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i,$$

where the usual notation is applied. The riskfree rate is $r_f = 0.02$, and you estimate that the market index has an expected return of $E[r_m] = 0.06$ and a standard deviation of $\sigma_m = \text{Std}[r_m] = 0.15$. Based on this model, you have estimated the expected returns and the variance-covariance matrix as shown in this table:

Stock	E[r]	Covariances		
		X	Y	Z
X	0.074	0.0976	0.0360	0.0288
Y	0.100	0.0360	0.1125	0.0180
Z	0.112	0.0288	0.0180	0.1744

For example, the covariance between stock Y and stock Z is $\text{Cov}[r_Y, r_Z] = 0.0180$.

- (a) For each stock, calculate the return standard deviation and the Sharpe ratio.

Carl has informed you that his target expected rate of return is 11%.

- (b) Determine the mean-variance efficient portfolio of the three stocks that achieves exactly this expected return. What is the standard deviation of the portfolio return?

You will try to convince Carl that he should include the riskfree asset in the portfolio.

- (c) Determine the portfolio weights of the three stocks in the tangency portfolio as well as the portfolio's expected return, standard deviation, and Sharpe ratio. Does the tangency portfolio weights make sense given the inputs?
 (d) Which fraction of his wealth should Carl invest in the riskfree asset and each of the three stocks to obtain an 11% expected return with the lowest possible variance? Calculate the standard deviation of this portfolio and compare with the answer to question (b) above.

Carl seems to be convinced that it makes sense to combine the three stocks with the riskfree asset. You decide to try to convince him that he should also invest in an ETF that perfectly replicates the stock market index.

- (e) Based on the pairwise covariances of the three stocks and the properties of the Single-Index Model, state three equations that you can solve (for example, by using Solver in Excel) for the beta-values of the stocks. What is the solution, i.e. β_X , β_Y , and β_Z ?
 (f) For each of the three stocks X, Y, and Z, determine the alpha-value and the standard deviation of the residual (i.e. $\text{Std}[\varepsilon_i]$).

Given the answers to the previous questions, you intend to apply the Treynor-Black model to find the optimal combination of the market fund, the three stocks, and the riskfree asset.

- (g) Determine the optimal active portfolio of the three stocks.
 (h) What is the optimal combination of the market fund and the active portfolio? Calculate the Sharpe ratio of this combined portfolio and compare with the Sharpe ratio of the tangency portfolio from question (c).
 (i) Which portfolio of the market fund, the three stocks, and the riskfree asset generates an expected return of 11% with the lowest possible standard deviation? Calculate the standard deviation of this portfolio and compare with the answers to questions (b) and (d) above.
 (j) Draw a diagram with standard deviation of returns along the horizontal axis and expected returns along the vertical axis. Show the location of the three stocks, the market index, the riskfree asset, the tangency portfolio of the three stocks, the optimal active portfolio, as well as each of the portfolios determined in questions (b), (d), and (i). Also draw the efficient frontier determined from (1) only the three stocks and the riskfree asset and from (2) the three stocks, the riskfree asset, and the market portfolio.

Exercise 13.9. Suppose returns follow the Single-Index Model, i.e.

$$r_i = r_f + \alpha_i + \beta_i(r_m - r_f) + \varepsilon_i.$$

You believe that all assets have zero alphas, except for three stocks whose relevant characteristics over the next year are stated in the following table.

	α_i	β_i	$\text{Std}[\varepsilon_i]$
Duff Beer	-0.08	0.8	0.4
Monsters, Inc.	0.04	1.4	0.2
Wayne Enterprises	0.08	1.0	0.4

You intend to apply the Treynor-Black model to find the optimal portfolio. The riskfree rate is 0.01, and the stock market index has an expected rate of return of 0.07 and a standard deviation of 0.16.

- (a) For each of the three mispriced stocks, determine the expected rate of return, the return variance, the information ratio, and the Sharpe ratio.
- (b) What is the optimal active portfolio of the three stocks? What is the optimal active portfolio's alpha, beta, expected rate of return, return standard deviation, and Sharpe ratio?
- (c) What is the optimal combination of the market portfolio and the active portfolio? Compute the Sharpe ratio of this optimally combined portfolio and compare with the Sharpe ratio of the market portfolio. Briefly discuss your results.

Exercise 13.10. The supplementary material for these lecture notes includes an Excel data file `ESG_portfolios_data.xlsx` to be used in this exercise. The file is taken from the homepage <https://www.lhpedersen.com/data> of Lasse Heje Pedersen, Professor of Finance at the Copenhagen Business School. The data was used in the paper [Pedersen, Fitzgibbons, and Pomorski \(2021\)](#).

The data file shows monthly returns on various portfolios of stocks sorted by either E, S, G, or an overall ESG score.

- (a) For each of the E-sorted portfolios E1, ..., E5 and the difference E5-E1 calculate the sample mean, standard deviation, skewness, and kurtosis, and find the minimum and maximum observation. Use the full sample period from June 2009 to March 2019. Do your findings indicate a systematic relation between firms' environmental performance and their stock returns? How is this question investigated in the Pedersen-Fitzgibbons-Pomorski paper and what is the authors' conclusion?
- (b) Now consider the S-sorted portfolios S1 and S2 and the difference S2-S1. Perform the same analysis as done in (a) above. Do this both for (i) the full sample from February 1963 to March 2019 and (ii) the most recent 10 years of observations. What can you conclude? Compare again with the analysis and conclusions in the Pedersen-Fitzgibbons-Pomorski paper.
- (c) Next consider the G-sorted portfolios G1, ..., G5 and the difference G5-G1. Perform the same analysis as done above. Do this both for (i) the full sample from February 1963 to March 2019 and (ii) the most recent 10 years of observations. What can you conclude? Compare again with the analysis and conclusions in the Pedersen-Fitzgibbons-Pomorski paper.
- (d) Finally consider the ESG-sorted portfolios ESG1, ..., ESG5 and the difference ESG5-ESG1. Perform the same analysis as done above. Do this for the full sample from February 2007 to March 2019. What can you conclude? Compare again with the analysis and conclusions in the Pedersen-Fitzgibbons-Pomorski paper.

CHAPTER 14

Forwards, futures, and swaps

Derivative securities were introduced in Section 1.4 which also presented some statistics on the size of the markets for derivatives. The main types of derivatives are forwards, futures, swaps, and options. This chapter digs deeper into forwards, futures, and swaps, whereas the subsequent chapter covers options. We explain how the different derivatives can be used for risk management and speculation. Furthermore, we develop some standard formulas and methods for the pricing of such assets. We deal with forwards in Section 14.1, futures in Section 14.2, and swaps in Section 14.3.

14.1 Forwards

14.1.1 Definition and characteristics

A forward contract or simply a **forward** is a binding agreement between two parties to transact a given asset at a pre-specified future point in time and at a pre-specified price. The party committing to buying the asset is said to have a *long position* or to buy the asset forward. The party committing to selling the asset is said to have a *short position* or to sell the asset forward. The pre-specified transaction price is called the *delivery price* and the time of delivery is called the *maturity date* of the forward. The asset to be transacted is referred to as the *underlying asset*.

Let $S(t)$ denote the price of the underlying asset at any time t , let K denote the pre-specified delivery price, and T the maturity date. An investor with a long position in the forward can buy the asset at time T at a price of K instead of buying it at the market price $S(T)$ prevailing at that time. Therefore, the forward contract has a value to him equal to the difference $S(T) - K$. We can think of this party receiving a payoff from the forward at time T equal to

$$S(T) - K.$$

Conversely, the party with a short position receives

$$K - S(T)$$

from the forward. Of course, the payoff is generally positive to one party and negative to the other, but the sign and size of the payoff are not known until time T . The payoffs are illustrated in Figure 14.1. The payoffs can be settled either by

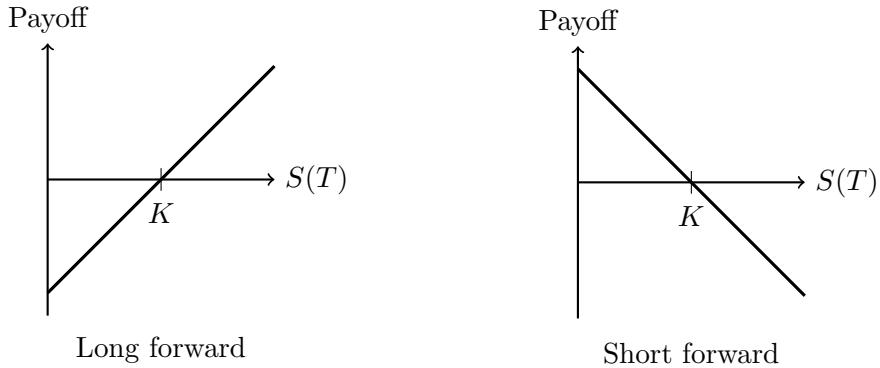


Figure 14.1: Forward payoffs.

The figure shows the payoff of a long forward contract in the left diagram and the payoff of a short forward contract in the right diagram, in both cases shown as a function of the price of the underlying asset at the maturity date. K denotes the pre-specified delivery price.

- (a) physical delivery: the investor with the short position delivers one unit of the underlying asset to the other party and receives a cash payment of K in return; or
- (b) cash settlement: only the amount $S(T) - K$ is exchanged, not the underlying asset.

The delivery price is set so that the present value of the contract is zero to both parties when the forward contract is written. The value of K that ensures this is called the **forward price** and it generally depends on the underlying asset and the maturity date. We denote the forward price at time t for maturity date T (where $t \leq T$) by $F(t, T)$ or simply F when no confusion about the dates should arise. Since the present value is then zero, no money is changing hands when the contract is made. Consequently, the only payment on a forward contract is on the maturity date where the net payments is plus or minus $S(T) - F(t, T)$, depending on the side of the trade. In particular, note that the forward price is not the price of entering into a forward contract, but simply a contract parameter important for the payoff at maturity.

At the time or maturity T the fair forward price must be equal to the spot price,

$$F(T, T) = S(T). \quad (14.1)$$

Why? If $F(T, T) < S(T)$ we could take a long position in the forward contract and immediately cash in a profit of $S(T) - F(T, T)$. Similarly, if $F(T, T) > S(T)$. Any difference between $F(T, T)$ and $S(T)$ would thus lead to an easy arbitrage profit. Below we derive formulas for the fair forward price $F(t, T)$ before maturity, but first we look at the applications of forward contracts.

14.1.2 Risk management and speculation using forwards

Forwards are useful for hedging risk. Consider an investor who knows today that he needs a certain asset at a future date T . If he waits and buys it at this date, he has to pay the prevailing spot price $S(T)$. Since this price is uncertain, he faces the risk of having to pay a high price. However, he can completely eliminate the risk by taking a long position in a forward contract on this asset with delivery at date T . He can effectively lock in the future purchase price at today's forward price $F(t, T)$. Conversely, an investor who knows today that he wants to sell an asset at a future date T can lock in the selling price

by taking a short position in a forward on the asset with maturity T . Note that there are no costs involved in taking the relevant position in the forward. Locking in the price to be paid or received at the future date implies that the investor will not benefit from a favorable development in the price of the underlying asset, but on the other hand the investor is protected against an unfavorable development.

Obviously forwards are most useful for hedging when the timing and the size of the future transaction is known with certainty. If that is not the case, a position in the forward cannot fully eliminate the risky position. If the investor enters a forward contract and later learns that the future transaction is cancelled, he can take an off-setting position in a new forward contract on the same asset and having the same delivery date. The difference between the original forward price and the new forward price determines the gain or loss to the investor. Obtaining the off-setting forward position is sometimes expensive or impossible, however, since forwards are not exchange-traded and the investor might have to search for a counterparty himself.

Example 14.1

Two months from now a U.S. based company expects to receive a payment of 20 million euro from a German customer. The company fears that the dollar-euro exchange rate develops unfavorably to them (leading to fewer dollars per euro). This risk is eliminated by entering a forward on the dollar-euro rate in which the company sells 20 million euro in exchange for dollars. The forward matures in two months.

Example 14.2

A U.S. producer of corn flakes needs five tons of corn in three months, but fears that the corn price will increase before then. The producer can lock in the purchase price today by taking a long position in a three-month forward on corn.

Forwards are also used by investors for speculating in a certain development in the price of the underlying asset. An investor expecting an increase in the price of an asset can take a long position in a forward on the asset. Taking this position is costless. If his expectations are realized, he will receive a positive payoff at maturity. More precisely, the payoff he receives is $S(T) - F$, so he has to compare the expected future spot price to the current forward price. Of course, he takes the risk that his expectations were wrong in which case he ends up with a negative payoff when the forward expires.

14.1.3 Forwards on non-dividend paying assets

To simplify the analysis, let us first assume the underlying asset pays no dividends before the maturity date of the forward. For example, this could be a non-dividend paying stock or a bond with no payments before the forward expires. As before, let $S(t)$ denote the price of the underlying at time t , let T denote the maturity date of the forward, and let $F(t, T)$ denote the forward price at time t for maturity T . The forward price is set at time t so that the present value of the payoff $S(T) - F(t, T)$ at maturity is zero. Afterwards, the forward contract generally has a non-zero value. We aim at determining the forward price as well as the value of a forward after inception.

Let $V(t, T, K)$ denote the value at time t of a long position in a forward with maturity

date T and delivery price K . The value of the corresponding short position is of course $-V(t, T, K)$. Let $Z(t, T)$ denote the price at time t of a default-free zero-coupon bond maturing at time T and having a face value of 1. This zero-coupon bond is the riskfree asset when the investment horizon ends at time T . With an investment of $Z(t, T)K$ at time t , you can buy K such zero-coupon bonds and thus get a total payoff of K at time T . The corresponding (annualized) yield $y(t, T)$ is given via the relation

$$Z(t, T) = (1 + y(t, T))^{-(T-t)}. \quad (14.2)$$

Hence $y(t, T)$ is the riskfree rate between t and T . Alternatively, we can think of a long position in the zero-coupon bond as a fixed-term (bank) deposit up to time T , where $y(t, T)$ is then the pre-set interest rate on the deposit. Conversely, a short position in the zero-coupon bond is like a fixed-term (bank) loan up to time T with a pre-set interest rate $y(t, T)$.

Theorem 14.1

If no arbitrage opportunities exist, then

$$V(t, T, K) = S(t) - Z(t, T)K. \quad (14.3)$$

The unique arbitrage-free forward price at time t is

$$F(t, T) = \frac{S(t)}{Z(t, T)}. \quad (14.4)$$

Proof

Consider investing in the following two portfolios at time t :

- A: a long position in the forward plus the amount $Z(t, T)K$ invested in the riskfree asset (i.e., the zero-coupon bond maturing at T),
- B: one unit in the underlying asset.

At time t the value of portfolio A is $V(t, T, K) + Z(t, T)K$, and the value of B is simply $S(t)$. At the maturity date of the forward, the value of A is $S(T)$, namely the sum of the payoff $S(T) - K$ from the forward plus K from the riskfree investment. The value of portfolio B is clearly also $S(T)$. Since the two portfolios have the same future payments, they must have the same value today (time t), i.e.,

$$V(t, T, K) + Z(t, T)K = S(t),$$

from which (14.3) follows. If this relation does not hold, there is a clear arbitrage opportunity in buying the cheaper of the two portfolios and selling the more expensive one and holding on to that position until the forward expires.

The time t forward price $F(t, T)$ is the value of K so that $V(t, T, K) = 0$, that is,

$$S(t) - Z(t, T)K = 0 \quad \Leftrightarrow \quad K = \frac{S(t)}{Z(t, T)},$$

which confirms (14.4).

Note that in terms of the yield on the relevant zero-coupon bond, we can express the forward price as

$$F(t, T) = S(t) (1 + y(t, T))^{T-t}, \quad (14.5)$$

so we can think of the forward price as the current spot price discounted forward in time to the delivery date using the appropriate riskfree rate for the period.

The proof of the above theorem claims that an arbitrage exist if you can enter into a forward contract (at zero cost) with a delivery price K different from the value $F(t, T)$ given by (14.4). If, for example, $K > S(t)/Z(t, T)$, the strategy is to borrow $Z(t, T)K$ at the riskfree rate, buy the underlying asset, and sell it forward with delivery price K . This leads to an immediate profit of $Z(t, T)K - S(t) > 0$, whereas the net payment at maturity is zero no matter what happens. And there are no payments between t and T . This is a riskfree profit, an arbitrage.

Conversely, if $K < S(t)/Z(t, T)$, then invest $Z(t, T)K$ in the riskfree asset, sell the underlying asset, and buy it forward with a delivery price of K . This provides an immediate payoff of $S(t) - Z(t, T)K$ and zero net payments in the future. Note that this requires selling the underlying asset. If you do not own the asset, maybe you can sell it short. If not, you cannot exploit the arbitrage opportunity. But as long as some investors can, they will, and prices change until the arbitrage opportunity vanishes and the relation (14.4) is established.

Example 14.3

The stocks of XYZ Inc. currently sell at \$710 per share. You are offered to buy the stock on a three-month forward with a delivery price of \$720 and at no cost today. The stock pays no dividends in that period. The three-month riskfree rate is 4% (annualized). Should you accept the offer?

According to (14.5), the arbitrage-free forward price is

$$710 \times (1.04)^{3/12} \approx 717.00.$$

Hence, you should decline the offer. If you could sell the stock forward with a delivery price of \$720, you should certainly do so because this leads to an arbitrage. Buy the stock now and sell it forward, and borrow $(1.04)^{-3/12} \times 720 \approx 712.97$ at the riskfree 3-month rate. This provides you with an immediate profit of \$2.97. (Scale up positions if you want a higher profit.) At maturity, the net payment is zero.

Note that it follows from Eqs. (14.3) and (14.4) that the time t value of a forward contract with delivery price K can be written as

$$V(t, T, K) = (F(t, T) - K)Z(t, T). \quad (14.6)$$

In particular, a forward contract entered into at time 0 with a forward price of $F(0, T)$ will have a time t value of

$$V(t, T, F(0, T)) = (F(t, T) - F(0, T))Z(t, T). \quad (14.7)$$

After the forward contract is made, its value generally moves away from zero.

14.1.4 Forwards on assets with known dividends

The formula for the value of a forward contract in Theorem 14.1 can be generalized to the case with known dividends from the underlying asset before maturity of the forward. The generalized formula reads

$$V(t, T, K) = \mathbf{PV}_t(S(T)) - Z(t, T)K, \quad (14.8)$$

where $\mathbf{PV}_t(S(T))$ is the present value at time t of getting an uncertain payment of $S(T)$ at time T and nothing at other dates. In other words, $\mathbf{PV}_t(S(T))$ is the amount that has to be invested at time t to end up with $S(T)$ at time T . In some relevant cases, this present value can be computed quite easily. The arbitrage-free forward price follows again by deriving the delivery price for which the current value of the forward equals zero:

$$F(t, T) = \frac{\mathbf{PV}_t(S(T))}{Z(t, T)}. \quad (14.9)$$

In the case where the underlying asset makes no payment before the forward maturity date, we have $\mathbf{PV}_t(S(T)) = S(t)$, so that we are back at the formulas (14.3) and (14.4). With a position in a forward maturing at time T , you care about the stock price at time T , which is the value of all the dividends paid after time T . In contrast, the current stock price reflects the value of both the dividends coming after time T and the dividend paid out between today and time T . So for computing the forward price, you have to subtract the value of these intermediate dividends from the current stock price. If we rewrite Eq. (14.9) as $Z(t, T)F(t, T) = \mathbf{PV}_t(S(T))$, we have the intuitive relation that the discounted arbitrage-free delivery price equals the present value of the price of the underlying asset at the delivery date.

Note that the expressions (14.6) and (14.7) are still valid.

For forwards on coupon bonds or stocks paying a dividend which is known already (or can be estimated with high precision), $\mathbf{PV}_t(S(T))$ is easily computed. Assume for simplicity that the underlying asset has a single payment between today t and the forward maturity date T . Let D denote the payment and let $t^* \in (t, T)$ denote the payment date. Then

$$\mathbf{PV}_t(S(T)) = S(t) - Z(t, t^*)D,$$

since this is the amount that has to be invested at time t to end up with $S(T)$ at time T and nothing at other dates. To see this, set up a portfolio of one unit of the underlying asset and a loan of $Z(t, t^*)D$. As the owner of the underlying asset you receive D at time t^* and you can use that payment to pay off the loan, which implies a zero net payment at time t^* .

Example 14.4

Let us find the arbitrage-free six-month forward price on a coupon bond. The bond has a face value of \$100 and makes a coupon payment of \$5 in four months. The spot price of the bond (including accrued interests) is \$96.30. The four-month interest rate is 3.6% and the six-month interest rate is 3.8%, both annualized.

The appropriate discount factors are

$$Z(t, t + 4/12) = (1 + 0.036)^{-4/12} \approx 0.9883, \quad Z(t, t + 0.5) = (1 + 0.038)^{-0.5} \approx 0.9815.$$

The present value of the coupon payment is thus $5 \times 0.9883 \approx 4.9414$. According to (14.9), the arbitrage-free forward price is then

$$F(t, t + 0.5) = \frac{96.30 - 4.9414}{0.9815} \approx 93.0782.$$

Another tractable case is when the underlying asset can be assumed to pay dividends continuously through time with the dividend payments being proportional to the value of the asset, i.e., with a constant dividend yield δ . This is a reasonable assumption when the underlying asset is a foreign currency where the foreign riskfree rate takes the role of δ because you can deposit the foreign currency and receive interests. It may also be a reasonable approximation for a broad stock index. As the dividend payment dates of the different companies are spread out over the year, the total dividends to a basket of many stocks is more or less a continuous stream. And for relatively short periods of time, it may be reasonable to assume the dividend yield is a constant δ . With a constant dividend yield δ it suffices to purchase $e^{-\delta(T-t)}$ units of the underlying asset at time t in order to end up with one unit of the asset at time T .¹ Consequently,

$$\mathbf{PV}_t(S(T)) = S(t)e^{-\delta(T-t)},$$

and the fair forward price is therefore

$$F(t, T) = \frac{S(t)e^{-\delta(T-t)}}{Z(t, T)} = S(t)e^{(y^c(t, T) - \delta)(T-t)}, \quad (14.10)$$

where $y^c(t, T)$ is the continuously compounded yield at time t for payments at time T , cf. the discussion around Eq. (5.17).

14.1.5 Currency forwards

A currency forward is a forward contract where the underlying asset is a specified number of units of a foreign currency. Currency forwards are also called forwards on foreign exchange or simply FX forwards. Note that having a long position in a forward contract on euros with a delivery price stipulated in dollars, you will end up receiving some amount in euros and paying some amount in dollars. This is the same as having a short position in a forward contract on dollars with a delivery price stipulated in euros.

For currency forwards we can also determine the present value $\mathbf{PV}_t(S(T))$ entering our valuation formulas. The owner of a unit of a foreign currency can invest in foreign zero-coupon bonds or deposit it in a bank and thus earn the riskfree rate on that currency, i.e., the riskfree rate in the country or region where the currency is used. Let $Z_f(t, T)$ denote the corresponding discount factor in the foreign country so that a riskfree investment of $Z_f(t, T)$ units of the foreign currency at time t grows to one unit at time T . Therefore

$$\mathbf{PV}_t(S(T)) = Z_f(t, T)S(t).$$

Here $S(t)$ is the spot exchange rate measured in units of domestic currency per unit of the

¹This is completely analogous to continuously compounded interest rates. If the annualized, continuously compounded interest rate is r , then 1 dollar deposited at time t grows to $e^{r(T-t)}$ dollars at time T . Conversely, you have to deposit $e^{-r(T-t)}$ dollars at time t to end up with 1 dollar at time T .

foreign currency. It follows that the value of a forward with delivery price K is

$$V(t, T, K) = Z_f(t, T)S(t) - Z(t, T)K, \quad (14.11)$$

and that the arbitrage-free forward price at time t is

$$F(t, T) = \frac{Z_f(t, T)}{Z(t, T)}S(t). \quad (14.12)$$

The forward price of a unit of foreign currency is referred to as the **forward exchange rate**.

If we let $y(t, T)$ and $y_f(t, T)$ denote the riskfree rate per year in the home country and the foreign country, respectively, we have

$$Z(t, T) = (1 + y(t, T))^{-(T-t)}, \quad Z_f(t, T) = (1 + y_f(t, T))^{-(T-t)}. \quad (14.13)$$

Now the value of the currency forward can be restated as

$$V(t, T, K) = S(t)(1 + y_f(t, T))^{-(T-t)} - K(1 + y(t, T))^{-(T-t)}$$

and the arbitrage-free forward exchange rate as

$$F(t, T) = S(t) \left(\frac{1 + y(t, T)}{1 + y_f(t, T)} \right)^{T-t}. \quad (14.14)$$

This is the so-called **covered interest rate parity**. Note again that this *has* to hold, otherwise an arbitrage exists. Nevertheless, some studies find that deviations from the parity have been frequent and sometimes sizeable during and after the global financial crisis in 2007–2009, see Cerutti, Obstfeld, and Zhou (2021) and Du, Tepper, and Verdelhan (2018).

The **forward premium** is the difference between the forward price and the spot price, i.e.,

$$F(t, T) - S(t) = S(t) \left[\left(\frac{1 + y(t, T)}{1 + y_f(t, T)} \right)^{T-t} - 1 \right]. \quad (14.15)$$

We can get a nice approximative formula for the right-hand side by using a first-order Taylor approximation (i.e. a linear approximation) of the function $f(x) = x^{T-t} - 1$ around $x = 1$. Note that $f(1) = 0$. Furthermore, $f'(x) = (T-t)x^{T-t-1}$ so $f'(1) = T-t$. Hence the Taylor approximation is

$$f(x) \approx f(1) + f'(1)(x - 1) = (T-t)(x - 1).$$

It follows that

$$\begin{aligned} \left(\frac{1 + y(t, T)}{1 + y_f(t, T)} \right)^{T-t} - 1 &\approx (T-t) \left(\frac{1 + y(t, T)}{1 + y_f(t, T)} - 1 \right) \\ &= (T-t) \frac{1 + y(t, T) - [1 + y_f(t, T)]}{1 + y_f(t, T)} \\ &= (T-t) \frac{y(t, T) - y_f(t, T)}{1 + y_f(t, T)}. \end{aligned}$$

The forward premium is thus approximated as

$$F(t, T) - S(t) \approx S(t)(T - t) \frac{y(t, T) - y_f(t, T)}{1 + y_f(t, T)}. \quad (14.16)$$

If $y_f(t, T)$ is close to zero, you might even use the approximation

$$F(t, T) - S(t) \approx S(t)(T - t) (y(t, T) - y_f(t, T)). \quad (14.17)$$

The forward premium reflects the difference in the interest rates of the two countries. If the home country has a higher interest rate than the foreign country, the forward price exceeds the spot price so that the forward premium is positive.

Example 14.5

Suppose the spot U.S. dollar-euro exchange rate is 1.35 dollars per euro and that the one-year interest rate is 0.55% on U.S. dollar deposits and 0.47% on euro deposits. Then the one-year forward exchange rate should be

$$F_{\text{USD/EUR}} = S_{\text{USD/EUR}} \frac{1 + y_{\text{USD}}}{1 + y_{\text{EUR}}} = 1.35 \times \frac{1.0055}{1.0047} \approx 1.351075 \text{ USD per euro}$$

corresponding to a forward premium of 0.001075 U.S. dollars per euro. The approximation (14.17) leads to a premium of

$$F - S \approx 1.35 \times 1 \times (0.0055 - 0.0047) = 0.001080 \text{ USD per euro},$$

which is very close to the exact value.

Sometimes continuously compounded yields are used in relation to currency forwards. If $y^c(t, T)$ and $y_f^c(t, T)$ denote the continuously compounded domestic and foreign yields or interest rates over the period from t to T , the fair forward price can be expressed as

$$F(t, T) = S(t) e^{(y^c(t, T) - y_f^c(t, T))(T - t)} \quad (14.18)$$

and the forward premium as

$$\begin{aligned} F(t, T) - S(t) &= S(t) \left(e^{(y^c(t, T) - y_f^c(t, T))(T - t)} - 1 \right) \\ &\approx S(t) (y^c(t, T) - y_f^c(t, T)) (T - t), \end{aligned} \quad (14.19)$$

where we are using the approximation $e^x - 1 \approx x$ for $x \approx 0$.

14.1.6 Commodity forwards

Commodity forwards are useful in the risk management of producers and consumers of the commodity, as well as for speculators trying to profit from their price predictions. A key difference between commodities (such as oil, sugar, or gold) and financial assets is that it is costly to own commodities due to the need for storage facilities and any associated insurance contracts. Let us simply refer to such costs as **storage costs**. Note that costs act like a negative dividend on the commodity.

If we let $L(t, T)$ denote the present value of the storage costs per unit of the underlying commodity over the period from t to T , we need to invest

$$\mathbf{PV}_t(S(T)) = S(t) + L(t, T)$$

at time t to end up with $S(T)$ at time T . According to (14.9), the arbitrage-free forward price is then

$$F(t, T) = \frac{S(t) + L(t, T)}{Z(t, T)}.$$

For example, if you have to pay storage costs of ℓ_1 at time t_1 , ℓ_2 at time t_2 , etc., up to ℓ_n at time t_n , where $t < t_1 < t_2 < \dots < t_n < T$, then $L(t, T) = \sum_{i=1}^n \ell_i Z(t, t_i)$. For simplicity, it is often assumed that the storage costs are proportional to the spot price of the commodity. If ℓ denotes the percentage cost per year, then $L(t, T) = (T - t)\ell S(t)$, and the forward price becomes

$$F(t, T) = S(t) \frac{1 + (T - t)\ell}{Z(t, T)}.$$

If we use a simple annualized riskfree rate y^{simp} so that $Z(t, T) = 1/[1 + (T - t)y^{\text{simp}}]$, then we can rewrite the forward price as

$$\begin{aligned} F(t, T) &= S(t) (1 + (T - t)\ell) \left(1 + (T - t)y^{\text{simp}} \right) \\ &\approx S(t) \left(1 + (T - t) (\ell + y^{\text{simp}}) \right), \end{aligned} \tag{14.20}$$

where the approximation disregards the term $(T - t)^2 \ell y^{\text{simp}}$ which is small when ℓ and y^{simp} are small. We can see that the relative storage cost ℓ is added to the interest rate. The sum is often called the **cost of carry**. Intuitively, if you borrow money at time t to buy the commodity and “carry” it until time T , you have to pay both interest and storage costs.

Again, continuous compounding is sometimes used in this context. If $y^c(t, T)$ denotes the continuously compounded riskfree interest rate per year and ℓ^c denotes the continuously compounded relative storage cost, then

$$Z(t, T) = e^{-(T-t)y^c(t, T)}, \quad L(t, T) = S(t) \left(e^{(T-t)\ell^c} - 1 \right), \tag{14.21}$$

so that the forward price is

$$F(t, T) = S(t) e^{(T-t)[\ell^c + y^c(t, T)]}. \tag{14.22}$$

The above expressions for the commodity forward price are again based on a no-arbitrage argument as the one presented in the proof of Theorem 14.1. Note that if the inequality

$$F(t, T) < \frac{S(t) + L(t, T)}{Z(t, T)} \tag{14.23}$$

holds, the arbitrage would involve selling the commodity spot and buying it forward. However, companies do not store commodities just because of their value as an investment, but also because they might need the commodities for production. Hence, they may not want to sell the assets to engage into the above strategy. Consequently, we cannot rule out the inequality (14.23).

We can measure the value that companies associate with having the commodities in storage by the difference between the two sides of the inequality (14.23). Again, this is often done applying continuous compounding. If we define δ implicitly through the equality

$$F(t, T) e^{(T-t)\delta} = \frac{S(t) + L(t, T)}{Z(t, T)}, \quad (14.24)$$

we can think of δ as a measure of the benefits of having easy access to the commodity. It can be interpreted as a dividend yield on the commodity and is often referred to as the **convenience yield** of the commodity. The convenience yield reflects the benefits of carrying the commodity. Other things equal, the convenience yield is higher when current inventories are small. In combination with (14.21), we can then express the forward price as

$$F(t, T) = S(t) e^{(T-t)[\ell^c + y^c(t, T) - \delta]} = S(t) e^{(T-t)\ell_{\text{net}}}, \quad (14.25)$$

where

$$\ell_{\text{net}} = \ell^c + y^c(t, T) - \delta \quad (14.26)$$

is the so-called **net cost of carry**, i.e., the difference between the cost of carry and the benefits of carry.

Typically several forwards on the same commodity but with different maturities are available simultaneously, and then we can consider how the forward prices $F(t, T)$ depend on the maturity date T . If forward prices are decreasing in maturity, the market is said to be in normal backwardation or simply **backwardation**. In particular, since $S(t) = F(t, t)$, the forward prices are then lower than the spot price. We can see from (14.25) that backwardation occurs when the net cost of carry is negative, which means relatively high convenience yields, low interest rates, and low storage costs.

Conversely, if forward prices are increasing in maturity, the market is said to be in **contango**. In this case, forward prices exceed the spot price. This situation requires a positive net cost of carry, i.e., relatively low convenience yields, high interest rates, and high storage costs. Sometimes forward prices are not monotonic in maturity, so a mix of backwardation and contango may occur.

Note that when the underlying is a non-dividend paying financial asset, forward prices are increasing in maturity (for positive interest rates), which implies contango.

The terms backwardation and contango are sometimes used to describe the relation between forward prices for different maturity dates and the expectation today of the spot price at these maturity dates. We know that expected returns should reflect systematic risk. If we let r denote the continuously compounded corresponding required rate of return per year (maybe computed using some CAPM-like model), the expectation at time t of the spot price at time T is

$$E_t[S(T)] = S(t) e^{(T-t)r}.$$

Then we can rewrite the commodity forward price (14.25) as

$$F(t, T) = E_t[S(T)] e^{(T-t)[\ell^c + y^c(t, T) - \delta - r]} = E_t[S(T)] e^{(T-t)(\ell_{\text{net}} - r)}. \quad (14.27)$$

The difference between the net cost of carry and the required return determines whether the forward price exceeds the expected spot price or the other way around.

Finally, note that for commodity forward contracts it is important to specify the quality of the commodity underlying the contract and also whether the contract is settled by cash or by physical delivery of the underlying asset (and in this case the exact time and location of delivery).

14.1.7 Forward markets

Forwards on various underlying assets are traded on semi-organized **over-the-counter (OTC) markets**, which are basically phone- or computer-based networks of mainly financial institutions, corporations, and fund managers. When retail investors enter into a forward contract it is often with their bank as the other party. Forward contracts can be customized to the special needs of an investor.

In general, no collateral is posted to ensure that the parties can actually make their promised payment at the forward maturity date. Being one party in the agreement you should really think about the **counterparty risk**, i.e., the risk that the counterparty does not live up to its obligation.

It would be highly impractical to trade forwards on an organized exchange. The current forward price is a contract parameter in all the forwards initiated on a given trading day for a specific underlying asset and a specific maturity date. The next day the forward price for the same underlying asset and maturity date is likely to be different (e.g. due to a change in the spot price) so trade in completely new contracts have to be opened up. This would quickly lead to a huge number of different forward contracts on the same underlying asset and the same maturity date, which would result in very little subsequent trade in most of the contracts. Next we look at the so-called futures contract which can be seen as a version of the forward contract that can be traded at organized exchanges.

14.2 Futures

14.2.1 Definition and characteristics

Futures contracts or simply **futures** are in many respects similar to forwards, but futures are standardized and often traded at organized exchanges. A futures is, as a forward, a binding agreement between two parties to transact a given asset in the future at terms known today. The main difference is that the gains and losses on futures are regularly settled and then the value of the futures is reset to zero, a procedure known as **marking-to-market**. Typically, futures are marked-to-market at the end of each trading day.

Associated with a futures contract with a given underlying asset and a given maturity date (or final settlement date) is a **futures price** that changes over time, just as the forward price for a fixed maturity and underlying asset changes over time. At maturity the futures price equals the spot price of the underlying asset. Each settlement involves a cash payment equal to the change in the futures price since the previous settlement. If you have a long position in the future (“buy the futures”), then you basically commit to buying the underlying asset. In this case, at each settlement you receive the increase in the futures price since the previous settlement. Of course, if the futures price has declined, you get a negative payment, i.e., you have to make a payment to the counterparty. Conversely, if you have a short position in the future (“sell the futures”), you will receive positive mark-to-market payments when the futures price decreases and you have to make payments when the futures price increases.

Here is an example of the mark-to-market procedure:

Example 14.6

An investor buys a futures on a stock. The contract matures in five trading days. The current (day 0) futures price is \$255. Suppose the futures price at the end of the following trading days are \$248, \$250, \$257, \$254, and \$257, respectively. At the end of each day,

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5
Futures price	255	248	250	257	254	257
Payment to investor	0	-7	+2	+7	-3	+3

Table 14.1: Futures cash flow.

The table shows the cash flow on the futures contract in Example 14.6.

the contract is marked-to-market, which leads to the cash flow illustrated in Table 14.1. Note that the total payout equals \$2, which is simply the difference between the terminal and the initial futures prices.

If the investor took a long position in a forward contract instead of a futures contract and the initial forward price would equal the initial futures price of \$248, the investor would get a terminal payoff of $\$257 - \$255 = \$2$ and no payments at other dates. The total payout of the forward and the futures are thus identical, only the distribution of the payout over time differs.

The futures price at a given point in time is set so that the present value of all future payments is zero and therefore taking a position in futures is costless. Future payments can be negative or positive, of course. Let $\varphi(t, T)$ denote the futures price at time t for a contract with maturity T . The futures price at the final settlement is set equal to the spot price of the underlying asset at that time, $\varphi(T, T) = S(T)$. At the next-to-last settlement date $T - 1$, the futures price $\varphi(T - 1, T)$ is set so that the present value of the terminal payment $S(T) - \varphi(T, T)$ is zero. At this date, there is no difference between a forward and a futures, so the futures price is identical to the forward price, $\varphi(T - 1, T) = F(T - 1, T)$.

At earlier dates, it gets more complicated. At time $T - 2$, the futures price $\varphi(T - 2, T)$ is set so that the present value of the mark-to-market payment $\varphi(T - 1, T) - \varphi(T - 2, T)$ at time $T - 1$ and the terminal payment of $S(T) - \varphi(T - 1, T)$ is zero. Note that this involve the futures price $\varphi(T - 1, T)$ which is still unknown. The futures price $\varphi(t, T)$ set at time t depends on all the future futures price, $\varphi(t + 1, T), \varphi(t + 2, T), \dots, \varphi(T - 1, T)$. Determining the futures price is generally a very involved recursive problem.

Nevertheless, it can be shown that under some conditions, the futures price is identical to the forward price, of course with the same underlying asset and the same maturity date:²

$$\varphi(t, T) = F(t, T). \quad (14.28)$$

When this holds, the formulas for forward prices developed in the preceding section also applies to the corresponding futures contracts.

The futures-forward identity (14.28) is true if interest rates are assumed to be constant until the maturity date, which might be a fair assumption for short-maturity contracts. If interest rates cannot be assumed constant, the result still holds as long as interest rates are uncorrelated with the spot price of the underlying asset in a very specific sense, which we will not try to explain here.

We can build some intuition for why the correlation with interest rates matters for the futures price from Example 14.6. Assume to begin with that the futures price and forward price at day 0 are identical. As we observed above the two contracts then lead to the same

²The original proof is due to Cox, Ingersoll, and Ross (1981b), but can also be found in most textbooks on derivatives such as Hull (2021).

total payment. Suppose the spot price is positively correlated with interest rates. Since futures prices increase with the spot price, other things equal, then futures prices are also positively correlated with interest rates. Having a long futures position, we thus typically receive positive mark-to-market payments when interest rates are high, whereas we need to make payments when interest rates are low. We can thus place positive futures payments at a high interest rate and finance the futures payments we have to make at low interest rates. This is better than having a forward, although it produces the same total payment. To ensure a zero present value of all future payments, the futures price has to exceed the forward price. Conversely, if the spot price is negatively correlated with interest rates, the futures price is smaller than the forward price.

14.2.2 Futures markets

The marking-to-market settlement mechanism implies that investors in futures, unlike investors in forwards, do not need to know when the contract was originally issued and what the futures price was at that date. New contracts with the same underlying and maturity do not have to be issued each day as the original futures contracts are continuously adjusted to reflect the current market situation. Therefore, futures contracts can be traded on organized exchanges and most often they are. In this case, the contracts are standardized with respect to

- (a) the underlying assets: only futures on specific assets are traded
- (b) the contract size: only futures on a certain number of units of the underlying asset are traded
- (c) the maturity dates: only futures with specific, regularly spaced maturity dates are traded, e.g., with one- or three-month intervals.

When you take a position in a futures on an organized exchange, the exchange (or the associated clearing central) takes the opposite position. Since you might have to make payments to the exchange if the futures price evolves unfavorable to you, the exchange is concerned with your ability to actually make those payments. Hence, you have to set up a **margin account** and make a certain deposit when you take the initial futures position. This deposit is called the initial margin. Any gains or losses are then settled via the margin account so that the balance of the account varies over time. The exchange requires a certain minimum balance at any point in time, the so-called maintenance margin. If your balance goes below the maintenance margin, you get a margin call, i.e., you are told to pay in additional funds to satisfy the margin constraint, otherwise your futures position will be eliminated.

Many exchanges are open for trade in certain futures on stock indices, some individual stocks, government bonds, and some particular interest rates. In addition, there are specialized futures exchanges around the world with trade in futures on foreign exchange and many different commodities such as precious metals, oil, electricity, sugar, and frozen concentrate orange juice.

14.2.3 Risk management with futures

Futures are used for managing and hedging risks in a very similar way as forwards. Due to the standardization of the contracts, it may not always be possible to obtain the desired position in futures because of a mismatch in the contract size, the maturity date, or even the underlying asset. Even when a perfect hedging is impossible, a significant risk reduction is often achievable by taken an appropriate position in futures on an asset highly

correlated with the desired asset and with a maturity date close to the desired date. In such cases the **basis** is important. The basis is defined as the difference between (a) the spot price of the asset which the hedger is exposed to and seeks protection towards and (b) the spot price of the underlying asset of the futures used for hedging. In the ideal situation with a perfect match, the basis is zero. Otherwise, the investor faces basis risk, i.e., uncertainty about the future size of the basis.

The presence of the basis risk implies that it is suboptimal to implement a one-to-one hedging in which the futures position involves exactly the same number of units of the underlying asset as the investor wants to hedge. The **hedge ratio** is defined as the ratio of the number of units of the underlying asset in the futures position relative to the number of units of the asset that the investor is exposed to. In the following the hedge ratio is denoted by h . For example, if the investor owns 80,000 barrels of oil and hedges using futures 60 contracts each written on 1,000 barrels of oil, the hedge ratio is $h = 60 \times 1,000 / 80,000 = 0.75$.

A simple and frequently applied method for determining the hedge ratio is **minimum-variance hedging**. Let ΔS denote the change in the spot price of the asset which the investor is exposed to, measured over the period where the investor wants to hedge. Let $\Delta\varphi$ denote the change in the futures price over the same period. With a hedge ratio of h , the total change in the investor's value is

$$\Delta V = \Delta S - h\Delta\varphi$$

or

$$\Delta V = h\Delta\varphi - \Delta S,$$

depending on the circumstances. Here we disregard the mark-to-market settlement feature and thus think of the futures as a true forward contract. In both cases the variance of the value change is

$$\text{Var}[\Delta V] = \text{Var}[\Delta S] + h^2 \text{Var}[\Delta\varphi] - 2h \text{Cov}[\Delta S, \Delta\varphi].$$

Minimizing the variance, the first-order condition is

$$2h \text{Var}[\Delta\varphi] - 2 \text{Cov}[\Delta S, \Delta\varphi] = 0,$$

leading to the minimum-variance hedge ratio

$$h = \frac{\text{Cov}[\Delta S, \Delta\varphi]}{\text{Var}[\Delta\varphi]} = \rho \frac{\sigma_S}{\sigma_\varphi}, \quad (14.29)$$

where σ_S and σ_φ are the standard deviations of changes in the spot and futures prices, respectively, and where ρ is the correlation between these changes. This hedge ratio leads to a variance of

$$\text{Var}(\Delta V) = (1 - \rho^2) \text{Var}(\Delta S). \quad (14.30)$$

Note that the higher the absolute value of the correlation, the lower the minimum variance.

14.3 Swaps

A **swap** refers to an agreement to exchange two specified streams of payments. We focus on two common types of swaps, namely interest rate swaps and currency swaps.

Time	LIBOR in %	From Y to X	From X to Y	Net from Y to X
Now	2.3	0	0	0
After 1 year	2.6	23,000	25,000	-2,000
After 2 years	3.0	26,000	25,000	1,000
After 3 years	2.9	30,000	25,000	5,000
After 4 years	(irrelevant)	29,000	25,000	4,000

Table 14.2: Interest rate swap cash flow.

The table shows an example of the cash flows generated by an interest rate swap.

14.3.1 Interest rate swaps

An **interest rate swap** is an exchange of two payment streams that are determined by certain interest rates. In the simplest and most common interest rate swap, a *plain vanilla* swap, two parties exchange a stream of fixed interest rate payments and a stream of floating interest rate payments. The payments are in the same currency and are computed from the same (hypothetical) face value or notional principal. The floating rate is usually a specific interbank rate, possibly augmented or reduced by a fixed margin. The interbank rate has traditionally been a LIBOR rate, and we will also refer to LIBOR rates in the examples below, but note that LIBOR rates are being replaced by other benchmark interbank rates, cf. the discussion in Section 1.3.2. The fixed interest rate is usually set so that the swap has zero net present value when the parties agree on the contract. While the two parties can agree upon any maturity, most interest rate swaps have a maturity between 2 and 10 years. The following example illustrates the cash flows of a swap.

Example 14.7

Two companies X and Y agree on a 4-year swap with annual interest payments derived from a face value of \$1,000,000. At every payment date, company X has to pay 2.5% of the face value, i.e., \$25,000, to company Y. The interest rate defining the payment from Y to X is the 1-year LIBOR rate at the beginning of each payment period. Hence, these payments are known one year in advance, but not earlier. If we suppose that the 1-year LIBOR rate is 2.3% at the date where the agreement is made, the payment from Y to X after the first year is \$23,000. Table 14.2 shows an example of the future LIBOR rates and the resulting swap payments.

The party paying a stream of fixed rate payments and receiving a stream of floating rate payments is said to have a **pay fixed, receive floating swap** or a **fixed-for-floating swap** or simply a **payer swap**. The counterpart receives a stream of fixed rate payments and pays a stream of floating rate payments. This party is said to have a **pay floating, receive fixed swap** or a **floating-for-fixed swap** or simply a **receiver swap**. Note that the names payer swap and receiver swap refer to the fixed rate payments.

Let us briefly look at the uses of interest rate swaps. A borrower can transform a floating rate loan into a fixed rate loan by entering into an appropriate swap, where he receives floating rate payments (netting out the payments on the original loan) and pays fixed rate payments. This is called a **liability transformation**. Conversely, a lender having lent money at a floating rate, i.e., owning a floating rate bond, can transform this to a fixed rate bond by entering into a swap, where he pays floating rate payments and receives fixed

Company	fixed rate loan	floating rate loan
High	7%	6 month LIBOR + 1.0%
Low	6%	6 month LIBOR + 0.4%

Table 14.3: Offered interest rates in Example 14.8.

The table shows the assumed interest rates offered on a fixed or a floating rate loan to two companies.

rate payments. This is an **asset transformation**. Hence, interest rate swaps can be used for hedging interest rate risk on both (certain) assets and liabilities. On the other hand, interest rate swaps can also be used for taking advantage of specific expectations of future interest rates, i.e., for speculation.

Swaps are often said to allow the two parties to exploit their **comparative advantages** in different markets. Concerning interest rate swaps, this argument presumes that one party has a comparative advantage (relative to the other party) in the market for fixed rate loans, while the other party has a comparative advantage (relative to the first party) in the market for floating rate loans. Here is an example:

Example 14.8

Two companies High and Low want a four-year loan of \$1,000,000. High prefers a fixed rate loan, whereas Low prefers a floating rate loan. The companies are offered the interest rates shown in Table 14.3. Although High must pay a higher rate than Low on both fixed and floating rate loan, High has a comparative advantage in floating rate loans where the interest rate differential is only 0.6 percentage points compared to 1.0 percentage points on fixed rate loans. Conversely, Low has a comparative advantage in fixed rate loans. Unfortunately, each company has a comparative disadvantage in its preferred type of loan.

Both parties can obtain their preferred loans and benefit from their comparative advantages as follows. High takes a floating rate loan and Low a fixed rate loan, and then they enter into an interest rate swap. Figure 14.2 illustrates a swap benefitting both parties. Low commits to paying the LIBOR rate to High, and High commits to paying a fixed rate of 5.8% to Low. In effect, High then has a fixed rate loan with an interest rate of 6.8%, whereas Low has a floating rate loan with an interest rate equal to the 6-month LIBOR rate plus 0.2 percentage points. Note that in this case both parties save 0.2 percentage points relative to the interest rate they were offered on their preferred loan type. The total savings of 0.4 percentage points is identical to the distance between the interest rate differentials in the two markets, namely $7\% - 6\% = 1\%$ on fixed rate loans and $(\text{LIBOR} + 1.0\%) - (\text{LIBOR} + 0.4\%) = 0.6\%$ on floating rate loans.

Often a swap is set up by a bank or another financial intermediary. Figure 14.3 illustrates a structure where a bank enters a swap with both parties. Here, the bank effectively earns a fee of 0.2%, whereas Low ends up with a floating rate loan with a rate of LIBOR plus 0.3 percentage points, and High ends up with a loan with a fixed rate of 6.9%. Both Low and High are thus still better off than if they directly accepted the offered rates on their preferred loan type.

The markets for fixed rate loans and floating rate loans are well integrated, and the exis-

tence of comparative advantages conflicts with modern financial theory and the efficiency of the money markets. Apparent comparative advantages can be due to differences in default risk premiums. In the example, the reason High has to pay a higher interest rate than Low must be that there is a higher risk that High cannot make the promised payments on a loan compared to Low. Hence, lenders demand a higher default risk premium on loans to High than on loans to Low. But why is the default risk premium on floating rate loans less than on fixed rate loans? The offered floating rates apply to the first 6-month period, and the lender can adjust the add-on to the LIBOR rate before the next 6-month period if the default risk changes. Therefore, the premium on the LIBOR rate reflects only the default risk over the first 6 months, whereas the fixed rates reflect the default risk over the full maturity. In the example, if High takes the floating rate loan, it may have to pay a higher premium on LIBOR before the loan matures than the current 1% premium. If that happens, the interest rate swap arrangement can end up being unfavorable to High. The apparent comparative advantage is illusionary. For details we refer the reader to the discussion in Hull (2021, Ch. 7).

Next, we discuss the valuation of swaps. Assume that the fixed rate payments and the floating rate payments occur at exactly the same dates throughout the life of the swap. This is true for most, but not all, traded swaps. For some swaps, the fixed rate payments only occur once a year, whereas the floating rate payments are quarterly or semi-annual. The formulas below can be adapted to such swaps. Denote the payment dates by $1, 2, \dots, n$ and assume that the swap contract is made at time 0. Let H denote the face value of the swap. For now let K denote the swap's fixed interest per period.

Each of the fixed payments equals HK and the time 0 value of the payments is

$$V_0^{\text{fix}} = \sum_{i=1}^n HKZ_{0,i} = HK \sum_{i=1}^n Z_{0,i}, \quad (14.31)$$

where $Z_{0,i}$ is the price at time 0 of a zero-coupon bond promising a payment of 1 at time i . Note that this zero-coupon bond refers to a promise made by the party paying the fixed rate payments in the swap and the zero-coupon bond price should therefore reflect the default risk of this party.

The floating interest payment to be made at time i equals the product of the face value H and the one-period interest rate $y(i-1, i)$ set at time $i-1$ for the period until time i . It might seem difficult to compute the present value of these yet unknown future payments. However, notice that the payments are equivalent to interest payments on a floating rate bond. In Section 5.10.1 we argued that exactly when the coupon rate is reset, the present value of a floating rate bond is equal to the face value. This includes the present value

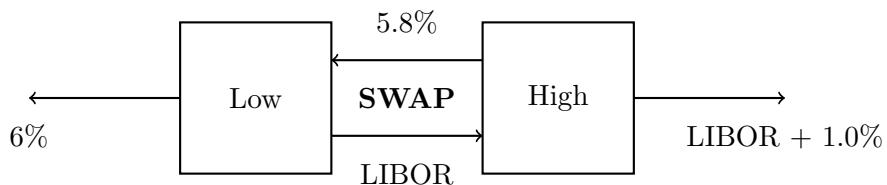


Figure 14.2: Exploiting comparative advantages.

The figure illustrates the construction of an interest rate swap that allows companies to exploit the comparative advantages in different loan markets. The figure refers to Example 14.8.

of the repayment of the face value at the maturity date n . In a swap, there is no such repayment at maturity so we have to subtract its present value. We conclude that the value at time 0 of the floating rate payments on a swap is

$$V_0^{\text{fl}} = H - HZ_{0,n} = H [1 - Z_{0,n}]. \quad (14.32)$$

The above arguments require that the both the interest rates $y(i-1,i)$ and the zero-coupon bond prices $Z_{0,i}$ reflect the default risk of the party promising to make the floating rate payments in the swap. Note that the same price $Z_{0,n}$ occurs in both the value of the fixed payments and the value of the floating payments. This assumes that the two parties have identical default risks.

It follows from the two preceding formulas that the time 0 of a payer swap is

$$\mathbf{P}_0 = V_0^{\text{fl}} - V_0^{\text{fix}} = H [1 - Z_{0,n}] - HK \sum_{i=1}^n Z_{0,i} = H \left(1 - \sum_{i=1}^n Y_i Z_{0,i} \right), \quad (14.33)$$

where

$$Y_i = \begin{cases} K, & i = 1, 2, \dots, n-1, \\ 1+K, & i = n \end{cases}$$

is equivalent to the payments on a bullet bond with a coupon rate of K and a face value of 1. The present value of the payer swap is thus equal to the face value minus the price of a specific bullet bond. The value of a receiver swap is simply the negative of the value of the payer swap, i.e.,

$$\mathbf{R}_0 = V_0^{\text{fix}} - V_0^{\text{fl}} = H \left(\sum_{i=1}^n Y_i Z_{0,i} - 1 \right). \quad (14.34)$$

The **swap rate** y_{swap} at time 0 for a swap with payments dates $i = 1, 2, \dots, n$ is defined as the unique value of the fixed rate that makes the present value of a swap starting at time 0 equal to zero. We determine the swap rate by solving the equation $\mathbf{P}_0 = 0$ (or, equivalently, $\mathbf{R}_0 = 0$) for K . The solution is

$$y_{\text{swap}} = \frac{1 - Z_{0,n}}{\sum_{i=1}^n Z_{0,i}}. \quad (14.35)$$

The swap rate is sometimes called the equilibrium swap rate or the par swap rate.

If we consider simultaneous swap rates for different maturities, we get a **term structure of swap rates**. Many financial institutions participating in the swap market offer swaps

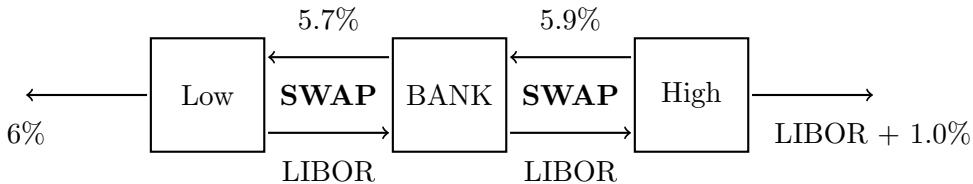


Figure 14.3: Exploiting comparative advantages.

The figure illustrates the construction of a pair of interest rate swaps that allows two companies and an intermediary to exploit the comparative advantages in different loan markets. The figure refers to Example 14.8.

Company	USD loan	DKK loan
Danish Airways	7.6%	6.2%
Gamma Airlines	7%	6%

Table 14.4: Interest rates offered in Example 14.9.

The table shows the assumed offered interest rates on loans in USD and DKK, respectively.

of varying maturities under conditions reflected by their posted term structure of swap rates.

Note that after the swap agreement has been made, the value of the swap will typically deviate from zero as the discount factors and associated discount rates change.

14.3.2 Currency swaps

A **currency swap** or FX swap is an exchange of a payment stream in one currency with a payment stream in a different currency. In the simplest version, each payment stream is equivalent to the payments on a bullet bond or fixed rate loan. Both interest payments and the final repayment of the face value are swapped. Note that all the payments are known in advance in the specified currency. However, the value measured in one currency of a payment fixed in the other currency varies with the exchange rate. Currency swaps typically have a maturity between 5 and 10 years.

Currency swaps are used for hedging long-term exchange rate risk and also for speculating in a certain change in the exchange rate. As for interest rates, a comparative advantage argument is often seen in relation to currency swaps. Here is an example:

Example 14.9

The Danish company Danish Airways has a steady income in U.S. dollars (USD). It plans to take a 5-year loan of 1,000,000 USD which will reduce its net dollar income and therefore its net exposure to the exchange rate between U.S. dollars and Danish kroner (DKK). On the other hand, the US-based company Gamma Airlines wants to take a 5-year loan of 5,500,000 DKK. We assume that the current exchange rate is 5.50 DKK per USD. The companies are offered the interest rates shown in Table 14.4.

Danish Airways has a comparative advantage in loans denominated in Danish kroner, whereas Gamma Airlines has a comparative advantage in loans denominated in U.S. dollars. The two companies agree that Danish Airways borrows 5,500,000 DKK at an interest rate of 6.2%, and Gamma Airlines borrows 1,000,000 USD at 7%. The companies exchange the proceeds immediately and further agrees on the currency swap illustrated in Figure 14.4. According to the swap agreement, Danish Airways pays a 7.3% interest on a 1,000,000 USD loan to Gamma Airlines. In return, Gamma Airlines pays a 6.1% interest in a 5,500,000 DKK loan to Danish Airways.

In total, Danish Airways pays a net interest of 0.1% on the 5,500,000 DKK and 7.3% interest on the 1,000,000 USD. This looks better than the offered 7.6% interest on a loan of 1,000,000 USD. Gamma Airlines receives a 0.3% net interest on the 1,000,000 USD and pays a 6.1% interest on the 5,500,000 DKK, which seems more attractive than paying the offered 6.0% on a loan of 5,500,000 DKK. Hence both companies apparently end up

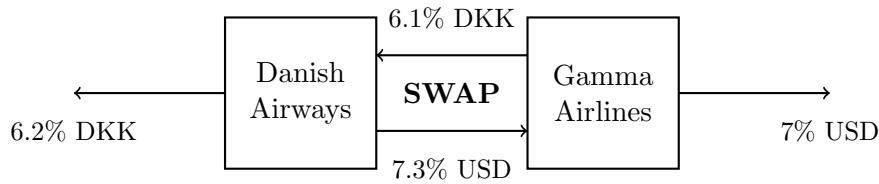


Figure 14.4: A currency swap exploiting the comparative advantages.

The figure illustrates the construction of a currency swap that allows two companies to exploit their comparative advantages in different loan markets. The figure refers to Example 14.9.

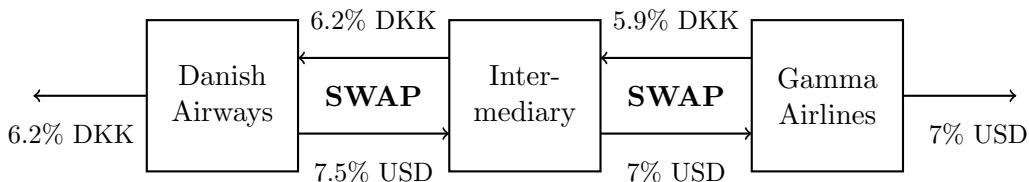


Figure 14.5: A currency swap exploiting the comparative advantages.

The figure illustrates the construction of a pair of currency swaps that allows two companies and an intermediary to exploit their comparative advantages in different loan markets. The figure refers to Example 14.9.

with a cheaper loan than they were offered. Note, however, that they are both exposed to exchange rate risk as the DKK/USD exchange rate may evolve unfavorably.

The exchange rate exposure of the companies is often eliminated by involving an intermediary who is willing to take on the exchange rate risk if sufficiently compensated. The intermediary is typically a bank or another large financial institution. Figure 14.5 shows an example of such an arrangement which really involves two swaps: one between Danish Airways and the intermediary, the other between Gamma Airlines and the intermediary. In total Danish Airways obtains a USD loan at a 7.5% interest rate, and Gamma Airlines obtains a DKK loan at 5.9%. Both companies save 0.1 percentage points relative to the interest rate originally offered in the desired currency. The intermediary receives a net interest of 0.5% of 1,000,000 USD and pays a net interest of 0.3% on 5,500,000 DKK. The intermediary takes the entire exchange rate risk, but may then decide to partly or fully hedge this risk through other transactions.

Just as for interest rate swaps, the comparative advantage motive can be criticized. The question is why the interest rate differential between the two companies is not the same for all currencies. In the above example you might argue that U.S. borrowers are unfamiliar with the Danish company, and therefore require a higher risk premium compared to the U.S. company. And vice versa for the U.S. company.

A currency swap can be seen as an exchange of two bullet bonds in different currencies. In principle, it is therefore simple to determine the value of a currency swap. If the swap is such that you receive payments in your own currency and make payments in the foreign currency, it is like having a long position in a domestic bond and a short position in a

foreign bond. The value at time t is

$$V_t = B_t - S_t B_t^f,$$

where B_t is the value of the domestic bond, B_t^f is the value of the foreign bond in the foreign currency, and S_t is the spot exchange rate in units of domestic currency per unit of foreign currency. Note that if the parties are subject to default risk, this risk should also be reflected by the bond prices.

14.3.3 Swap markets

The swap markets are over-the-counter markets. Most often a company wanting a specific swap contract approaches its bank which might then itself take the role of the counterparty in the swap or find another financial institution willing to do so. Therefore the company does not have to search for a counterparty on its own. Often some or all of the risk exposure caused by the swap is taken by the bank which can then try to balance its exposure by issuing other contracts related to the same risks.

The demand for interest rate swaps and some currency swaps is large enough that somewhat standardized contracts are offered by several financial institutions. In the recent decades the international swap markets have grown explosively both in terms of the number of contracts issued and traded and in terms of the variety of contracts being offered.

14.4 Exercises

Exercise 14.1. An investor wants to buy a six-month forward on a stock. The spot price of the stock is \$50, and the annualized six-month riskfree rate is 5%. Assume the stock does not pay any dividends within the next six months.

- (a) What is the theoretically fair forward price on the stock?

The investor buys the forward with the forward price just found. However, after two months the investor wants to get rid of the forward contract. The spot price of the stock is now \$42, and the annualized four-month riskfree rate is 6%.

- (b) What is the value of the forward contract after the two months?

Exercise 14.2. The spot exchange rate between Euros and Canadian dollars is 0.6835 EUR per CAD. The annualized six-month interest rate is 0.267% on Euros and 0.980% on CAD. What is the six-month forward EUR/CAD exchange rate?

Exercise 14.3. The S&P 500 stock index is currently at 1,787. The index pays a dividend yield of 2% per year, and the one-year riskfree interest rate is 0.5%. The current futures price on the S&P 500 index is 1,758 for contracts maturing in one year.

- (a) Compute the theoretically fair forward price for a 1-year forward on the S&P 500 index.
- (b) If you assume that the theoretically fair futures price on the index equals the theoretically fair forward price, is there an arbitrage opportunity? If so, explain how to make a riskfree profit. Are there any real-life considerations that might make the apparent arbitrage strategy less attractive?

Exercise 14.4. The gold spot price is 1,283 USD per ounce. Suppose the storage costs are 5 USD per ounce per quarter to be paid up front. Further, assume the annualized riskfree rate is 4% using quarterly compounding. What is the theoretical forward price on gold with delivery in three months? What if delivery is in 6 months? Or 9 months?

Exercise 14.5. Soybean meal is the residue left from extracting soybean oil from soybeans and is used as an ingredient in animal food. At the Chicago Mercantile Exchange Soybean Meal is traded at a spot price of 382.6 USD per unit, whereas the futures price for delivery in 12 months is

346.9 USD per unit. Suppose the interest rate over the next year is 1% and that the storage costs are 2%, both calculated with continuous compounding. Estimate the convenience yield of soybean meal and discuss the result.

Exercise 14.6. Two banks consider entering a 3-year interest rate swap with semi-annual payments. In the swap Bank A pays a floating rate to Bank B, whereas Bank B pays a fixed annualized interest rate of r to Bank A. The floating rate that applies at a given date is the six-month money market rate at that date. The hypothetical face value is \$1,000,000. The current annualized money market rates for various maturities, which are also the appropriate discount rates for the two banks, are as follows:

t	0.5	1	1.5	2	2.5	3
$y(0, t)$	0.2%	0.4%	0.6%	0.8%	1.0%	1.2%

- (a) What is the present value of the floating rate payments in the swap?
- (b) What is the present value of the fixed rate payments in the swap if the fixed rate is 1.0%?
- (c) What is the present value of the swap?
- (d) What is the fair swap rate, i.e., the fixed rate that leads to a zero swap value?

CHAPTER 15

Options

This chapter focuses on options, which is a class of derivative securities distinctively different from the forwards, futures, and swaps treated in the preceding chapter. Options and option markets were briefly introduced in Section 1.4. In this chapter we explain the characteristics, applications, and pricing of different types of options.

Section 15.1 introduces the key properties, characteristics, and terminology related to options. Section 15.2 explains how options can be used for either reducing risks (hedging) or taking risks (speculating). The markets for options are described in Section 15.3. The following sections deal with the pricing of options. First, Section 15.4 derives some general properties of option prices that have to hold under very weak assumptions. However, to pin down a unique price for a given option, we need to apply a pricing model. The leading option pricing models are the binomial model and the Black-Scholes model, which are presented in Sections 15.5 and 15.6, respectively. Section 15.7 discusses the returns on option investments. Finally, Section 15.8 offers some concluding remarks.

15.1 Definition, characteristics, and terminology

An **option** is an asset giving the owner the right, but not the obligation, to perform a certain transaction in the future at terms specified today. Typically this transaction is to purchase or sell a given *underlying* asset at a pre-set price at or before a given future date. This pre-set price is referred to as the **strike price** or **exercise price** of the option. For now it may be useful just to think about the underlying asset as a share in a given company but, as we discuss in Section 15.3, options with many other underlying assets are also traded. The owner of the option is said to have a **long** position in the option or simply to be long the option. The issuer (or *writer*) of the option is said to have a **short** position in the option or simply to be short the option.

Options are categorized as call or put options. A **call option** (or simply: a call) is an option giving the owner the right to purchase some underlying asset, whereas a **put option** (or simply: a put) is an option giving the right to sell some underlying asset. If the owner of the option decides to make use of his right to carry out the pre-specified transaction, the option is said to be exercised.

Options are also categorized as European or American options. A **European option** grants its owner the right to execute the transaction in the underlying at a single specified future date, whereas an **American option** gives the right to do so at any date up to

and including some specified future date. Note that the labels ‘European’ and ‘American’ do not indicate where the option is traded but only the time span of the right granted by the option. Both European and American options are traded all over the world. The final date at which an option can be exercised is referred to as the **expiration date** or **maturity date** of the option.

We divide options into four main types: (1) European call options, (2) American call options, (3) European put options, and (4) American put options.

An example is an option granting its owner the right to purchase one share of Apple stocks at a price of \$90 within the next three months. This is an American call option on Apple stocks with an exercise price of \$90 and three months to expiry. If Apple shares sell at \$94 today, the owner of the option could exercise it right away and obtain a payoff of $\$94 - \$90 = \$4$, but he could also wait and hope that the price of Apple shares will increase within the next three months in which case he would obtain a higher payoff. Of course, Apple shares may also drop below \$90 and then the call option would not be exercised.

Note that the owner of an option is not forced to pay anything. If exercising the option would lead to a negative payoff, the owner simply refrains from exercising. On the other hand, with some probability it will be profitable for the owner to exercise the option. When the owner exercises the option, the issuer of the option has to make the payment stipulated in the option contract. Hence, the payoff of the option writer ends up being zero or negative. Therefore, obtaining an option comes at a cost. Upon purchase of the option, the buyer has to pay a cash **premium** to the issuer. This premium must reflect the value of having the option. We are obviously interested in determining what the fair premium or fair price of any given option is, just as we in earlier chapters have discussed fair prices for stocks and bonds.

We will make use of the following notation:

T	the expiration date of the option
$S(t)$	the price at time t of the underlying asset
X	the exercise price
$C(t)$	the price at time t of a European call option with expiration date T and exercise price X
$C^a(t)$	the price at time t of an American call option with expiration date T and exercise price X
$P(t)$	the price at time t of a European put option with expiration date T and exercise price X
$P^a(t)$	the price at time t of an American put option with expiration date T and exercise price X

Note that when writing the option price, we suppress the exercise price and the expiration date of the option.

In what follows we assume that the option gives the right to transact one unit of the underlying asset. In reality options are typically written on a large number of units of the underlying, say 1,000 or 10,000, but option prices are quoted as the price per unit of the underlying asset. Option prices are proportional to the number of units of the underlying so a simple scaling of the price of an option on one unit of the underlying leads to the price of an option on multiple units of the underlying. For example, the price of a call option on 100 shares of Apple stocks (with a certain exercise price and expiration date) is equal to 100 times the price of a call option on one share of Apple stock (with the same exercise price and expiration date). This holds whether the option is European or American.

Let us be specific about the potential payoffs of the different types of options. By exercising a European call option at the expiration date T , the owner receives one unit of the underlying asset having a market value of $S(T)$. In return the owner has to pay the exercise price X so the net payoff to the owner is $S(T) - X$. Obviously, the owner will only exercise his right if it leads to a positive payoff, i.e., if $S(T) > X$, in which case his net payoff is $S(T) - X$. This is what he saves by exploiting his right to purchase the underlying asset at the pre-set price X instead of the prevailing market price $S(T)$. If $S(T) < X$, exercise would lead to a negative payoff, so the owner of the call option refrains from exercising. If $S(T) = X$, the owner is indifferent between exercising and not exercising since he ends up with nothing anyway. We can write the terminal payoff (which equals the terminal value) of the European call option compactly as

$$C(T) = \max(S(T) - X, 0). \quad (15.1)$$

The profit to the option owner is the difference between the payoff and the premium originally paid for the option.¹

Similarly, the terminal payoff and value of the European put option can be written as

$$P(T) = \max(X - S(T), 0) \quad (15.2)$$

since the owner of the put option will only exercise his right to sell the underlying asset at the price of X if the market price is lower, i.e., $S(T) < X$.

The terminal payoffs and the profits from European options are illustrated graphically in Figure 15.1. Note that the payoff and profit to the issuer (with the short position in the option) is simply minus the payoff and profit to the owner (with the long position). The issuer receives the initial premium, but then has to pay the option payoff to the owner upon exercise. For American options similar diagrams apply at any date where the option is exercised.

Next, we introduce additional terminology often used in relation to options. The value of an option at a given date is sometimes decomposed into an intrinsic value and a time value. The **intrinsic value** is the value of the option if it expired at that date, whereas the **time value** is the remaining value of the option which must reflect the value of the possibility that the option payoff becomes better than it is right now. For example, for a call option with a value of $C(t)$ at time $t \leq T$ the intrinsic value is $\max(S(t) - X, 0)$ and the time value is residually determined as $C(t) - \max(S(t) - X, 0)$.

An option is said to be **in the money** if exercising it immediately would lead to a positive payoff, that is if its intrinsic value is strictly positive. Hence, a call option is in the money if the current price of the underlying asset exceeds the exercise price. In contrast, a put option is in the money if the current price of the underlying asset is below the exercise price. Similarly, an option is said to be **out of the money** if exercising it immediately would lead to a negative payoff and **at the money** if exercising it would lead to a zero payoff.

Options can be settled either physically or in cash. **Physical settlement** means that when the option is exercised the underlying asset is physically delivered by one party to the other party or at least the formal ownership of the underlying asset is transferred. In return the other party pays the exercise price in cash or by an appropriate bank transfer.

¹The payoff and the premium are paid at different points in time. Therefore, when computing the profit, we should discount the payoff back to the purchase date or discount the premium forward to the exercise date. But most options are issued and purchased with a relatively short time to expiration (often a few months) so the discounting would typically not matter much and is left out in this context.

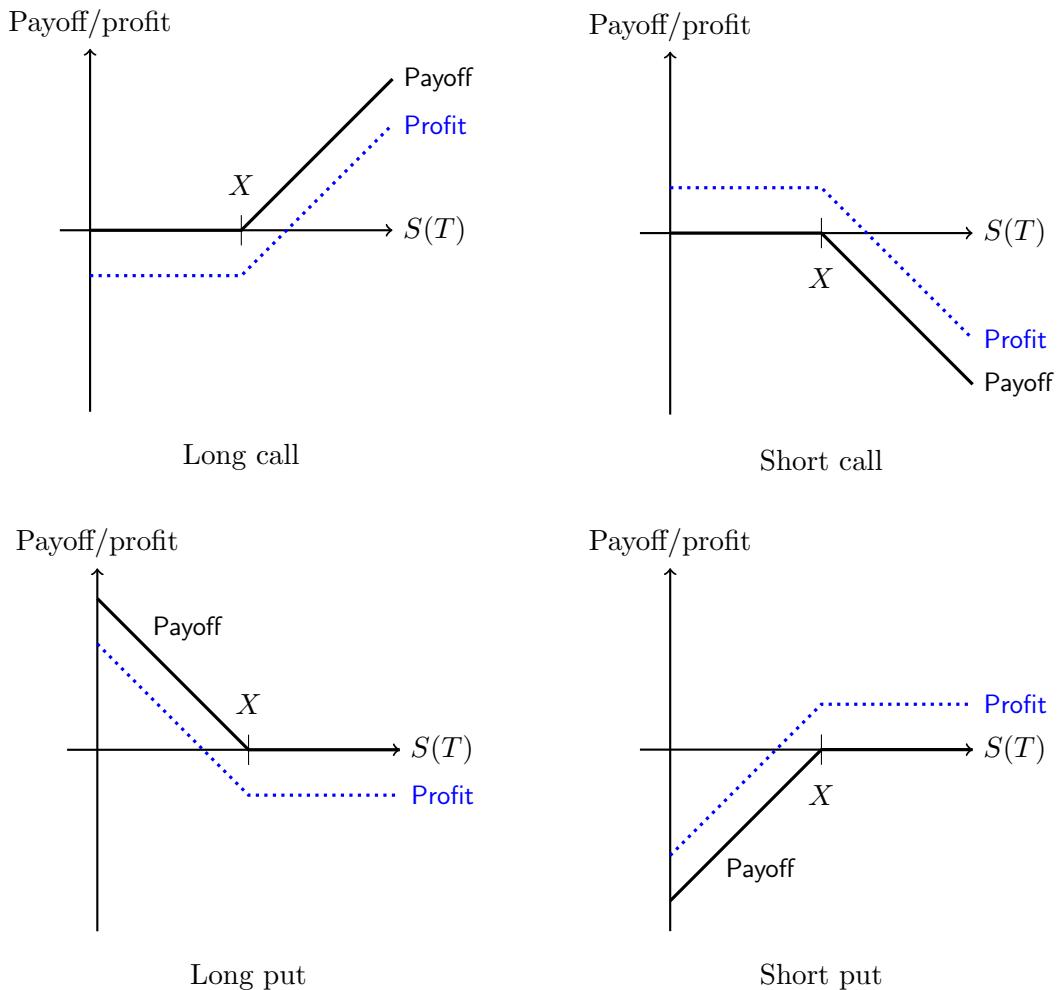


Figure 15.1: Option payoffs and profits.

The figure shows payoffs and profits from positions in single options. The black solid line shows the payoff of the option and the dotted blue line the profit (payoff minus price), in both cases as a function of the price of the underlying asset at the option maturity date.

In contrast, a **cash settlement** involves only a cash payment or bank transfer of the difference between the price $S(T)$ of the underlying asset and the exercise price X . For exchange-traded options the manner of settlement is clearly stated in the contract terms.

Some options are not really written on an underlying asset but on some other underlying variable, like the value of a specific interest rate, which would then play the role of $S(T)$ in the payoff formulas above. Such options are necessarily settled in cash. A call option on an interest rate is simply a contract with a payoff of the form (15.1). Likewise, a put option on an interest rate is a contract with a payoff of the form (15.2).

15.2 Applications

Options are used for hedging (risk reduction) and speculation (risk seeking).

15.2.1 Hedging

Let us consider the **hedging** application first, which is particularly transparent when the underlying asset is a commodity. A simple example illustrates the idea. A company producing popcorn estimates that it is going to need 50 tons of corn to cover production in the first quarter of next year. The company can do (at least) the following three things:

1. *Naked (unhedged) position:* The company does nothing now and then purchases the corn just before it is needed in the production at the spot price of corn prevailing at that date. (The term spot price refers to the price for immediate delivery.) But, seen from today, this approach involves substantial risk as the company does not know what the spot price at the purchase date will be. It may end up higher than expected and thus increase the costs of the company. If we let $S(T)$ denote the spot price of the 50 tons of corn when the company needs it, the payoff to the company with this strategy is $-S(T)$.
2. *Hedging with forwards:* The company buys a forward contract on the 50 tons of corn and thus effectively locks in the price to be paid for the corn. A long forward contract with delivery price X gives a terminal payoff of $S(T) - X$ so the total payoff to the company is $-S(T) + [S(T) - X] = -X$, i.e., the purchase price is locked in at X . If X equals the current forward price, the company does not pay anything to enter the forward contract.
3. *Hedging with options:* The company buys a call option on corn and thus obtains the right to buy the corn when needed at a price fixed today. With options, the company still benefits if the corn spot price falls, since then the option simply expires worthless and the company can buy the corn at the low spot price. If the corn spot price increases, the call option compensates the company for the increase in the spot price above the exercise price of the option. In symbols, the call option provides a payoff of $\max(S(T) - X, 0)$ which together with the naked position leads to a total payoff of

$$-S(T) + \max(S(T) - X, 0) = \max(-X, -S(T)) = \begin{cases} -X, & \text{if } S(T) > X, \\ -S(T), & \text{if } S(T) \leq X, \end{cases}$$

so the company ends up paying at most X and possibly less.

The three strategies are illustrated in the left panel of Figure 15.2. The naked strategy is the best if the future spot price ends up below X , whereas the forward is best if the spot price exceeds X , but each of these strategies might also turn out to be much worse than

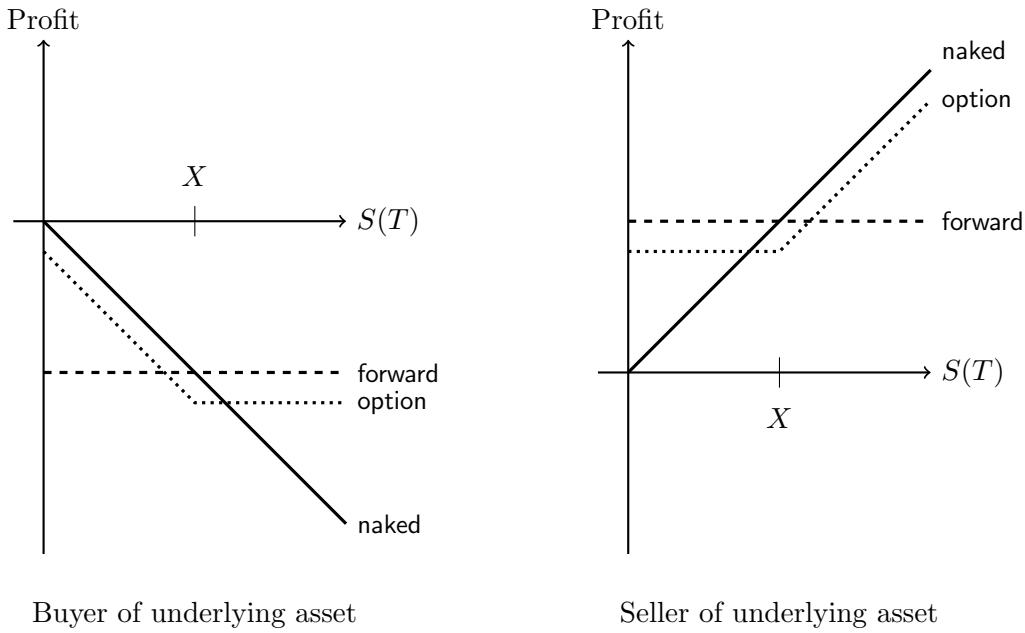


Figure 15.2: Profits from different hedging strategies.

The left diagram refers to the case corresponding to a short position in the underlying asset. The right diagram is for the case with a long position in the underlying asset. See the main text for an explanation of the different lines.

the other. The option hedging strategy limits the downside risk if the spot price ends up high and still delivers a decent net result if the spot price ends up low.

In describing the hedging strategies above, we assumed that the company knew precisely which quantity of the underlying asset it needs and exactly when it needs it. This information is necessary to exactly lock in the price using forwards. But maybe the 50 tons is just an estimate. Maybe it turns out that 60 tons of corn is needed, and then the company still might risk losing money despite buying the forward contract on 50 tons. On the other hand, buying a forward on 60 tons might lead to a loss if only 50 tons are needed. The option strategy is easier to accommodate to such uncertainty. By purchasing a call option on 60 tons of corn instead of 50 tons, any downside risk is eliminated even if the company should end up wanting 60 tons. But, of course, the call option on 60 tons of corn is more expensive than the call option on 50 tons (at least if they have identical exercise prices and expiration dates). Likewise, there might be uncertainty about exactly when the company needs the corn. Also in this case hedging with a forward cannot completely remove the downside risk, whereas options are more flexible. If the company might need the underlying at any time during the next six months, it can obtain an American call expiring in six months.

Conversely, a company producing a certain commodity fears that the spot price of this good drops before the company is ready to sell it. It can lock in the selling price by taking a short position in a forward on that commodity. Again, if the quantity and maturity date of the forward exactly matches the need of the producer, the company has eliminated all price risk, but also the chance of making a profit from a favorable development in the spot price of the commodity. In case of uncertainty about the quantity produced and the exact selling date of the commodity, the forward does not eliminate all risks. By purchasing put options on an appropriate quantity of the commodity and with an appropriate exercise

date, the company can eliminate any downside price risk and maintain the chance to make a profit, but of course the option premium paid should be considered as well. The hedging strategies for this case are illustrated in the right panel of Figure 15.2.

To sum up, options are a more flexible hedging instrument than forwards (or futures). Used appropriately, options can remove any downside risk, and still leaves a chance of a profit. On the other hand, the hedger has to pay up front for the options.

In the examples above the hedger was exposed to the price of a given commodity and therefore could consider purchasing a commodity option. It is easy to come up with similar examples where a company or other investor is exposed to risks related to exchange rates, interest rates, bond prices, or stock prices, and might then consider investing in options written on such quantities in order to reduce or eliminate these risks.

For example a German company having to pay 2,500,000 USD in three months faces the risk of an unfavorable development in the Euro/dollar exchange rate (more Euros per dollar), since it then effectively would have to pay more in Euro terms. It can eliminate the risk by purchasing a call option on the stated amount of USD with an appropriate exercise price. Note that such a foreign exchange option is really the option to exchange a certain amount of one currency to a certain amount of another currency. A call option on 1 USD with an exercise price of, say, 1.25 EUR/USD gives the holder the right to obtain 1 USD by delivering 1.25 EUR. This is equivalent to the right of selling $1/1.25 = 0.8$ EUR and receiving 1 USD, which from a Euro-based investor is a put option on Euro struck in USD. While this may seem confusing at first, note that an option always involves the right to exchange two assets. A call option on gold grants the owner the right to obtain a certain quantity of gold in exchange of some pre-specified cash amount of a given currency.

Many stock investors use stock options to adjust their exposure to stock prices. If you consider investing in a given stock, you probably believe that the company issuing the stock is going to perform better than general market expectations in the near future. But still you might fear that the stock price might fall for some reason. Then, in addition to the stock, you can buy a put option on the stock. If the stock price $S(T)$ at option maturity ends up lower than the exercise price X , you exercise the put and receive $X - S(T)$ so that your total value is X . If the stock price ends up above the exercise price, you simply let the put option expire so that your total value is $S(T)$. Effectively, you ensure a lower bound on the future value of your stock position. This is referred to as a **protective put** strategy. The terminal value is illustrated in the left panel of Figure 15.3. The gross value is simply the total value of the stock and the put when the put expires. The net value is the gross value less the premium paid up front for the put. The kinked line labeled net profit shows the net value less the initial purchase price of the stock. Clearly, by following a protective put strategy, the investor benefits if the underlying stock price ends up high and suffers a modest and limited loss if the stock price ends up low or moderate. The protective put establishes a lower bound on the value of the position but, in return, gives up some value for higher stock prices.

Another frequently used strategy is a **covered call** which involves buying a stock and writing a call option on that stock. This strategy is often followed by the issuers of call options on stocks. An investor only issuing (or shorting) a call option receives the initial option premium, but faces an unlimited potential loss. If the price of the underlying stock booms, the holder of the call option is entitled to a large payment of $S(T) - X$ from the option issuer. To reduce the risk of having to make such large future payments on the option, the issuer can purchase and hold on to the stock until the option matures. Should the option finish in the money, the option issuer can simply sell the underlying stock at the high price and use the proceeds to pay off the option holder. The outcome is

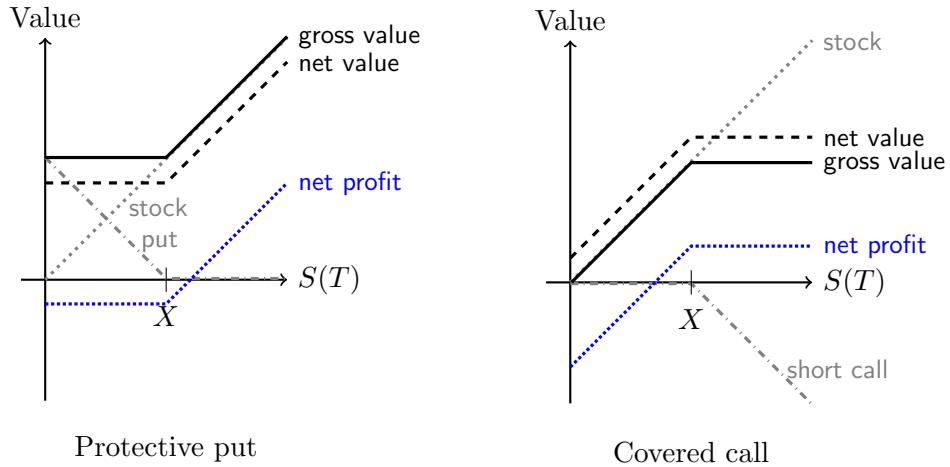


Figure 15.3: Hedging strategies involving stock options.

The left diagram represents the case where the investor owns or buys a stock and buys a put option on the stock. The right diagram is for the case where the investor owns or buys a stock and sells a call option on the stock. See the main text for an explanation of the different lines.

illustrated in the right panel of Figure 15.3. The gross value is the total value of the stock and the written call at the expiration date of the call. The net value is the gross value plus the premium received when selling the call option initially. Again the kinked line labeled net profit shows the net value less the initial purchase price of the stock. By following a covered call strategy, the investor makes a modest profit if the price of the underlying stock ends up moderate or high and suffers a (limited, but possibly substantial) loss if the underlying stock price ends up low. The covered call strategy is often used by existing owners of the stock who are willing to give up some of the capital gains should the stock price boom in return for receiving the call premium.

15.2.2 Speculation

Investors also use options for speculation. An investor expecting an increase in the price of a given stock can profit from investing in call options on the stock. Of course, he could just buy the stocks today and wait for the price to increase, but because the option premium is typically much lower than the stock price, the percentage return is much larger using the options, at least if his expectations are realized.

Suppose, for example, that Apple stocks today sell at \$90 per share, but you expect them to increase to \$120 over the next six months. Buying the stocks, you would expect to realize a return of 33.3%. You could also buy call options maturing in six months with an exercise price of \$90. If your expectations are correct, the option gives a payoff of $\$120 - \$90 = \$30$. Say the options are priced at \$10. Then your return is 200%, much higher than the return on the stock investment. Of course, the option investment is also much more risky. If apple stocks fall to \$80 per share, the option would be worthless, so if you have invested in options, you would lose your entire investment, corresponding to a return of -100% . Had you invested in the stock itself, your loss would be \$10 on a \$90 investment, that is a return of -11.1% .

Continuing this example, suppose you plan to invest \$90,000 over six months in some way so that you benefit if the price of Apple stocks increase in that period. Consider the

Stock price	All stock		All options		Bonds + options	
	Value	Return	Value	Return	Value	Return
50	50,000	-44.4%	0	-100%	80,800	-10.2%
60	60,000	-33.3%	0	-100%	80,800	-10.2%
70	70,000	-22.2%	0	-100%	80,800	-10.2%
80	80,000	-11.1%	0	-100%	80,800	-10.2%
90	90,000	0%	0	-100%	80,800	-10.2%
100	100,000	11.1%	90,000	0%	90,800	0.9%
110	110,000	22.2%	180,000	100%	100,800	12.0%
120	120,000	33.3%	270,000	200%	110,800	23.1%
130	130,000	44.4%	360,000	300%	120,800	34.2%

Table 15.1: A comparison of investment strategies.

The table shows how the terminal value and corresponding rate of return depend on the stock price for three different investment strategies. The initial stock price is assumed to be 100. The first strategy is to invest all your money in the stock. The second strategy is to invest all your money in call options on the stock. The third strategy is to invest in a combination of riskfree bonds and call options on the stock. See the main text for more information.

following three alternatives:

1. *All stocks*: With a current stock price of \$90 per share, you can purchase 1,000 shares.
2. *All options*: With a current call premium of \$10, you can purchase 9,000 options.
3. *Bonds + options*: You spend \$80,000 on bonds that offers a rate of return of 1% (non-annualized) over the next six months, and you spend the remaining \$10,000 on buying 1,000 call options on Apple stocks.

Table 15.1 shows the values and the corresponding rates of return on the three strategies for various possible Apple stock prices in six months. The returns are illustrated in Figure 15.4. The “all options” strategy has by far the largest upside potential, but is also most risky since you lose your entire investment for a wide range of possible stock prices. The “all stocks” strategy has both a strong upside potential and a significant risk of losing a substantial fraction of the investment. The “bonds plus options” strategy has a lower upside potential but also less risk since the worst-case scenario is to lose ‘only’ 10.2% of the investment. None of the three strategies dominate the other two for all stock prices. Each strategy has a range of stock prices for which it is the most profitable strategy. So which of the strategies you should prefer, depends on your willingness to take risks and your expectations about the future price of Apple stocks.

15.2.3 Strategies involving multiple options

Various payoff structures can be constructed with strategies involving multiple options. An example is a **straddle** which consists of a long position in both a call option and a put option with the same underlying asset, exercise price, and maturity date. The solid line in the left panel of Figure 15.5 illustrates the payoff of a straddle, whereas the dashed line shows the profit. Such a strategy is profitable if the price of the underlying asset at option expiry is either far above or far below the exercise price of the options. If you expect significant news about the company issuing a stock, but you are uncertain whether the news will be positive or negative, then a straddle might be a strategy to consider.

Another example is a **butterfly spread** which consists of a long position in a call option with a relatively low exercise price X_1 and a call with a relatively high exercise

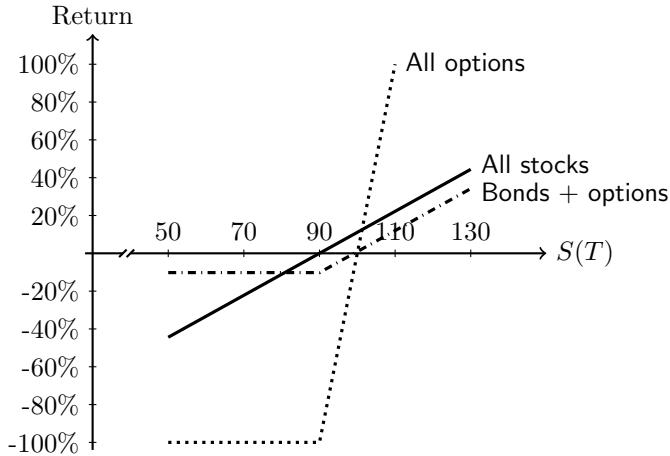


Figure 15.4: A comparison of investment strategies.

The figure shows how the rate of return depends on the stock price for three different investment strategies. Illustrated by the solid line, the first strategy is to invest all your money in the stock. The second strategy, represented by the dotted line, is to invest all your money in call options on the stock. The third strategy, depicted by the dash-dotted line, is to invest in a combination of riskfree bonds and call options on the stock. See the main text for more information.

price X_3 , together with a short position in two calls with an intermediate exercise price of $X_2 = (X_1 + X_3)/2$. All the options share the same underlying asset and maturity date. In the right panel of Figure 15.5 the gray dotted and dash-dotted lines depict the payoffs of the individual options and the solid black line shows the payoff of the combined strategy. As the total payoff is never negative and may turn out positive, the price of the combined strategy has to be positive. The dashed black line shows the profit (the payoff less the price) from the butterfly spread. If you firmly believe that the price of the underlying asset ends up close to X_2 , a butterfly spread might be an investment strategy to pursue.

Many other option strategies are discussed in specialized textbooks on derivative securities, see, e.g., Hull (2021, Ch. 12).

15.3 Option markets

Apparently the first option-style deal ever took place in ancient Greece.² The Greek philosopher and scientist Aristotle tells the story in his books *Politics* published in 332 BC. The intelligent but poor Thales from the city of Miletus lived approximately from 624 to 546 BC. A philosopher and scientist like Aristotle, Thales was known as one of the seven wise men of Greece. According to the story, one winter Thales concluded from observations of the weather conditions that next summer's olive harvest would be exceptionally good and, therefore, olive presses would be in high demand. He did not have sufficient funds for buying or producing olive presses himself, but in the winter he approached owners of existing olive presses and for small up-front fees they were willing to grant him the rights to use the presses the following summer at a fixed and fairly low price. In effect, Thales bought call options on olive presses. Indeed, the summer harvest was abundant, and Thales could charge the olive growers a high price for the use of the presses and thus made a sizeable profit. While it is generally not difficult to make money if you are able to

²The short description of the history of options given here is based primarily on Poitras (2009a, 2009b).

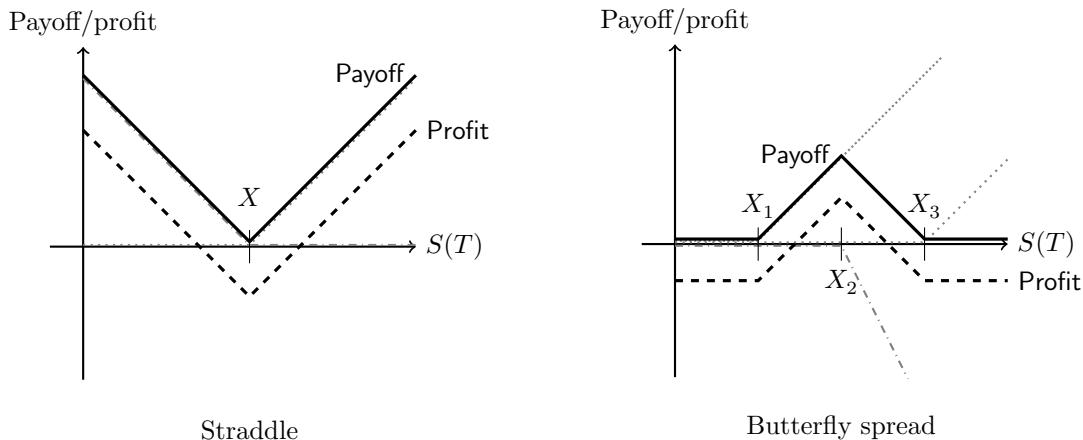


Figure 15.5: Payoffs and profits from option strategies.

The left diagram shows the payoff and profit of a straddle, i.e. a portfolio of a call and a put option on the same asset and with identical exercise prices X and maturity dates T . The right diagram shows the payoff and profit of a butterfly spread, a specific portfolio of call options on the same asset and with identical maturity dates T . The butterfly spread has a long position in a call with exercise price X_1 and a call with exercise price X_3 , and a short position in two calls with exercise price X_2 , where $X_2 = (X_1 + X_3)/2$.

predict the future, the story is an example that you can do it with options at a relatively low price—and with a limited downside in the case that your prediction fails.³

Somewhat organized option markets existed for shorter periods in more recent history. For example, commodity options were presumably traded by speculators on the Antwerp Exchange in the 16th century. In the 17th century, stock options were traded on the Amsterdam Bourse primarily by speculators, and there are also indications of option-like contracts on tulip bulbs trading during the famous Dutch tulip mania in 1636-1637 where prices of tulip bulbs grew to extremely high levels. Organized markets for call and put options on stocks also existed in London from 1687 (maybe somewhat earlier) to 1733, where legislation banned option trading which nevertheless seemed to have continued. Stock options traded on the Paris bourse in the 19th century. In the United States trading in stock options apparently began in 1790. Option-like commodity contracts were traded in Chicago from the mid 19th century, and both call and put options on stocks were traded over the counter in New York from 1872. Option trading has often been controversial and blamed for contributing to financial crises, and has thus been made illegal for various periods in different countries.

The modern era of options dates back to 1973 where the Chicago Board of Exchange (CBOE) and the Options Clearing Corporation were formed and from April 26 introduced exchange-based trading of call options on 16 stocks. In 1975, other U.S. exchanges began trading call options on stocks, and put options were added in 1977. Non-U.S. option markets soon followed and introduced trading in options in selected commodities, government bonds, and stock indices in addition to the more traditional options on individual stocks. The markets have since experienced substantial innovation in terms of the underlying assets or variables upon which options are traded and variations in the exact contract terms. Options are now available for underlying assets or variables such as

³Thales was also known for predicting the solar eclipse in 585 BC, but apparently he did not know how to profit on that prediction.

- stocks in individual companies,
- broad stock indices,
- a measure of the volatility (magnitude of price movements) of a stock index,
- exchange-traded funds,
- specific government bonds,
- foreign exchange rates,
- key interest rates, such as the yield of the 10-year government bullet bond or the 3-month Eurodollar rate,
- futures contracts on specific agricultural products, precious metals, energy, nationwide or regional house price indices, or weather-related quantities.

Note that an option on a futures contract is a compound derivative: a derivative on a derivative.

Typically several options on the same underlying asset are traded simultaneously. These options differ with respect to their exercise price and their time to expiration. Most options are relatively short-dated and mature within six months after they are initially introduced. Still, some longer-term options are traded, for example the so-called LEAPS (Long-term Equity AnticiPation Securities) traded on CBOE. These are options on individual stocks or stock indices with an initial time-to-maturity of up to three years.

A large number of options are traded on organized exchanges around the world, and in addition there are many, many options traded OTC, that is in over-the-counter markets. Exchange-traded options are standardized and are settled via a clearing house associated with the exchange, and the clearing house guarantees that option holders receive what they are entitled to according to the contract terms. Some semi-organized OTC markets for options on certain underlying assets exist, whereas other OTC markets are little organized. In principle, any two parties can enter an option-like contract, but most often at least one of the parties is a bank or another financial institution. The exchange-traded options are European or American options written on, or at least closely linked to, other exchange-traded assets. While OTC markets also offer similar “plain vanilla” options, numerous more exotic options are also traded.

An **exotic option** has one or more non-standard features which, for example, may refer to the underlying variable, the way of computing the option payoff, or the possible exercise dates. Some examples are:

1. *Bermudan option*: allows the owner to exercise at several pre-specified dates so in a sense it falls between a European and an American option, but is closer to an American than a European option—just as Bermuda is between Europe and America, but closest to America.
2. *Asian option*: the payoff is determined from the average price of the underlying asset in a pre-specified period.
3. *Binary option*: here the payoff is either zero or some pre-specified fixed amount, e.g., \$100, depending on whether some specific event is realized when the option expires. As an example, the event could be that S&P 500 stock index is above 2,000. Buying a binary option is very much like taking a simple bet that the event happens.

Numerous financial assets and other contracts have **embedded option features**. Many mortgage loans grant the borrower the right to prepay the loan before maturity by paying the entire outstanding debt. Such a prepayable loan is a package of a non-prepayable loan and an option. Likewise, most corporate bonds are callable in the sense that the issuing company retains the right to buy back the bonds at a pre-specified price, often the par value plus a premium of a few percent. A convertible corporate bond is a corporate bond

giving its holder the right to exchange the bond for a certain number of stocks in the company issuing the bond.

The so-called **executive stock options** are call options issued by a company to a manager of the company granting the manager the right to purchase stocks in the company. The options should supposedly give the manager an incentive to work harder and take decisions that contribute to increasing the stock price, which would benefit both the manager and the owners of the company. Executive stock options are typically long-term American options (often issued with a maturity of 5-10 years) but have an initial vesting period of, e.g., 3 years where the options cannot be exercised. The options are not transferable or tradable because if the manager would sell the options, the incentive effect would disappear. Most large companies in the U.S. and other leading economies offer their top managers lucrative packages of executive stock options in addition to a generous base salary and various benefits. [Hall and Murphy \(2002\)](#) report that 94% of the S&P 500 companies issued options to management in 1999. The grant value of options awarded (computed using the Black-Scholes formula discussed below) constituted 47% of annual top management compensation and for the industrial companies in S&P 500 the median market value of top management equity instrument holdings was \$31 million. The use of executive stock options for CEO compensation has declined since then. [Edmans, Gabaix, and Jenter \(2017\)](#) find that in 2014 the grant value of options constituted only 16% of the total compensation to CEOs in the S&P 500 companies. Option compensation programs are sometimes blamed for providing managers with incentives to take large risks, which might be destabilizing to industries and the society at large.⁴ Sometimes companies grant options to a wider range of employees than top management, in which case the options are referred to as employee stock options.

Options frequently show up in relation to productive investment projects. Such options are often referred to as **real options** or **strategic options**. An example is the right the owner of a gold mine has to extract gold from the mine. This right can be seen as an American-style call option. The exercise price consists of the cost of extracting the gold, whereas gold is the underlying asset. By exercising the option, the mine owner must pay the costs, but obtains gold that can be sold at current market prices. Similarly, oil fields can be considered as options on oil. At the point in time where geologists or geophysicists detect an underground area where an oil deposit seems to be located, it is generally uncertain how much oil the deposit has and how expensive it is to extract the oil. The first step is thus a costly exploration of the deposit and subsequently extraction might take place. The ownership of the areas is initially like a compound option: an option to explore the deposit which, if exercised, gives an option to extract the oil.

More generally, investment projects often allow some option-like flexibility such as the possibility to expand or shrink production, the possibility to temporarily suspend production, or the possibility to abandon the project. Before a project is initiated, there might be a timing option in terms of a possibility to postpone the starting date. This flexibility is important for the valuation of investment opportunities, and companies should identify and include any such option in their capital budgeting considerations.

The concept of an option is also useful in other contexts. An example from corporate finance is that the stocks of a levered company is like a call option on the assets of the company. Why? Suppose the company has a single loan of X to be paid back at time T .

⁴There is a sizeable theoretical literature studying whether option compensation of the manager is optimal seen from the perspective of the owners of the company. Some theoretical models support option compensation, others do not. See [Flor, Frimor, and Munk \(2014\)](#) and the references stated in that paper for further information.

The owners of the company can decide not to pay back the loan in which case the creditor takes over the assets of the company, and the stocks are typically worthless. If the owners do pay back the loan, they have possession of the assets of the company. If we let $V(T)$ denote the value of the assets at time T , the owners will decide to pay back the loan if and only if $V(T) > X$. In other words, owning the stocks is equivalent to having a call option on the assets with an exercise price X and maturity date T . As we shall see below, options are more valuable when the uncertainty about the value of the underlying at expiration is high. Therefore, the stock holders of the company will benefit from increasing the uncertainty about the value of the assets of the company, and they can do that by undertaking more risky projects. This is known as **asset substitution**. If the projects fail, the creditors will take part of the downside to the extent that they do not receive the full face value of the loan. If the projects succeed, the stock holders will keep any upside in asset value above the debt X . Of course, a rational creditor anticipates that the stock holders have incentives to increase risk, and may therefore introduce some covenants disciplining the manager/owners in the debt contract or increase the interest rate on the loan to compensate for the higher risk of default.

15.4 Option pricing: general properties

This section explores what we can say about option prices under relatively weak assumptions. Most importantly, we are able to derive the so-called put-call parity which gives a tight relation between the price of a call option and a put option written on the same underlying asset and having the same exercise price and expiration date. Further, we derive bounds on the price of a call option with the bounds being stated in terms of the exercise price and the price of the underlying asset. Similar price bounds are obtained for put options. As these bounds are typically quite wide, there are not very informative about the option prices, though.

The key assumption we make in this section is that prices are set to that no arbitrage opportunities exist. In addition, we abstract from any transactions costs, portfolio constraints, and other market imperfections.

15.4.1 The put-call parity

We start by stating and deriving the **put-call parity**. While the modern version of the parity is due to [Stoll \(1969\)](#), a closely related formula was published already in 1908 by the Italian mathematician Vinzenz Bronzin, cf. [Hafner and Zimmermann \(2009\)](#). Moreover, there is evidence that as early as the 17th century, some option traders understood that some relation of that form had to hold between the price of the call option, the put option, and the underlying asset, cf. the discussion in [Poitras \(2009a\)](#).

Consider a European call and a European put on the same underlying asset. Both options mature at time T and have an exercise price of X . Assume for now that the underlying asset does not pay any dividends between today (time $t \leq T$) and option maturity. Let $Z(t, T)$ denote the market-based discount factor for discounting back sure payments from time T to time t . This corresponds to the time t price of a zero-coupon bond paying one unit of account at time T for sure (and, of course, nothing at other dates). In particular, the present value of the exercise price is the product $Z(t, T)X$. We can also think of $Z(t, T)$ as what you need to deposit at time t on a riskfree bank account in order to have 1 at time T . If the effective annual interest rate over the period is r , we have

$$Z(t, T) = (1 + r)^{-(T-t)}.$$

If the interest rate r is continuously compounded, we have

$$Z(t, T) = e^{-r(T-t)}.$$

The put-call parity is stated in the following theorem:

Theorem 15.1

If no arbitrage opportunities exist, the following relation for European options on a non-dividend paying asset must hold:

$$C(t) + Z(t, T)X = P(t) + S(t). \quad (15.3)$$

Why is the put-call parity useful? Suppose that you know the price of the put option. Of course, you know X and T , and you will generally also know the market price $S(t)$ of the underlying asset as well as the appropriate discount factor $Z(t, T)$. Then the put-call parity immediately delivers the call price. In other words, it allows us to focus on determining either the fair put price or the fair call price, then the price of the other option follows from the put-call parity. The proof of the theorem goes as follows:

Proof

The left-hand side of (15.3) is the price at time t of a portfolio of the call option and a riskfree investment equal to the present value of the exercise price. The right-hand side is the price of a portfolio of the put option and the underlying asset. Table 15.2 shows that the two portfolios have the same value at the option expiration date no matter what the price of the underlying asset might be at that date. There are no other payments from the portfolios, so if they give the same payoff in any case, they must have the same price today. It is crucial for this argument that the options are written on a tradeable asset. If the options are written on a variable like an interest rate or the temperature in Miami, you cannot implement the portfolio corresponding to the right-hand side.

Suppose that the left-hand side portfolio was less expensive than the right-hand side portfolio, i.e., that $C(t) + Z(t, T)X < P(t) + S(t)$. Then it is easy to construct an arbitrage opportunity: take a long position in the left-hand side portfolio and a short position in the right-hand side portfolio. This means buying the call and the riskfree asset (make a deposit in the bank) and selling (shorting) the put and the underlying asset. You receive a net payment of $P(t) + S(t) - C(t) - Z(t, T)X$ which by assumption is positive. At the option expiry, you can exactly cover your short positions with the deposit, which has grown to X , and the payoff from the call option. Not all investors may be able to take the required short positions, but as long as some investors can, they will do so, and prices have to change until the arbitrage opportunity no longer exists.

If, on the other hand, $C(t) + Z(t, T)X > P(t) + S(t)$, just take the opposite positions: buy the put and the underlying asset, sell the call, and sell the riskfree asset (i.e., borrow the present value of X from the bank). Again this is an arbitrage since it provides a positive payoff now and all future net payments are zero.

Portfolio	Investment at t	Value at T	
		if $S(T) \leq X$	if $S(T) > X$
<i>Left-hand side portfolio</i>			
Call	$C(t)$	0	$S(T) - X$
Riskfree	$Z(t, T)X$	X	X
Total	$C(t) + Z(t, T)X$	X	$S(T)$
<i>Right-hand side portfolio</i>			
Put	$P(t)$	$X - S(T)$	0
Underlying	$S(t)$	$S(T)$	$S(T)$
Total	$P(t) + S(t)$	X	$S(T)$

Table 15.2: The put-call parity.

The table shows the investments and terminal values of the two portfolios set up in the proof of the put-call parity (15.3).

Example 15.1

Consider European options written on a stock. The exercise price is $X = 800$ and the options expire in three months from now. The stock pays no dividends in this period. Suppose the annualized three-month interest rate is 6% with continuous compounding so that the relevant discount factor is $Z(t, T) = e^{-0.06 \times 0.25} \approx 0.9851$. Assume that the stock price is $S(t) = 805$. If we know that the call option is traded at a price of $C(t) = 24.20$, it follows from the put-call parity that the price of the put option has to be

$$P(t) = 24.20 - 805 + 0.9851 \times 800 \approx 7.29.$$

If the put price is below 7.29, we are in the situation where $P(t) < C(t) - S(t) + Z(t, T)X$. Hence, we can lock in an arbitrage profit by purchasing the put, selling the call, purchasing the stock, and borrowing $Z(t, T)X$ at the riskfree rate. This provides an immediate profit of $-P(t) + C(t) - S(t) + Z(t, T)X > 0$ and, according to the proof of the put-call parity, the future net payments on this portfolio are zero no matter what.

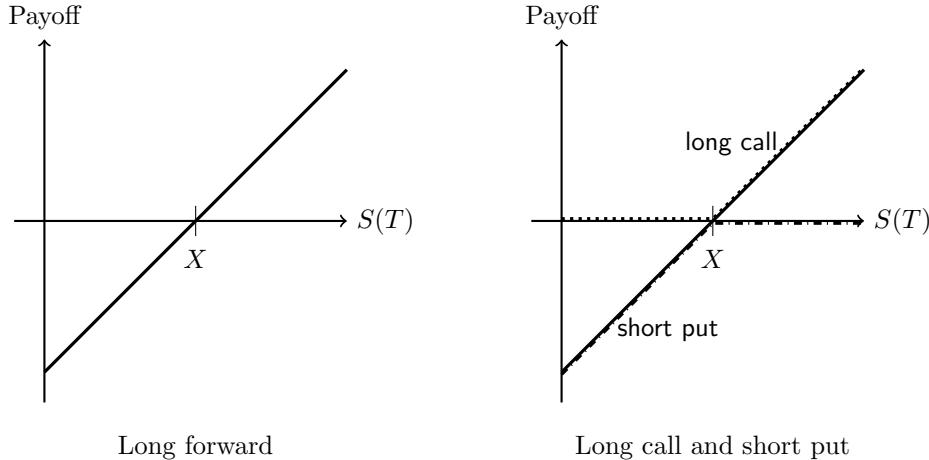
Conversely, if the put price exceeds 7.29, we sell the put, buy the call, sell the stock, and deposit $Z(t, T)X$ at the riskfree rate. This is an arbitrage.

The put-call parity can be restated as

$$C(t) - P(t) = S(t) - Z(t, T)X, \quad (15.4)$$

which emphasizes the fact that the parity pins down the difference between the call price and the put price. This relation holds because a portfolio of a long call and a short put always give a payoff equal to $S(T) - X$. Exactly the same payoff can be obtained by purchasing the underlying asset and borrowing the present value of X .

Note that it follows from (15.4) that if the exercise price X equals $S(t)/Z(t, T)$, then the call and the put have the same price. Recall from (14.4) that the ratio $S(t)/Z(t, T)$ is exactly the forward price $F(t, T)$ at time t for delivering one unit of the underlying asset at time T . So when the exercise price equals the forward price of the underlying, the call and put price are identical. An option with an exercise price equaling the forward price

**Figure 15.6: Forward vs. options.**

The left diagram shows the payoff of a forward contract as a function of the terminal value of the underlying asset. The right diagram shows the payoff of a combination of a long call option (dotted line) and a short put option (dash-dotted line) with identical exercise prices X . This combination has exactly the same payoff as the forward, no matter what the terminal value of the underlying asset is.

of the underlying is sometimes said to be *forward at the money*. Terms like *forward in the money* and *forward out of the money* are also used. If the exercise price falls below the forward price, we have $S(t) > Z(t, T)X$ so the call is more expensive than the put. Conversely if the exercise price exceeds the forward price. We can even replace $S(t)$ in the put-call parity by $Z(t, T)F(t, T)$ and rewrite it as

$$C(t) - P(t) = Z(t, T) [F(t, T) - X]. \quad (15.5)$$

A forward contract with delivery price X results in a terminal payoff of $S(T) - X$. As shown in Figure 15.6 you can obtain exactly the same payoff by taking a long position in a call and a short position in a put, both with exercise price X . The price of this option portfolio is given by the left-hand side of (15.5). Of course, when the delivery price of the forward equals the prevailing forward price $F(t, T)$, the terminal payoff is $S(T) - F(t, T)$. Therefore, having a forward with delivery price X instead of the ‘fair’ delivery price $F(t, T)$ provides an extra terminal payoff of $[S(T) - X] - [S(T) - F(t, T)] = F(t, T) - X$. The present value of this amount is exactly the right-hand side of (15.5).

The put-call parity is easy to adapt to the case where the underlying pays dividends in the life of the options as long as these dividends are known in absolute terms or are a known fraction of the price of the underlying asset at the dividend payment date. As in our analysis of forward prices in Section 14.1.4, we let $\mathbf{PV}_t(S(T))$ denote the present value at time t of getting an uncertain payment of $S(T)$ at time T and nothing at other dates. In other words, $\mathbf{PV}_t(S(T))$ is the amount that has to be invested at time t to end up with $S(T)$ at time T . Of course, if no dividends are paid before option expiration, then

$$\mathbf{PV}_t(S(T)) = S(t).$$

But if, for example, the underlying asset has a single dividend payment of D at time t^*

between today t and the option maturity date T , and D is known already at time t , then

$$\mathbf{PV}_t(S(T)) = S(t) - Z(t, t^*)D.$$

At time t you can buy the underlying asset $S(t)$ and borrow the present value of D , that is $Z(t, t^*)D$, so the net investment is $S(t) - Z(t, t^*)D$. At time t^* you repay the loan with the dividend received, and at time T you still have the underlying asset being worth $S(T)$ and you have no obligations.

The put-call parity stated above holds also with dividends if $S(t)$ is simply replaced by $\mathbf{PV}_t(S(T))$. For example, Eq. (15.3) becomes

$$C(t) + Z(t, T)X = P(t) + \mathbf{PV}_t(S(T)). \quad (15.6)$$

As pointed out by Merton (1973b), the put-call parity does not hold for American options. It can be shown that the **put-call inequalities for American options**

$$S(t) - X \leq C^a(t) - P^a(t) \leq S(t) - Z(t, T)X \quad (15.7)$$

must hold, otherwise an arbitrage can be constructed. This assumes no dividends on the underlying asset before options expiry and a non-negative discount rate so that $Z(t, T) \leq 1$. Comparing (15.7) with (15.4), we can see that the call-put price difference is generally smaller for American options than for European options.

15.4.2 Results for the price of call options

The put-call parity relates the price of the call, the put, and the underlying asset. But what can we say about the price of a call option without involving the put, again just by using the no-arbitrage principle? It turns out that we can derive an upper and a lower bound, i.e., an interval for the call price. If the underlying asset does not pay dividends in the life of the option, the price bounds are as stated in the following theorem:

Theorem 15.2

The time t price of a European call option on a non-dividend paying asset must satisfy

$$\max(S(t) - Z(t, T)X, 0) \leq C(t) \leq S(t), \quad (15.8)$$

otherwise an arbitrage opportunity exists.

We leave the case with known dividends before option expiry to the interested reader.

Proof

The call price is non-negative, $C(t) \geq 0$, since a rational option holder never has to pay anything after obtaining the option. It is also clear that the call option cannot be more expensive than the underlying asset, $C(t) \leq S(t)$: having the right to acquire the underlying at a cost of $X \geq 0$ cannot be worth more than already owning the underlying.

It remains to show that $C(t) \geq S(t) - Z(t, T)X$ or, equivalently, that

$$S(t) \leq C(t) + Z(t, T)X. \quad (15.9)$$

	Investment at t	Value at T	
		if $S(T) < X$	if $S(T) \geq X$
Buy call	$C(t)$	0	$S(T) - X$
Riskfree investment	$Z(t, T)X$	X	X
Short underlying asset	$-S(t)$	$-S(T)$	$-S(T)$
Total	$C(t) + Z(t, T)X - S(t) \leq 0$	$X - S(T) > 0$	0

Table 15.3: Call price bounds.

The table shows the investment and terminal value of an investment strategy used to show the call price bounds in Eq. (15.8).

To verify this, we show that if the opposite was true, i.e.,

$$S(t) > C(t) + Z(t, T)X, \quad (15.10)$$

we could construct an arbitrage opportunity. Simply buy the call, invest $Z(t, T)X$ in the riskfree asset (deposit the amount in the bank), and short the underlying asset. This leads to a positive net payoff today. At the option maturity date, the value of your position depends on the market price of the underlying as shown in Table 15.3.

In any case we receive X from the riskfree investment, and we have to buy back the underlying asset at a price of $S(T)$. Together, these two elements yield $X - S(T)$ at time T . If $X > S(T)$, the call expires worthless, so the total payoff is the $X - S(T)$, which is positive in this case. If $S(T) \geq X$, exercising the call option gives $S(T) - X$ and then the total payoff from the position is 0. Now it is clear that the strategy is an arbitrage: initially we receive a non-negative amount and in the future we never have to pay anything and, in addition, we have a chance of a strictly positive payoff. So if (15.10) would hold, an arbitrage would exist. Hence, we conclude that (15.9) and thus (15.8) must hold.

Example 15.2

Consider a 3-month European call option on a stock that pays no dividends before the option expires. The exercise price is $X = 800$ and the stock trades at $S(t) = 810$. Suppose the annualized 3-month interest rate is 6% with continuous compounding so that the relevant discount factor is $Z(t, T) = e^{-0.06 \times 0.25} \approx 0.9851$. From Theorem 15.2 the upper bound on the call price equals the stock price of 810. The lower price bound is

$$\max(810 - 0.9851 \times 800, 0) \approx 21.91.$$

If the call price is not in the interval between 21.91 and 810, an arbitrage can be constructed. If the call price is higher than 810, simply sell the call and buy the stock. If the call price is lower than 21.91, buy the call, short the stock, and invest the present value of X , that is $0.9851 \times 800 \approx 788.09$, at the riskfree rate.

In the example, the bounds for the call price are very, very wide. Knowing that the call

price falls between 21.91 and 810 is not really useful. Unfortunately, this is true in most situations. In fact, if the option is forward at the money or forward out of the money so that $S(t) \leq Z(t, T)X$, the lower bound equals zero, and the call price can be anywhere between zero and the price of the underlying!

To obtain more precise information on the value of the call option, we need to make additional assumptions beyond the no-arbitrage principle, namely assumptions about the probability distribution of the price of the underlying asset at the option maturity date. We consider the two leading option pricing models in the subsequent sections.

What can we say about the price of an American call option based only on the no-arbitrage pricing principle? It is clear that an American call is worth at least the same as the corresponding European call, where ‘corresponding’ means same underlying asset, same exercise price, and same maturity date. Having the European call, you have the right to exercise only at the maturity date T . In addition, the American call gives you the right to exercise at any date before the maturity date. So we have

$$C^a(t) \geq C(t).$$

By combining this with the lower bound on the European call stated in (15.8) we get

$$C^a(t) \geq C(t) \geq S(t) - Z(t, T)X.$$

The lower bound assumes that the underlying asset pays no dividends between t and T . If the discount rate is non-negative, we have $Z(t, T) \leq 1$, and thus we can conclude that

$$C^a(t) \geq S(t) - X.$$

Note that the right-hand side is exactly the payoff if you exercise the American call option right away. The inequality thus shows that the American call is worth at least as much as what you get if exercising immediately. The argument holds at any point in time before the maturity date so it is never strictly advantageous to exercise the American call prematurely. Consequently, the extra rights offered by the American call are not worth anything, and the price of the American call must equal the price of the European call. We summarize this important result in a theorem:

Theorem 15.3

If the underlying asset pays no dividends before the option maturity date and the discount rate is non-negative, then it is never advantageous to exercise an American call option before the expiration date, and

$$C^a(t) = C(t).$$

There are two reasons why early exercise is not profitable. First, by exercising early, you pay the exercise price sooner, which is costly when interest rates are positive. Second, other things equal, options are worth more with high uncertainty about the price of the underlying asset. Since the price of the underlying asset is more uncertain in the distant future than in the near future, it is typically worthwhile to wait exercising. If the underlying asset has a positive expected return, you expect the price of the underlying to increase which will lead to higher payoffs on the call option.

The no-dividend assumption is crucial for Theorem 15.3 to hold. Suppose the underlying

asset pays a dividend, say, at time t' between today t and the expiration date T . The spot price of the asset then drops at time t' by an amount corresponding to the dividend payment, which also reduces the payoff from exercising the call option. In this case, it may be optimal to exercise the American call just before the dividend payment date, but not at any other date before the final expiration date. Exercising just before the dividend payment date is only optimal if the dividend is sufficiently large to outweigh the benefits from keeping the option alive, but it is not possible to write down a simple and practically applicable rule for what exactly “sufficiently large” means. A pricing model like the binomial model considered below has to be implemented to determine under which circumstances the option should be exercised pre-maturely.

If the underlying asset pays a continuous stream of dividends as might be an appropriate approximation for a large stock index or foreign exchange, early exercise may be relevant at any date if the price of the underlying is sufficiently large. At any time t , let $\bar{S}(t)$ denote the hurdle value of the underlying so that the American call is optimally exercised at time t if and only if the price $S(t)$ of the underlying exceeds the hurdle value $\bar{S}(t)$. Looking at the hurdle values across time gives an exercise boundary $t \mapsto \bar{S}(t)$. The exercise boundary has to be determined in conjunction with the value of the American call option by using a pricing model like the binomial model.

One statement about the prices of American options that is easy to show is the following. The longer the maturity of the American call, the more rights you have, and the larger the price of the option must be. To highlight the dependence on the expiration date, let us write the price of the American call price expiring at time T as $C^a(t; T)$. Then we have

$$T' > T \quad \Rightarrow \quad C^a(t; T') > C^a(t; T).$$

This is true whether the underlying asset pays dividends or not. But if the underlying asset does not pay dividends in the life of the options, we know from the theorem above that the same relation holds for European call options:

$$T' > T \quad \Rightarrow \quad C(t; T') > C(t; T).$$

In words, the price of European call options on non-dividend paying assets is increasing in the time to expiration.

15.4.3 Results for the price of put options

For European put options the following price bounds hold:

Theorem 15.4

The time t price of a European put option on a non-dividend paying asset must satisfy

$$\max(Z(t, T)X - S(t), 0) \leq P(t) \leq Z(t, T)X, \quad (15.11)$$

otherwise an arbitrage opportunity exists.

	Investment at t	Value at T	
		if $S(T) \leq X$	if $S(T) > X$
Buy put	$P(t)$	$X - S(T)$	0
Riskfree loan	$-Z(t, T)X$	$-X$	$-X$
Buy underlying asset	$S(t)$	$S(T)$	$S(T)$
Total	$P(t) - Z(t, T)X + S(t) \leq 0$	0	$S(T) - X > 0$

Table 15.4: Put price bounds.

The table shows the investment and terminal value of an investment strategy used to show the put price bounds in Eq. (15.11).

Proof

It is clear that the put option has a non-negative value. Moreover, the value of the put cannot exceed the present value of the exercise price because the exercise price is the maximum payoff of the put. It remains to show that

$$Z(t, T)X - S(t) \leq P(t). \quad (15.12)$$

Assume the opposite, i.e., that

$$d(t, T)X > P(t) + S(t), \quad (15.13)$$

and consider the portfolio consisting of a put option, one unit of the underlying asset, and a riskfree loan of $Z(t, T)X$. This leads to a positive net payoff today (a negative net investment is required). At the option maturity date, the value of the position depends on the price of the underlying as shown in Table 15.4.

As seen in the table the terminal value is either zero or positive. Since the price is negative, we have identified an arbitrage. Since we assume that no arbitrage opportunities exist, our assumption (15.13) cannot be correct and thus Eq. (15.12) has to hold.

Example 15.3

Let us find the no-arbitrage price bounds of a 3-month European put option on a stock. The current price of the stock is $S(t) = 785$. As in Examples 15.1 and 15.2 assume an exercise price of $X = 800$ and a discount factor of $Z(t, T) = 0.9851$. The upper bound on the put price is the discounted exercise price

$$Z(t, T)X = 0.9851 \cdot 800 \approx 788.09,$$

and the lower bound is

$$\max(0.9851 \times 800 - 785, 0) \approx 3.09.$$

If the put price is not between 3.09 and 788.09, a riskfree profit can be obtained.

As the example illustrates, the bounds on the European put price are not very informative as they leave a wide interval of possible prices, exactly as we saw it for call options.

Turning to American put options, it is again clear that the American put is at least as valuable as the corresponding European put:

$$P^a(t) \geq P(t). \quad (15.14)$$

Even when the underlying asset pays no dividends in the life of the option, exercising the American put is optimal if the price of the underlying is sufficiently low. Consider the extreme case where the price of the underlying drops to zero. Then immediate exercise of the put generates a payoff of X , and the payoff can never become bigger since the price of the underlying cannot go negative. With non-negative discount rates, exercising the American put is therefore optimal. In general, at any time t , a hurdle value $\underline{S}(t)$ exists so that the American put is optimally exercised at time t if and only if the price of the underlying $S(t)$ is below the hurdle value $\underline{S}(t)$. Seen across time, the hurdle values form an exercise boundary $t \mapsto \underline{S}(t)$, but this can only be determined in conjunction with the American put price itself, e.g. by using the binomial pricing model described below.

If the underlying asset pays lump-sum dividends, the American put should never be exercised immediately before such a dividend payment. The dividend payment leads to a drop in the price of the underlying, which increases the payoff from exercising the put.

15.4.4 Factors influencing option prices

As seen above, just assuming absence of arbitrage is insufficient to pin down a unique price of an option. We have to make further assumptions about the possible evolution of the price of the underlying asset over the life of the option. In the following sections we introduce the two most popular option pricing models, namely the binomial model and the Black-Scholes model.

Before going into specific models, it is useful to think about which factors that should influence option prices and how. Of course, the price of the underlying asset is crucial. A call option is more valuable, other things equal, the higher the price of the underlying asset. Having the right to buy the underlying asset for a certain price X is clearly worth more when the underlying itself is more valuable. Conversely, the put option is less valuable, the higher the price of the underlying asset. The right to sell the underlying asset at a certain price X is worth less when the underlying is very valuable. These considerations hold for both European and American options. The price effects are indicated in the upper row of Table 15.5. Here '+' means that the option price is increasing in the factor and '-' means the option price is decreasing in the factor.

Concerning the exercise price, the value of a call is decreasing in the exercise price which the option owner has to pay in case he decides to exercise the option. The value of the put is increasing in the exercise price which the option owner receives upon exercise.

The price of an American option is increasing in the time to expiration as explained above. If the underlying asset does not pay dividends, the price of a European call equals the price of the corresponding American call and is thus also increasing in the time to expiration. With dividends, we cannot prove the same holds, but we expect it to hold except for extreme cases. This is indicated by the symbol '(+)' in the table. Similarly for European put options whether the underlying makes dividend payments or not.

The discount rate has a positive effect on call prices and a negative effect on put prices. The reason is that a higher discount rate decreases the present value of the exercise price which is paid by the call holder and received by the put holder upon exercise.

Factor	Effect on call price		Effect on put price	
	European	American	European	American
Price of underlying, $S(t)$	+	+	-	-
Exercise price, X	-	-	+	+
Time to expiration, $T - t$	(+)	+	(+)	+
Discount rate, r	+	+	-	-
Dividends	-	-	+	+
Volatility, σ	+	+	+	+

Table 15.5: Factors influencing option prices.

The symbol ‘+’ indicates that the option price is increasing in the factor whereas the symbol ‘-’ means that the option price is decreasing in the factor. The symbol ‘(+)’ means that, when the underlying asset pays dividends in the life of the money, the relation will be increasing in most cases, but maybe not in some extreme cases.

Other things equal, dividends reduce the price of the underlying as they are paid. This is beneficial for an owner of a put option, but hurts the owner of a call option.

The final factor listed in Table 15.5 is the volatility of the underlying asset, which is denoted by the symbol σ . The volatility of an asset is a measure of the magnitude of movements in the price of the asset over time and is typically estimated by the standard deviation of past price movements. If the underlying asset has a high volatility, the future values of the underlying asset are more uncertain.

A higher volatility implies that relatively extreme prices of the underlying in the future are more likely. This makes both call and put options more valuable. Why? Suppose that the exercise price X of the options equals the median price of the underlying asset at the option maturity date for any volatility we consider. Now consider what happens to the call option payoffs if we increase the volatility of the underlying. High values of the underlying asset and thus high payoffs of the call become more likely. On the other hand, low values of the underlying also become more likely, but here the call payoff is zero anyway. So in total the expected payoff of the call—and thus its price—increases with the volatility. Similarly for put options that benefit from the more likely very low values of the underlying and are not hurt by the more likely very high values of the underlying.

Example 15.4

Consider three month European options on a stock. The options have an exercise price of \$80. Table 15.6 shows two different probability distributions of the stock price at the expiration of the option. Both distributions are symmetric around the expected value of \$80. The first distribution has a high probability that the stock price ends up at or close to the expectation and is thus corresponding to a low volatility. The other distribution has higher probabilities of stock price far away from the expectation and is thus corresponding to a higher volatility than the first distribution. For the call the expected payoff using the low and high volatility, respectively, is

$$0.20 \times 10 + 0.10 \times 20 = 4,$$

$$0.15 \times 10 + 0.15 \times 20 + 0.10 \times 30 = 7.5.$$

Stock	Value or payoff at time T		Probability	
	Call	Put	Low volatility	High volatility
50	0	30	0.00	0.10
60	0	20	0.10	0.15
70	0	10	0.20	0.15
80	0	0	0.40	0.20
90	10	0	0.20	0.15
100	20	0	0.10	0.15
110	30	0	0.00	0.10

Table 15.6: Effects of volatility on option values.

The left part of the table shows the payoff of a call and a put for different terminal values of the stock. The options have an exercise price of $X = 80$. The right part of the table shows two examples of probabilities of the different payoffs, one example corresponding to a low volatility and the other corresponding to a high volatility of the underlying asset.

For the put the expected payoff using the low and high volatility, respectively, is

$$0.10 \times 20 + 0.20 \times 10 = 4,$$

$$0.10 \times 30 + 0.15 \times 20 + 0.15 \times 10 = 7.5.$$

Hence, the expected payoff of either option is higher in the high-volatility case than in the low-volatility case.

Note that by increasing the time to expiration of the option, the distribution of the price of the underlying asset at expiration generally becomes more dispersed so intuitively this leads to the same effect on option values as an increase in the volatility. However, for European put options this effect is countered by a decrease in the present value of the exercise price caused by the longer time to expiration so that the net effect is not obvious.

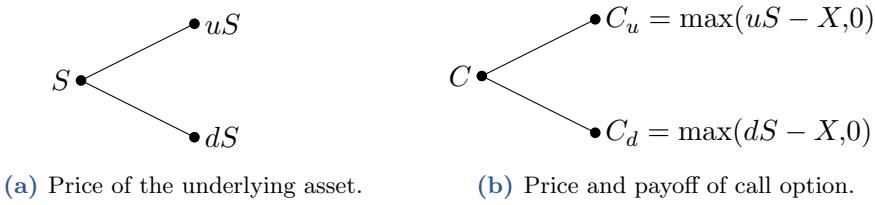
The above analysis only pins down the qualitative effect each relevant factor has on option values. Without applying a specific pricing model we cannot figure out by how much the option price increases or decreases as one of the listed factors is changed.

15.5 Option pricing: the binomial model

This section explains the binomial option pricing model introduced by [Rendleman and Bartter \(1979\)](#) and [Cox, Ross, and Rubinstein \(1979\)](#). The model is a discrete-time model in the sense that the time to expiration of the option is divided into a finite number of periods, and the model is only concerned about the price of the underlying asset and the corresponding option value at the end of each of these periods, not at any time points within periods. We start with the simple one-period version of the model to fix main ideas and subsequently extend the method to multiple periods.

15.5.1 The one-period binomial model

The one-period binomial model considers only the current date and the option expiration date. The length of the period equals the time to maturity of the option. The current price of the underlying is S and the key assumption of the model is that there are only

**Figure 15.7: The one-period binomial model**

The left picture shows how the underlying asset price can evolve over the period. The right picture shows the initial price and terminal payoff of a call option.

two possible values of the underlying at the end of the period, namely uS and dS where $u > d$. Consequently, there are only two possible payoffs of the option. The multipliers u and d are referred to as the up-factor and the down-factor, respectively. Each outcome happens with a strictly positive probability. The underlying does not pay any dividends during this period. Figure 15.7 illustrates the binomial model for a call option.

We assume investors have access to a riskfree asset, e.g. borrowing or lending from a bank. The rate of return on the riskfree asset over the period is denoted by r_f . To avoid an arbitrage, we must have

$$u > 1 + r_f > d. \quad (15.15)$$

For example, if $u > d \geq 1 + r_f$, the underlying asset gives a return at least as large as the riskfree asset with a chance of an even higher return. Then borrowing money and investing in the underlying asset would produce a sure profit. Conversely, if $1 + r_f \geq u > d$, short the underlying asset and invest the proceeds in the riskfree asset.

How can we determine the current price C of the call option? We can replicate the payoff of the option by a portfolio of the underlying asset and a riskfree investment. The portfolio is constructed at the beginning of the period and held until the end. Let H denote the number of units of the underlying asset in the portfolio and let B denote the amount invested in the riskfree asset. Since we generally assume that prices in financial markets do not generate arbitrage opportunities, the price of the call option must equal the price of setting up the replicating portfolio. The following theorem presents the results.

Theorem 15.5

In the one-period binomial option pricing model, the price of a European call option is

$$C = \frac{1}{1 + r_f} (qC_u + (1 - q)C_d), \quad (15.16)$$

where

$$q = \frac{1 + r_f - d}{u - d}. \quad (15.17)$$

The replicating portfolio consists of H units of the stock and an investment of B in the riskfree asset, where

$$H = \frac{C_u - C_d}{(u - d)S}, \quad B = \frac{uC_d - dC_u}{(u - d)(1 + r_f)}. \quad (15.18)$$

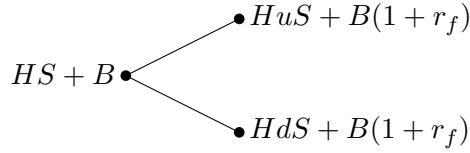


Figure 15.8: Replicating portfolio.

In the setting of the one-period binomial model, the picture shows the value of a portfolio consisting of H units of the underlying asset and a riskfree investment of B . If H and B are chosen as shown in (15.18), the portfolio is replicating the call option.

Proof

Figure 15.8 illustrates the value of the portfolio in the binomial tree. We need to choose H and B so that

$$HuS + B(1 + r_f) = C_u, \quad HdS + B(1 + r_f) = C_d. \quad (15.19)$$

This means that the portfolio has the same value at the end of the period as the call option whether the price of the underlying asset ends up at uS or dS so in this sense the portfolio replicates the option. If we subtract the second equation from the first, we get $(u - d)HS = C_u - C_d$, which is easily solved for H . Next, we can substitute this value of H into one of the equations and solve for B . The solution is given by (15.18). To rule out arbitrage, the price of the call option then has to be

$$\begin{aligned} C &= HS + B = \frac{C_u - C_d}{(u - d)S} S + \frac{uC_d - dC_u}{(u - d)(1 + r_f)} \\ &= \frac{C_u - C_d}{u - d} + \frac{uC_d - dC_u}{(u - d)(1 + r_f)} \\ &= \frac{(1 + r_f)C_u - (1 + r_f)C_d}{(u - d)(1 + r_f)} + \frac{uC_d - dC_u}{(u - d)(1 + r_f)} \\ &= \frac{1}{1 + r_f} \left(\frac{1 + r_f - d}{u - d} C_u + \frac{u - (1 + r_f)}{u - d} C_d \right) \\ &= \frac{1}{1 + r_f} (qC_u + (1 - q)C_d), \end{aligned}$$

where q is given by Eq. (15.17).

Because of the inequality (15.15), q will be a number between 0 and 1, so we can interpret it as a probability. We refer to q as the **risk-neutral up-probability** for reasons explained below. Then $qC_u + (1 - q)C_d$ can be interpreted as the expected payoff of the option, and the call price in (15.16) is thus the discounted expected payoff.

The number H of units of the underlying asset in the portfolio replicating the option is called the “hedge ratio” or the “Delta of the option.” Note that we can write H as

$$H = \frac{\Delta C}{\Delta S}, \quad (15.20)$$

where $\Delta C = C_u - C_d$ and $\Delta S = uS - dS$. The hedge ratio H measures the sensitivity of the option price or payoff at the end of the period relative to the price of the underlying asset. Observe that ΔS is positive by construction and for a call option ΔC is surely non-negative and $\Delta C \leq \Delta S$. So for a call option $H \in [0, 1]$. It can be shown that $B \in [-X(1 + r_f)^{-1}, 0]$, so in particular B is negative (or zero). The portfolio replicating a call involves a long position in the underlying asset and a loan.

Continuously compounded interest rates are frequently used in option pricing. If r denotes the annualized, continuously compounded interest rate, then the relevant discount factor over a period of length Δt is

$$1 + r_f = e^{r\Delta t}.$$

We can thus rewrite the risk-neutral probability as

$$q = \frac{e^{r\Delta t} - d}{u - d} \quad (15.21)$$

and the option pricing formula as

$$C = e^{-r\Delta t} (qC_u + (1 - q)C_d). \quad (15.22)$$

Here, in the one-period binomial model, Δt is simply the time-to-maturity of the option.

How should we value a put option in the one-period binomial model? If you go through the arguments above, notice that they do not rely on how the possible option payoffs C_u and C_d are actually computed from the price of the underlying asset. Therefore the same approach applies to the put option (or any other derivative depending only on the same underlying asset). If we write the possible payoffs of the put as

$$P_u = \max(X - uS, 0), \quad P_d = \max(X - dS, 0),$$

we get the following results.

Theorem 15.6

In the one-period binomial option pricing model, the price of a European put option is

$$P = \frac{1}{1 + r_f} (qP_u + (1 - q)P_d), \quad (15.23)$$

where q is given by Eq. (15.17). The replicating portfolio consists of H units of the stock and an investment of B in the riskfree asset, where

$$H = \frac{P_u - P_d}{(u - d)S} = \frac{\Delta P}{\Delta S}, \quad B = \frac{uP_d - dP_u}{(u - d)(1 + r_f)}. \quad (15.24)$$

In this case $P_u \leq P_d$ with $|\Delta P| \leq \Delta S$, so $H = \frac{\Delta P}{\Delta S} \in [-1, 0]$. It can be shown that $B \in [0, X(1 + r_f)^{-1}]$. The portfolio replicating a put involves a short position in the underlying asset and an investment in the riskfree asset (depositing/lending money).

The next example illustrates the numerical calculations of replicating portfolios and option values.

Example 15.5

Consider options on a non-dividend paying stock with a current price of $S = \$100$. The options mature in one year and have exercise prices of $X = \$105$. The riskfree rate is 10% per year, i.e., $r_f = 0.1$. Assume that the stock price at the end of the year is either $\$120$ or $\$90$ so that $u = 1.2$ and $d = 0.9$. See Figure 15.9 for an illustration.

First, focus on a call option. The payoff is $C_u = \max(120 - 105, 0) = 15$ in the up-state and $C_d = \max(90 - 105, 0) = 0$ in the down-state. To price the option, we compute the risk-neutral probability

$$q = \frac{1.1 - 0.9}{1.2 - 0.9} = \frac{2}{3}.$$

The call price is thus

$$C = (1.1)^{-1} \left(\frac{2}{3} \times \$15 + \left(1 - \frac{2}{3}\right) \times \$0 \right) \approx \$9.09.$$

The replicating portfolio consists of H stocks and B invested riskfree, where

$$H = \frac{\$15 - \$0}{(1.2 - 0.9) \times \$100} = 0.5, \quad B = \frac{1.2 \times \$0 - 0.9 \times \$15}{(1.2 - 0.9) \times 1.1} \approx -\$40.91,$$

i.e., a loan of $\$40.91$. The value of the portfolio in the up- and down-states are

$$\begin{aligned} 0.5 \times \$120 - 1.1 \times \$40.91 &= \$15, \\ 0.5 \times \$90 - 1.1 \times \$40.91 &= \$0, \end{aligned}$$

matching the payoff of the option. Hence, the portfolio does replicate the call option. Note that the price of the replicating portfolio is

$$HS + B = 0.5 \times \$100 - \$40.91 \approx \$9.09$$

in accordance with the price of the call option.

Next, we consider a put option on the same stock, also maturing in one year and having an exercise price of $X = \$105$. The payoff of the put is $\$0$ in the up-state and $\$15$ in the down-state so the current price is

$$P = (1.1)^{-1} \left(\frac{2}{3} \times \$0 + \left(1 - \frac{2}{3}\right) \times \$15 \right) \approx 4.55.$$

The replicating portfolio consists of

$$H = \frac{\$0 - \$15}{(1.2 - 0.9) \times \$100} = -0.5, \quad B = \frac{1.2 \times \$15 - 0.9 \times \$0}{(1.2 - 0.9) \times 1.1} \approx \$54.55,$$

i.e., a deposit of $\$54.55$. You can check that this portfolio replicates the put. The price of the portfolio is $-0.5 \times \$100 + \$54.55 = \$4.55$, identical to the put price above.

15.5.2 The risk-neutral probability

The one-period binomial option pricing formula (15.16) does not involve the true or real-world probabilities of the possible prices of the underlying asset at the end of the

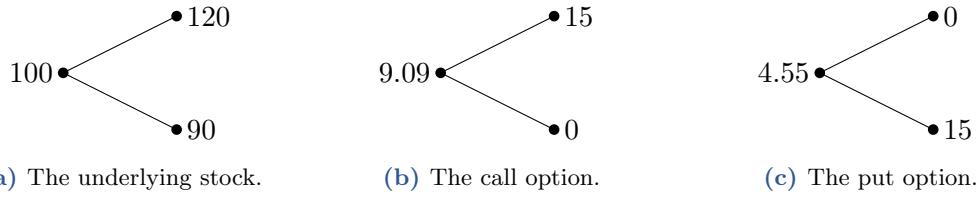


Figure 15.9: Stock and option prices and payoffs in Example 15.5.

The diagrams refer to a one-period binomial model explained in Example 15.5. The options have an exercise price of 105.

period. In fact, we have not mentioned that probability at all in the above analysis. The probability q defined in (15.17) is therefore generally different from the true probability of an up-movement, so in that sense it is an artificial probability. In fact, q is referred to as the risk-neutral up-probability, and then $1 - q$ is the risk-neutral down-probability. One reason for this terminology is that only a risk-neutral investor would value an uncertain future cash flow (in our case C_u or C_d) by taking the expected cash flow and discounting back using the riskfree rate, that is without any adjustment for risk.

Let us consider the expected return on the option. The rate of return on the call option is $(C_u/C) - 1$ in the up-state and $(C_d/C) - 1$ in the down-state. Hence, the expected return using the probabilities q and $1 - q$, respectively, is

$$\begin{aligned} E_{rn}[r_C] &= q \left(\frac{C_u}{C} - 1 \right) + (1 - q) \left(\frac{C_d}{C} - 1 \right) \\ &= \frac{qC_u + (1 - q)C_d}{C} - 1 = 1 + r_f - 1 = r_f, \end{aligned} \quad (15.25)$$

where the second-to-last equality comes from the pricing formula Eq. (15.16). Here the notation E_{rn} indicates that we calculate the expectation using the risk-neutral probabilities. So, when using the q -probability, the expected option return is indeed equal to the riskfree rate. The expected return on the underlying asset is also equal to the riskfree rate:

$$\begin{aligned} E_{rn}[r_S] &= q \left(\frac{uS}{S} - 1 \right) + (1 - q) \left(\frac{dS}{S} - 1 \right) = q(u - 1) + (1 - q)(d - 1) \\ &= qu + (1 - q)d - 1 = q(u - d) + d - 1 = 1 + r_f - d + d - 1 = r_f, \end{aligned} \quad (15.26)$$

where the second-to-last equality comes from the definition of q in Eq. (15.17).

Now think of a hypothetical risk-neutral world, i.e. a world in which all investors are risk-neutral. In such a world, prices must be set so that all assets have the same expected return. Why? Say one asset has a higher expected return than all the other assets. Since the risk-neutral investors disregard risks and rank assets solely by their expected returns, all investors would invest their money in this asset with the highest expected return and nothing in any other assets. This cannot be an equilibrium. The only equilibrium in a risk-neutral world is where all assets have identical expected returns. In particular, when a riskfree asset exists, the expected returns on all assets must equal the riskfree return. The binomial option pricing formula (15.16) is thus similar to how assets would be priced in a risk-neutral world and this motivates the name ‘risk-neutral probability’ for q . It is important to note, however, that when deriving the binomial option pricing formulas, we did not assume that investors are risk neutral. We were only assuming that the price of

the underlying asset could be described by the binomial tree, that investors can implement the replicating portfolio, and that prices respect the no-arbitrage principle.

It might seem strange why the option price apparently is independent of the real-world probabilities of the up- and down-movements of the underlying asset price. Note, however, that the pricing formulas really price the option relative to the current price of the underlying asset. For example, if we substitute the expressions for the call payoffs C_u and C_d into Eq. (15.16), we get

$$C = \frac{1}{1+r_f} (q \max(uS - X, 0) + (1-q) \max(dS - X, 0)),$$

which clearly involves S , the current price of the underlying asset. Of course, the possible future returns on the underlying and their real-world probabilities are likely to impact how the market sets S . The uncertainty is thus already accounted for through S and should not be accounted for again when determining the option price relative to S .

Most risky assets have a real-world expected return exceeding the riskfree return. This should normally be true for the stock market index and, following the CAPM, for all assets with a positive market-beta. Suppose this is the case for the underlying asset of an option. Then the real-world up-probability p has to exceed the risk-neutral up-probability q . Of course, this implies that the real-world down-probability $1-p$ must be lower than the risk-neutral down-probability $1-q$. Let use the term ‘good state’ to refer to a state in which the underlying asset has a relatively high value. Likewise, a ‘bad state’ means the underlying asset has a low value. In a one-period binomial model, the upper node at time 1 is a good state and the lower node is a bad state. The risk-neutral probabilities thus weigh good states less and bad states more than the real-world probabilities. The call option provides a positive payoff in good states and zero payoff in bad states. When using the real-world probabilities, the expected return on the call option will therefore exceed the riskfree return. In contrast, a put option delivers a positive payoff in bad states and zero payoff in good states. Therefore, the expected real-world return on a put option is below the riskfree return. We discuss option returns further in Section 15.7.

15.5.3 The two-period binomial model

The one-period binomial model is very simplistic by assuming only two possible values of the underlying asset at the maturity date of the option. However, we can extend the general approach to multi-period trees. Let us start with a two-period binomial tree. Here the time until expiration of the option is split into two periods of equal length. Over each period the price of the underlying is multiplied either by u or by d . We shall assume that the same values of u and d apply throughout the tree, and that the one-period riskfree rate r_f is also the same in all periods. Then also the risk-neutral up-probability q in (15.17) is constant throughout the tree. See Figure 15.10 for an illustration. Note that our assumptions lead to a recombining tree in the sense that the price of the underlying ends up the same whether it first moves up and then down or the other way around. Now there are three end-nodes in the tree representing three possible prices of the underlying asset at the option maturity date. The three possible payoffs of the option are easily determined from the prices of the underlying in the end-nodes. For a call option they are

$$C_{uu} = \max(u^2 S - X, 0), \quad C_{ud} = C_{du} = \max(uds - X, 0), \quad C_{dd} = \max(d^2 S - X, 0).$$

To price the option in the two-period tree we work backwards period by period using

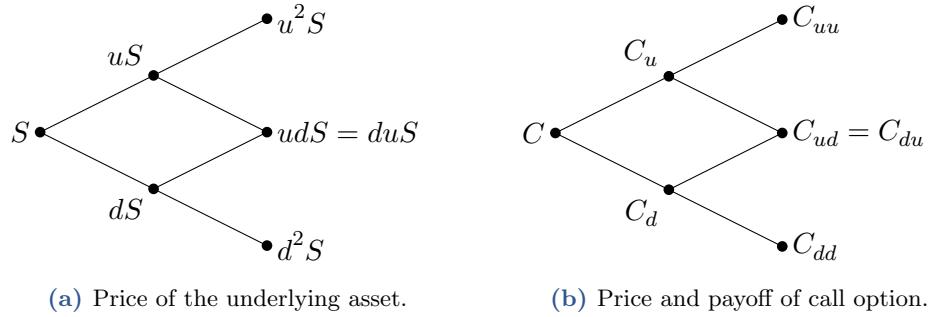


Figure 15.10: The two-period binomial model.

The diagrams show how the price of the underlying asset (left diagram) and the call option (right diagram) evolve over time in the two-period binomial model.

the technique developed for the one-period model above. The option value at the end of the first period is therefore

$$C_u = \frac{1}{1+r_f} (qC_{uu} + (1-q)C_{ud}), \quad (15.27)$$

$$C_d = \frac{1}{1+r_f} (qC_{ud} + (1-q)C_{dd}), \quad (15.28)$$

in the up- and the down-state, respectively. Subsequently we can determine the initial value of the option as

$$\begin{aligned} C &= \frac{1}{1+r_f} (qC_u + (1-q)C_d) \\ &= \frac{1}{(1+r_f)^2} (q^2C_{uu} + 2q(1-q)C_{ud} + (1-q)^2C_{dd}). \end{aligned} \quad (15.29)$$

We can interpret q^2 as the risk-neutral probability of two up-moves in a row and thus receiving an option payoff equal to C_{uu} . Likewise, $(1-q)^2$ is the risk-neutral probability of two down-moves in a row and thus an option payoff of C_{dd} . The risk-neutral probability of going up and then down is $q(1-q)$ and the risk-neutral probability of going down and then up is $(1-q)q$, which is the same. In both cases the option provides the payoff $C_{ud} = C_{du}$. The total risk-neutral probability of obtaining that payoff is thus $2q(1-q)$. Consequently, we can interpret the formula (15.29) for the call price as the risk-neutral expectation of the option payoff discounted back over two periods using the riskfree rate.

The option valuation formula (15.29) is again based on the idea of replicating portfolios and no-arbitrage pricing. In a two-period tree the replicating portfolio is not the same in all nodes. Based on our one-period analysis, we conclude that if we are in the up-node after the first period, the replicating portfolio over the following period is given by

$$H_u = \frac{C_{uu} - C_{ud}}{(u-d)uS}, \quad B_u = \frac{uC_{ud} - dC_{uu}}{(u-d)(1+r_f)}. \quad (15.30)$$

In the down-node the replicating portfolio is

$$H_d = \frac{C_{du} - C_{dd}}{(u-d)dS}, \quad B_d = \frac{uC_{dd} - dC_{du}}{(u-d)(1+r_f)}. \quad (15.31)$$

Over the first period the replicating portfolio is

$$H = \frac{C_u - C_d}{(u - d)S}, \quad B = \frac{uC_d - dC_u}{(u - d)(1 + r_f)}. \quad (15.32)$$

Again the formulas for the call option also apply to the put option, only the computation of the terminal payoffs differ.

Example 15.6

Consider the two-period tree for the price of a stock in Figure 15.11. The up- and down-factors are $u = 1.25$ and $d = 0.9$. Assume that the riskfree rate is $r_f = 0.01$ per period. The risk-neutral up-probability is therefore $q = (1.01 - 0.9)/(1.25 - 0.9) \approx 0.3143$.

We seek to value a European put option expiring at the end of the second period. The exercise price is $X = 240$. The possible payoffs are therefore 0, 15, and 78 depending on the stock price when the put expires. The current put price can be computed directly as

$$\begin{aligned} P &= (1.01)^{-2} \left((0.3143)^2 \times 0 + 2 \times 0.3143 \times (1 - 0.3143) \times 15 + (1 - 0.3143)^2 \times 78 \right) \\ &\approx 42.30. \end{aligned}$$

Alternatively, we can work backwards through the tree. In the up-state after the first period, the replicating portfolio is given by

$$H_u = \frac{0 - 15}{312.50 - 225} \approx -0.1714, \quad B_u = \frac{1.25 \times 15 - 0.9 \times 0}{(1.25 - 0.9) \times 1.01} \approx 53.04,$$

so that the put value is

$$P_u = H_u \times uS + B_u = -0.1714 \times 250 + 53.04 \approx 10.18.$$

In the down-state after the first period, the replicating portfolio is given by

$$H_d = \frac{15 - 78}{225 - 162} = -1, \quad B_d = \frac{1.25 \times 78 - 0.9 \times 15}{(1.25 - 0.9) \times 1.01} \approx 237.62,$$

so that the put value is

$$P_d = H_d \times dS + B_d = -1 \times 180 + 237.62 \approx 57.62.$$

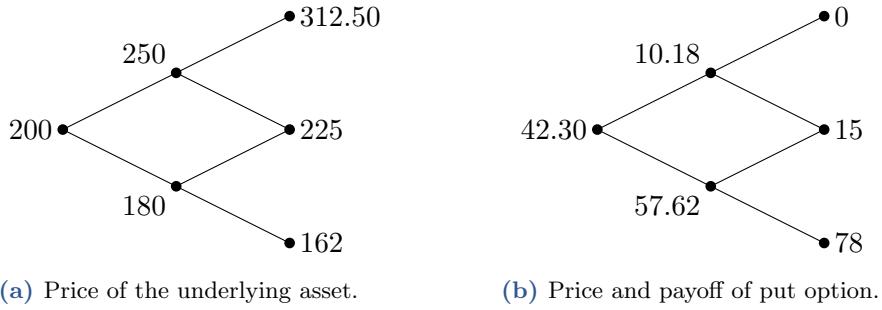
Notice that in this node the replicating portfolio consists of shorting the stock and investing the present value of the exercise price, $(1.01)^{-1} \times 240 \approx 237.62$, in the riskfree asset. This is because you know the put is ending up in the money whether the stock price goes up or down over the subsequent period. Hence, you know the payoff will be $X - S_T$, where S_T is the stock price at maturity, and you replicate this payoff in the way just described.

Finally, in the first period the replicating portfolio is given by

$$H = \frac{10.18 - 57.62}{250 - 180} \approx -0.6777, \quad B = \frac{1.25 \times 57.62 - 0.9 \times 10.18}{(1.25 - 0.9) \times 1.01} \approx 177.84,$$

from which we get a put value of

$$P = H \times S + B = -0.6777 \times 200 + 177.84 \approx 42.30.$$

**Figure 15.11: The two-period binomial model of Example 15.6.**

The left diagram shows the underlying asset price and the right diagram the value of a European put option with an exercise price of 240.

15.5.4 The multi-period binomial model

Also the two-period binomial model is simplistic, but we can add further periods. An N -period recombining binomial tree has $N + 1$ end nodes representing different prices of the underlying. Assuming the same up-factor u , down-factor d , and riskfree rate r_f in all periods, the risk-neutral up-probability in any period is given by q . The option payoff is determined by the number of up- and down-movements in the underlying asset. With j up-moves and $N - j$ down-moves the price of the underlying ends up at $u^j d^{N-j} S$ and the payoff is then $\max(0, u^j d^{N-j} S - X)$ for the call and $\max(0, X - u^j d^{N-j} S)$ for the put. The probability that the price of the underlying moves up in j and down in $N - j$ periods is

$$\text{Prob}(j \text{ ups}) = \frac{N!}{j!(N-j)!} q^j (1-q)^{N-j},$$

where $\frac{N!}{j!(N-j)!}$ is the number of ways to pick j out of N periods. Recall that for any positive integer n , we define “ n factorial” as $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. Outcomes with almost all moves being in the same direction are less likely than outcomes with a similar number of ups and downs, so the distribution of the terminal price of the underlying has kind of a bell-shape. However, the distribution has a fatter right tail than left tail and, when increasing the number of periods N , the distribution of the terminal price of the underlying asset will, in fact, approach a lognormal distribution.

The price of a European option still equals the discounted risk-neutral expectation of the option payoff, which leads to the following theorem:

Theorem 15.7

In the N -period binomial model, the prices of a European call and put option are

$$C = \frac{1}{(1+r_f)^N} \sum_{j=0}^N \frac{N!}{j!(N-j)!} q^j (1-q)^{N-j} \max(0, u^j d^{N-j} S - X), \quad (15.33)$$

$$P = \frac{1}{(1+r_f)^N} \sum_{j=0}^N \frac{N!}{j!(N-j)!} q^j (1-q)^{N-j} \max(0, X - u^j d^{N-j} S), \quad (15.34)$$

where q is given by Eq. (15.17).

Although the formulas look somewhat complicated, they are easy to program on a computer. For any other derivative depending only on the same underlying asset, the same formula applies if the payoff function is changed accordingly. The risk-neutral valuation approach and thus the option pricing formulas above are really based on the replication argument, and a replicating portfolio for each one-period sub-tree could be computed in each node by working backwards through the tree. As before, the portfolio varies with time and the price of the underlying asset, so it has to be rebalanced. If the issuer of the option wants to hedge it, a dynamic hedging strategy in the underlying asset and the riskfree asset is therefore necessary.

15.5.5 Handling American options

So far we have focused on the pricing of European options, but the binomial model can also be used for pricing American options as follows:

1. Go backwards through the tree, period-by-period.
2. As soon as you have computed the option value in a given node, compare it with the payoff from immediate exercise.
3. If the exercise payoff is higher, then exercise in that node is optimal. Overwrite the option value by the exercise value before you move on to earlier periods.

When you reach the initial node (corresponding to the present point in time), you will know both (a) the optimal exercise strategy, i.e., in what future situations you should exercise the option, and (b) the current value of the option when you take into account the possibility of early exercise. While we cannot write up a closed-form expression for the American option price, the procedure is relatively easy to program on a computer.

Example 15.7

We take the same two-period binomial model as in Example 15.6, but now we consider an American put option instead of a European put. The exercise price is still $X = 240$ and the payoff of the American put (if still un-exercised) at maturity is the same as for the European put. We proceed backwards through the tree and derive the put values shown in Panel (b) of Figure 15.12.

Start at the up-node after the first period. If we do not exercise the American put here, we have no chance to exercise it before maturity, so this decision gives us a value equal to the price $P_u = 10.18$ of the European put in this node. If we would exercise the American put in this node, we would get an immediate payoff of $X - uS = 240 - 250 = -10$, so of course we keep the put alive. Consequently, the value of the American put in this node is $P_u^a = P_u = 10.18$. The American put can be replicated over the subsequent period with the same portfolio that replicates the European put, i.e., $H_u^a \approx -0.1714$ and $B_u^a \approx 53.04$.

Now look at the down-node after the first period. If we do not exercise the American put here, it has a value equal to that of the European put in this node, i.e., $P_d = 57.62$. If we exercise the American put in the down-node, we receive $240 - 180 = 60$, which is higher. Therefore, the optimal decision is to exercise the American put if the stock price goes down during the first period. The early-exercise decision is marked by the red node in Figure 15.12. The value of the American put in the down-node is $P_d^a = 60$. It does not make sense to talk about a replicating portfolio over the subsequent period, since the option no longer exists.

Finally, consider the initial node. If we do not exercise the put right away, it will have a

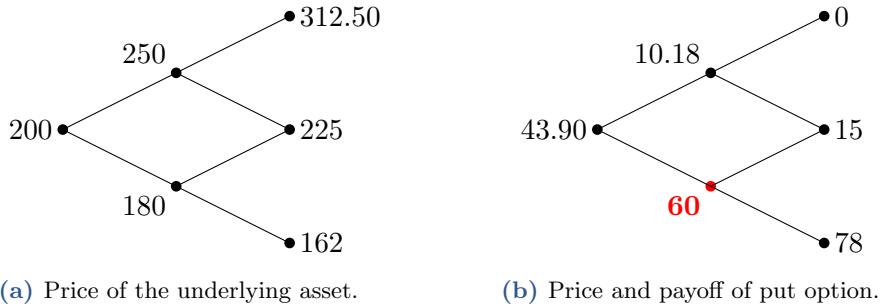


Figure 15.12: American option in a two-period binomial model.

The figure refers to Example 15.7. In the setting of a two-period binomial model, the left diagram shows the possible evolution in the price of the underlying stock. The right diagram shows the value of an American put option with an exercise price of 240. The red-colored node and value of 60 indicate that the put option is optimally exercised in this node before the option matures.

value of either $P_u^a = 10.18$ or $P_d^a = 60$ at the end of the first period. The present value is

$$(1 + r_f)^{-1} (qP_u^a + (1 - q)P_d^a) = (1.01)^{-1} (0.3143 \times 10.18 + (1 - 0.3143) \times 60) \approx 43.90.$$

We can replicate the put values at the end of the first period with the portfolio

$$H^a = \frac{10.18 - 60}{250 - 180} \approx -0.7117, \quad B^a = \frac{1.25 \times 60 - 0.9 \times 10.18}{(1.25 - 0.9) \times 1.01} \approx 186.24,$$

which costs $-0.7117 \times 200 + 186.24 \approx 43.90$ in accordance with the option value. We still have to check whether it is optimal to exercise the option right away. If we do that, we receive a payoff equal to $240 - 200 = 40$, which is less than the un-exercised value of 43.90. Hence, the optimal decision is not to exercise the put in the initial node, and the initial value of the American put option is

$$P^a = 43.90.$$

Recall from Example 15.6 that the value of the corresponding European put is $P = 42.30$. The difference of $P^a - P = 1.60$ is the so-called *early exercise premium*, which captures the value of the right to exercise before maturity. This right is valuable in the example because we will make use of it should the stock price fall over the first period.

15.5.6 Other issues

If you want to apply the binomial model to price a given option, you need to specify the relevant inputs. Of course, the exercise price and time to maturity of the option, as well as the initial price of the underlying asset, are easy to find. The relevant interest rate can be backed out from prices of traded bonds. It remains to specify the number of periods N and the up- and down-factors u and d . A relatively large number of periods is necessary to obtain a realistic distribution of the future price of the underlying asset. Often somewhere between 20 and 50 periods are used.

The values of u and d are more tricky. Higher values of u and lower values of d lead to

more extreme prices of the underlying, so the values of u and d are closely related to the volatility of the underlying asset. The volatility σ is typically defined as the annualized standard deviation of the log-returns on the underlying asset, and therefore $\sigma\sqrt{\Delta t}$ is the standard deviation of the log-return over a period of length Δt , cf. Section 3.7.2. This can be estimated from the observed price movements of the underlying asset in the recent years. Then u and d in a binomial tree with period length Δt are often chosen as

$$u = e^{\sigma\sqrt{\Delta t}}, \quad d = e^{-\sigma\sqrt{\Delta t}}. \quad (15.35)$$

In particular, $d = 1/u$ with this choice, which implies that after an up-move and a down-move the underlying is back at the original price. Other ways of fixing u and d have been suggested, however.

In the above description of the binomial model we have assumed that the underlying asset does not make any dividend payments during the life of the option, but the binomial model can be extended to dividends. The main idea is still to set up a replicating portfolio, but now we have to take into account that by holding the portfolio involving the underlying asset we also get a dividend payment, which we do not get by holding the option.

The simplest case is when the underlying asset can be assumed to pay dividends continuously through time with the dividend payments being proportional to the value of the asset, i.e., with a constant dividend yield δ . This is a reasonable assumption when the underlying asset is a foreign currency where the foreign riskfree rate r_{for} takes the role of δ because you can deposit the foreign currency and receive interests. It may also be a reasonable approximation for option on a broad stock index. As the dividend payment dates of the different companies are spread over the year, the total dividends to a basket of many stocks is more or less a continuous stream. And for relatively short periods of time, it may be reasonable to assume the dividend yield is a constant δ . With a constant dividend yield δ , the binomial analysis and results go through as long as the risk-neutral probability q in (15.21) is replaced by

$$q = \frac{e^{(r-\delta)\Delta t} - d}{u - d} \quad (15.36)$$

and then the option price can be computed using (15.22) as before.

Example 15.8

A company based in the Euro-zone needs around 1,000,000 USD in three months to pay a US-based supplier and worries that EUR/USD exchange rate increases. It thus considers buying a Euro-denominated 3-month call option on USD. The current spot exchange rate is 0.7467 EUR per USD, and the exercise price of the option is 0.75 EUR per USD. The annualized 3-month interest rate is 0.17% in the Euro-zone and 0.24% in the US, and we assume both are computed using continuous compounding.

Let us compute the price of the European call option using a 3-period binomial tree with $u = 1.05$ and $d = 0.95$, meaning that the exchange rate either increases or decreases by 5% each month. The binomial tree for the exchange rate is shown in the left panel of Figure 15.13 with all numbers rounded to four digits. The risk-neutral up-probability is

$$q = \frac{e^{(0.0017-0.0024)\times 3/12} - 0.95}{1.05 - 0.95} \approx 0.49825.$$

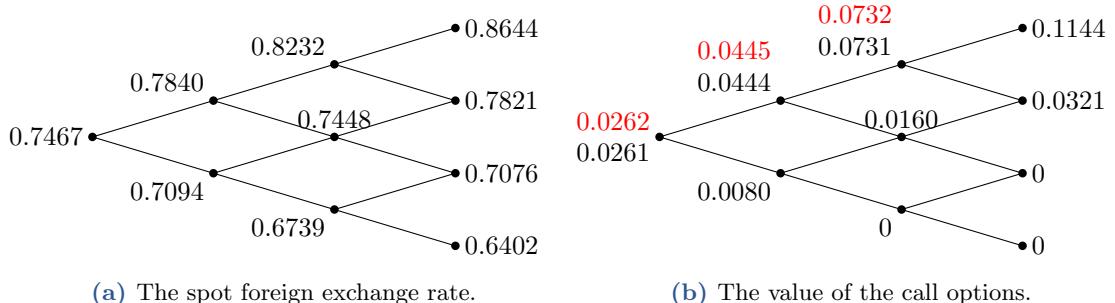


Figure 15.13: Currency options in a binomial model.

The figure refers to Example 15.8. The left diagram shows the EUR/USD spot exchange rate. The right diagram shows the values of a European call option in black, as well as the values of an American call option in red in the nodes where the American option is worth more than the European option.

The possible terminal payoffs of the European call are depicted at the end of the binomial tree in the right panel of Figure 15.13. Given these payoffs we can value the European call directly as

$$\begin{aligned} C &= e^{-0.0017 \times 3/13} \left((0.49825)^3 \times 0.1144 + 3 \times (0.49825)^2 \times (1 - 0.49825) \times 0.0321 \right) \\ &\approx 0.026125. \end{aligned}$$

This is the price in EUR per USD so a call on 1,000,000 USD costs 26,125 EUR. Alternatively, we can price the option by going backwards through the tree period by period.

To price the corresponding American call option we go backwards period by period and in every node check whether early exercise is optimal. In the right panel of Figure 15.13, the numbers written in red show the value of the American call whenever it is worth more than the European call. Early exercise is optimal if the exchange rate goes up in each of the first two months, although the exercise profit of $0.823237 - 0.75 = 0.073237$ is only slightly larger than the value of keeping the option alive, which is 0.073082. Since this node is only reached with a probability of q^2 , the effect on the initial option price is even smaller. The initial price of the American call is 0.026163 compared to the European call price of 0.026125. But it does give an example where early exercise of an American call option is optimal, which may happen when the underlying asset pays dividends.

The binomial model can also be accommodated to other types of dividend payments as shown for example by Hull (2021).

15.6 Option pricing: the Black-Scholes model

The Black-Scholes option pricing model was introduced by Black and Scholes (1973) and Merton (1973c). You can think of the model as the limit of the binomial model when the number of periods goes to infinity so that the price of the underlying asset changes continuously in time. This may sound complicated, but the resulting option pricing formula is relatively simple. We present and discuss the pricing formula first and then dig into the precise assumptions of the model.

15.6.1 The option pricing formula

First, let us fix some notation. Let T denote the time to maturity and X the exercise price of the option. Let r denote the continuously compounded, annualized riskfree rate so that e^{-rT} is the value today of getting a payment of one for sure at the option maturity date. The underlying asset has a current price of S and a volatility of σ ; we discuss what the ‘volatility’ really means in the next subsection. For now, we assume that the underlying asset pays no dividends before the option matures. Now we can state the Black-Scholes option pricing formula.

Theorem 15.8

In the Black-Scholes option pricing model, the prices of a European call and put option on an asset not paying dividends are

$$C = SN(d_1) - Xe^{-rT}N(d_2), \quad (15.37)$$

$$P = Xe^{-rT}N(-d_2) - SN(-d_1), \quad (15.38)$$

where N denotes the cumulative distribution function for a standard normal distribution and

$$d_1 = \frac{\ln(S/X) + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \quad d_2 = d_1 - \sigma\sqrt{T}. \quad (15.39)$$

In the notation used in this chapter up to this point, T has represented the maturity date of the option, whereas in the above formula it represents the time to maturity. The two representations are equivalent if the point in time at which we are valuing the option is time $t = 0$. Alternatively, if we think of T as being a fixed maturity date, we can write the call price at a general time $t < T$ as

$$C_t = S_t N(d_{1t}) - X e^{-r(T-t)} N(d_{2t}) \quad (15.40)$$

with

$$d_{1t} = \frac{\ln(S_t/X) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}, \quad d_{2t} = d_{1t} - \sigma\sqrt{T-t}. \quad (15.41)$$

Example 15.9

The stocks of a given company are traded at \$720 per share. What is the price of a two-month European call option on the stock if the exercise price is \$725? Suppose the volatility is 25% per year so that $\sigma = 0.25$. The riskfree rate is 5% per year using continuous compounding. The company does not pay dividends during the next two months.

To apply the Black-Scholes formula, we first determine d_1 and d_2 :

$$d_1 = \frac{\ln(720/725) + (0.05 + 0.5 \times (0.25)^2) \times 2/12}{0.25 \times \sqrt{2/12}} \approx 0.0649,$$

$$d_2 = 0.0649 - 0.25 \times \sqrt{2/12} \approx -0.0372.$$

Next, we determine the probabilities $N(d_1)$ and $N(d_2)$ by using a spreadsheet, advanced calculator, or by applying an appropriate table of the standard normal distribution:

$$N(d_1) \approx 0.52587, \quad N(d_2) \approx 0.48516.$$

By substitution into the Black-Scholes' formula (15.37) we find that

$$C = \$720 \times 0.52587 - \$725 \times e^{-0.05 \times 2/12} \times 0.48516 \approx \$29.80$$

is the price of the option.

The Black-Scholes formula is relatively easy to apply. The hardest part is to compute the two probabilities $N(d_1)$ and $N(d_2)$ from the normal distribution, but this distribution function is standard in relevant computer software and even many calculators. In contrast, an application of the multi-period binomial option pricing formula (15.33) requires more calculations. After learning the tractable Black-Scholes formula in the early 1970's, investors got much more confident in how to value options, which was a crucial element in the initiation of modern option markets in 1973 and their subsequent developments.

The Black-Scholes option pricing formulas can be extended to the case in which the underlying asset pays dividends before the expiration of the option. Just replace S by $S - PV(\text{dividends})$ both directly in (15.37) and in the calculation of d_1 and d_2 . The simplest case is when the underlying pays a continuous dividend with a constant dividend yield δ . As discussed in Section 14.1.4 this is relevant when the underlying asset is a foreign currency or a broad stock index. Then $S - PV(\text{dividends}) = e^{-\delta T} S$ and the adjusted Black-Scholes option pricing formula is

$$C = Se^{-\delta T} N(d_1) - Xe^{-rT} N(d_2), \quad (15.42)$$

where

$$d_1 = \frac{\ln(S/X) + (r - \delta + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \quad d_2 = d_1 - \sigma\sqrt{T}. \quad (15.43)$$

When applied to foreign exchange, where δ is the foreign riskfree interest rate, the option pricing formula is sometimes referred to as the Garman-Kohlhagen formula as it was first published by [Garman and Kohlhagen \(1983\)](#).

Note that the Black-Scholes formula is for a European option. For an American call option on a non-dividend paying asset, we can still use the Black-Scholes formula because in this case we know that the option should never be exercised before maturity and, consequently, the American call price coincides with the European call price. But if the underlying asset pays dividends, we cannot use the Black-Scholes formula for American call options. And whether or not the underlying pays dividends, we cannot use the Black-Scholes formula for American put options. To price American options we generally have to resort to the binomial models.

15.6.2 Assumptions and derivation of the pricing formula

As the binomial model, the Black-Scholes model is based on a probabilistic assumption about how the price of the underlying asset may evolve over the life of the option. But in contrast to the discrete-time binomial model, the Black-Scholes model is a continuous-time model in the sense that it considers the price of the underlying and of the option

at *any* point in time until the maturity date. In addition to the standard no-arbitrage assumption, the key assumptions are that the riskfree rate is constant and that the return on the underlying asset over any tiny time period, say from time t to time $t + dt$, is given by

$$\frac{S(t + dt) - S(t)}{S(t)} = \mu dt + \sigma \tilde{\varepsilon} \sqrt{dt}, \quad (15.44)$$

where $\tilde{\varepsilon}$ is a random variable following a standard $N(0,1)$ normal distribution. Moreover, μ and σ are constants where σ has to be positive and μ is typically also positive. The expectation and variance of this “instantaneous” return are

$$\begin{aligned} E\left[\frac{S(t + dt) - S(t)}{S(t)}\right] &= E[\mu dt] + E[\sigma \tilde{\varepsilon} \sqrt{dt}] = \mu dt + \sigma \sqrt{dt} E[\tilde{\varepsilon}] = \mu dt, \\ \text{Var}\left[\frac{S(t + dt) - S(t)}{S(t)}\right] &= \text{Var}[\mu dt] + \text{Var}[\sigma \tilde{\varepsilon} \sqrt{dt}] = \sigma^2 dt \text{Var}[\tilde{\varepsilon}] = \sigma^2 dt, \end{aligned}$$

so μ represents an annualized expected return and σ is the annualized standard deviation of the instantaneous return. The parameter σ is generally referred to as the volatility of the underlying asset.

It can be shown that the assumption (15.44) implies that the terminal value $S(T)$ of the underlying asset’s price is lognormally distributed. More precisely, given the price $S(t)$ at time t we have

$$\ln S(T) \sim N\left(\ln S(t) + \left[\mu - \frac{1}{2}\sigma^2\right](T-t), \sigma^2[T-t]\right) \quad (15.45)$$

or, equivalently,

$$\ln\left(\frac{S(T)}{S(t)}\right) \sim N\left(\left[\mu - \frac{1}{2}\sigma^2\right](T-t), \sigma^2[T-t]\right) \quad (15.46)$$

from which it is clear that $\sigma^2[T-t]$ is the variance of the log-return over the period. Note that this is similar to the assumptions on risky asset made in Merton’s basic model for long-term investments in Section 8.1. It can also be shown that the expectation at time t of the underlying asset’s price at time T is $E_t[S(T)] = S(t)e^{\mu(T-t)}$ and that the expected rate of return on the underlying asset is $e^{\mu(T-t)} - 1 \approx \mu(T-t)$. A lognormal distribution is not a bad approximation to the distribution of stock prices or returns observed empirically, but it may be a terrible description of the price distribution of other assets and then the Black-Scholes formula should not be applied to options on such assets.

As in the binomial model, the option price in the Black-Scholes model is derived by forming a portfolio of the underlying asset and the riskfree asset that replicates the option payoff at expiration. In the Black-Scholes model this portfolio has to be rebalanced continuously through time. It can be shown that the initial replicating portfolio of the call option consists of H units of the underlying asset and a riskfree investment of B where

$$H = \frac{\partial C}{\partial S} = N(d_1), \quad B = -Xe^{-rT}N(d_2). \quad (15.47)$$

It is easy to see that the price of this portfolio coincides with the call price given by the Black-Scholes formula (15.37). Note that $H \in [0, 1]$ and $B \in [-Xe^{-rT}, 0]$ similarly to the binomial model. The replicating portfolio is therefore a levered position in the underlying asset. As the price of underlying asset S and the time-to-expiration T change, so does the replicating portfolio. For the put option the hedge ratio $H = -N(-d_1)$ is negative, and

the riskfree investment $B = Xe^{-rT}N(-d_2)$ is positive.

For the binomial model we saw that the replication argument leads to the principle of risk-neutral valuation by which the option is priced as the discounted risk-neutral expectation of the option payoff. This is also true in the Black-Scholes framework. Therefore, the price of the European call can be computed as

$$C = e^{-rT} \mathbb{E}_{\text{rn}} [\max(S(T) - X, 0)], \quad (15.48)$$

where the subscript ‘rn’ indicates that the expectation is computed using the risk-neutral distribution of the terminal price $S(T)$ instead of the real-world distribution. In a risk-neutral world investors would choose investments only according to their expected returns so in equilibrium all assets would have to offer the same expected return and thus an expected return equal to the riskfree return. Therefore, in a risk-neutral world we replace the expected return parameter μ in the distribution (15.45) by the riskfree rate r . The expectation appearing in Eq. (15.48) can then be computed in closed form as shown in Theorem A.5 in Appendix A. By applying this theorem and subsequently discounting back, Eq. (15.48) does in fact lead to the Black-Scholes option pricing formula (15.37).

Given the Black-Scholes formula (15.37) for the call option, the put price in (15.38) follows by applying the put-call parity from Theorem 15.1 and the following manipulations:

$$\begin{aligned} P &= C + e^{-rT}X - S \\ &= SN(d_1) - Xe^{-rT}N(d_2) + e^{-rT}X - S \\ &= Xe^{-rT}(1 - N(d_2)) - S(1 - N(d_1)) \\ &= Xe^{-rT}N(-d_2) - SN(-d_1). \end{aligned}$$

Here the last equality is due to the relation $N(-z) = 1 - N(z)$ that holds because the standard normal distribution is symmetric around zero.

The Black-Scholes option pricing formula does not involve the parameter μ describing the expected return on the underlying asset. This is equivalent to our earlier observation that the binomial option pricing formula does not involve the real-world probabilities of the possible movements in the underlying asset price. While it may seem surprising that the expected return on the underlying is irrelevant for the option value, note that the Black-Scholes option pricing formula involves the current price of the underlying which presumably is set so that the expected return provides the fair compensation for risk.

15.6.3 Relation to the binomial model

The Black-Scholes model published in 1973 precedes the binomial model that was not published until 1979. In fact, the binomial model was motivated as a simplified and pedagogical approach to option pricing.

While the binomial model and the Black-Scholes model appear very different, they are closely related. The Black-Scholes model can be seen as the limit of the binomial model if the number of periods in the tree goes to infinity, at least if the u and d factors in the binomial tree are chosen according to (15.35). Figure 15.14 illustrates the convergence for a specific European call option. The Black-Scholes option price is \$8.57 in this case. With 20 or more periods the price generated by the binomial model deviates less than 1% from the Black-Scholes price. With more than 100 periods, the binomial price stays between \$8.55 and \$8.59. The nice convergence indicates that if you want to price an American option and you believe in the assumptions behind the Black-Scholes model,

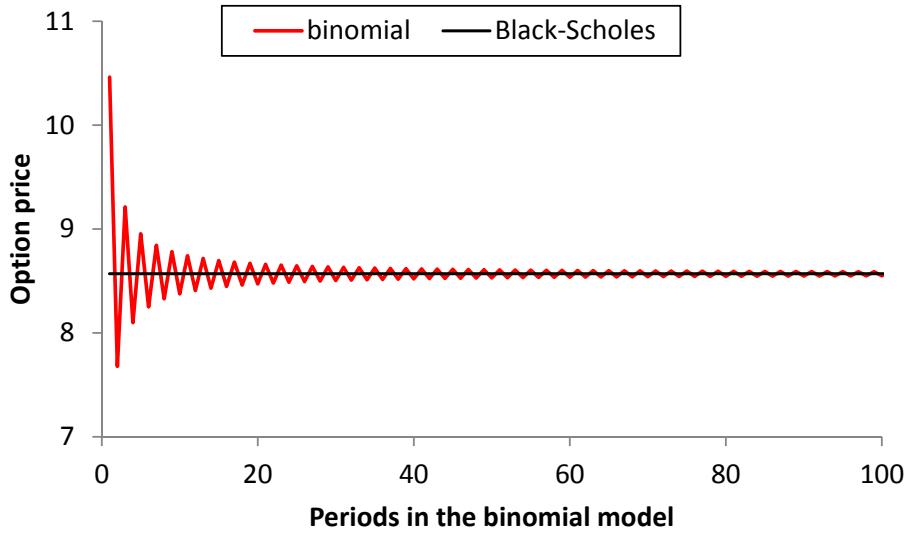


Figure 15.14: Convergence of the binomial option price to the Black-Scholes option price.

The option is a 1-year European call option on a stock currently traded at \$50. The exercise price is \$50, the continuously compounded interest rate is 3% per year, and the volatility is 40%.

then you can compute the option price by setting up a binomial model with 20 or more periods (depending on the required precision) and with u and d chosen as in (15.35).

15.6.4 The implied volatility and shortcomings of the Black-Scholes model

The only input to the Black-Scholes formula which is not directly observable is the volatility σ of the underlying asset. Given the market price of an option, the **implied volatility** of the option is the value of σ you need to put into the Black-Scholes formula to obtain that option price. In fact, market traders often quote option prices in terms of their implied volatilities. Since the Black-Scholes option price is increasing in the value of σ , there is a unique value of σ that does the job. It is impossible to mathematically isolate σ in the formula as it enters in a complicated way in d_1 and d_2 , but it is easy to solve for it using standard computer software, for example by applying **Goal seek** or **Solver** in Excel. Due to the put-call parity, the implied volatility should be the same for a European call and a European put written on the same underlying asset and having identical exercise prices and maturity dates.

If the Black-Scholes model was correct, *all* options on the same asset should have the same implied volatility. In reality, options on the same asset, but with different exercise prices or different maturity dates, often have different implied volatilities. If you fix the underlying asset and the maturity date, you can depict implied volatilities as a function of the exercise price of the options. The relation is referred to as a **volatility smile** since it is often U-shaped with a minimum for an exercise price close to the current price of the underlying asset. Near-the-money options have low implied volatilities, and the implied volatility then increases with the distance between the exercise price and the current price of the underlying asset. The exact shape of the volatility smile seems to depend on the nature of the underlying asset. For options on foreign exchange, the smile is relatively symmetric. For options on a stock index or an individual stock, the smile is

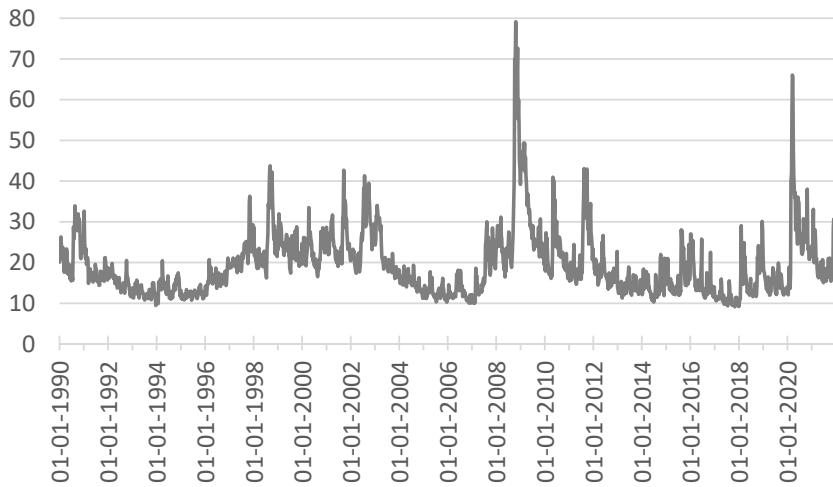


Figure 15.15: The VIX from January 1990 to December 2021.

Weekly observations of the VIX were downloaded from Yahoo Finance on February 4, 2022.

often asymmetric or skewed and then referred to as a volatility skew or volatility smirk.

The non-flat implied volatility curves indicate that the assumptions of the Black-Scholes model do not hold in real life. Empirically, stock volatilities seem to vary stochastically over time peaking during major economic or political crises as shown by Bloom (2009), among others. Hence, we should really use option pricing models allowing for stochastic volatilities. Hull and White (1987), Heston (1993), and others have analyzed such extensions of the Black-Scholes model, but such models are quite complicated and beyond the scope of this presentation.

A frequently used indicator of the market's view on the future volatility of the stock market index is the so-called **VIX**, the volatility index computed by the Chicago Board Options Exchange. Sometimes called the fear index, the main VIX index refers to the volatility of the S&P 500 stock index and is derived from prices of a range of S&P 500 options maturing in 30 days, but there are also VIX-measures for other time horizons and other underlying stock indices. Figure 15.15 shows how the VIX has evolved from the beginning of 1990 to the end of 2021. We see that there are some long periods in which the VIX is low and relatively stable, but there are also periods where the VIX is high and fluctuates a lot. The VIX spiked around the start of the financial crisis in the Fall of 2008 and in March 2020 where the Covid-19 pandemic took off.

Another shortcoming of the Black-Scholes model is that the lognormal distribution for the underlying asset price implies only a tiny probability of large changes of the typically short life of an option. However, we do occasionally observe large, sudden movements in the stock index, the price of an individual stock, or in prices of other types of underlying assets. For the stock index, we have seen several "crashes" where the index drops dramatically during one or a few days. The largest single-day drop in a leading U.S. stock index was the 22% drop in the Dow Jones on 'Black Monday,' October 19, 1987. At the same day, the S&P 500 fell by more than 20%. More recently, the S&P 500 dropped by almost 10% on March 12, 2020 and by another 12% four days later during the beginning of the Covid-19 pandemic. Of course, there are also examples of days with large positive index changes. For example, the S&P 500 rose by almost 17% on March 15, 1933 and recently by more

than 9% on both March 13 and 24 in 2020.

Such extreme movements are highly unlikely with the Black-Scholes assumptions. [Merton \(1976\)](#), [Madan, Carr, and Chang \(1998\)](#), and others have suggested models that extend the Black-Scholes model by incorporating possible jumps in the price of the underlying assets. Also these models and the resulting option pricing formulas are complicated so we will just indicate how jumps affect option prices. Other things equal, the possibility of upward jumps makes a call option more valuable, while the possibility of downward jumps implies that put options are more valuable. Recall that option prices rely on the risk-neutral probability distribution of the price of the underlying asset, not the real-world probability distribution. In general, the risk-neutral probability of bad states exceeds the real-world probability, while the opposite holds for good states. A deep-out-of-the-money put option on the stock index provides insurance against a large drop in the index. While the real-world probability of such a drop might be moderate, the risk-neutral probability of the drop—and thus the put price—can be substantial.

15.7 Option returns

How do we derive the real-world probability distribution of the return r_C on a European call option under the assumptions of the Black-Scholes model? What is the expected option return? We assume that we purchase the call at time 0 and hold it until maturity at time T .

Theorem 15.9

Assume that the Black-Scholes model holds, and let C and P denote the current price of a European call and put option, respectively. Then the return r_C on holding a European call option until maturity satisfies

$$\text{Prob}(r_C = -100\%) = N(-\hat{d}_2), \quad (15.49)$$

$$\text{Prob}(r_C < k) = N\left(\frac{\ln(1 + (1+k)\frac{C}{X})}{\sigma\sqrt{T}} - \hat{d}_2\right), \quad \text{for } k > -1, \quad (15.50)$$

$$E[r_C] = \frac{S(0)e^{\mu T}N(\hat{d}_1) - XN(\hat{d}_2)}{S(0)e^{rT}N(d_1) - XN(d_2)} e^{rT} - 1, \quad (15.51)$$

where

$$\hat{d}_1 = \frac{\ln(S(0)/X) + (\mu + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, \quad \hat{d}_2 = \hat{d}_1 - \sigma\sqrt{T}.$$

The return r_P on the put option satisfies

$$\text{Prob}(r_P = -100\%) = N(\hat{d}_2), \quad (15.52)$$

$$\text{Prob}(r_P < k) = N\left(\hat{d}_2 - \frac{\ln(1 - (1+k)\frac{P}{X})}{\sigma\sqrt{T}}\right), \quad \text{for } k > -1, \quad (15.53)$$

$$E[r_P] = \frac{XN(-\hat{d}_2) - S(0)e^{\mu T}N(-\hat{d}_1)}{XN(-d_2) - S(0)e^{rT}N(-d_1)} e^{rT} - 1. \quad (15.54)$$

Note that \hat{d}_1 and \hat{d}_2 deviate from d_1 and d_2 only by μ replacing r . The real-world

probability that $S(T) > X$ —so that the call ends up in the money and the put out of the money—is $N(\hat{d}_2)$. In contrast, the term $N(d_2)$ that enters the Black-Scholes option pricing formula represents the *risk-neutral* probability that $S(T) > X$. Whenever $\mu > r$, we will have $\hat{d}_2 > d_2$ and thus $N(\hat{d}_2) > N(d_2)$.

Proof

The call expires worthless if $S(T) < X$. In that case, the option owner gets a zero payoff and thus a return of -100% . The probability that this happens is

$$\begin{aligned}\text{Prob}(S(T) < X) &= \text{Prob}(\ln S(T) < \ln X) \\ &= \text{Prob}\left(\frac{\ln S(T) - \ln S(0) - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} < \frac{\ln X - \ln S(0) - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}\right) \\ &= N\left(\frac{\ln X - \ln S(0) - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}\right) = N(-\hat{d}_2),\end{aligned}$$

where we apply Eq. (15.45) with $t = 0$. For any $k > -1 = -100\%$, we find

$$\begin{aligned}\text{Prob}(r_C < k) &= \text{Prob}(S(T) - X < (1+k)C) = \text{Prob}(\ln S(T) < \ln(X + [1+k]C)) \\ &= N\left(\frac{\ln(X + [1+k]C) - \ln S(0) - (\mu - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}\right) \\ &= N\left(\frac{\ln(1 + (1+k)\frac{C}{X})}{\sigma\sqrt{T}} - \hat{d}_2\right),\end{aligned}$$

from which we can calculate the probabilities of the return ending up in any given interval.

The rate of return on the call is the payoff divided by the price and then one is subtracted. Hence, to find the expected call option return, we first determine the expected payoff of the call, i.e. $E[\max(S(T) - X, 0)]$. By the Black-Scholes assumptions, $S(T)$ is lognormally distributed. We can then use Theorem A.5 in Appendix A to get

$$E[\max(S(T) - X, 0)] = S(0)e^{\mu T}N(\hat{d}_1) - XN(\hat{d}_2).$$

This is similar to the calculation of the risk-neutral expected payoff that leads to the Black-Scholes pricing formula, but since we are now interested in the real-world expectation, we keep the parameter μ representing the real-world expected return on the underlying asset. The expected rate of return on the option is then

$$\begin{aligned}E[r_C] &= \frac{E[\max(S(T) - X, 0)]}{C} - 1 \\ &= \frac{S(0)e^{\mu T}N(\hat{d}_1) - XN(\hat{d}_2)}{S(0)N(d_1) - Xe^{-rT}N(d_2)} - 1 \\ &= \frac{S(0)e^{\mu T}N(\hat{d}_1) - XN(\hat{d}_2)}{S(0)e^{rT}N(d_1) - XN(d_2)} e^{rT} - 1.\end{aligned}$$

The formulas (15.52)–(15.54) for the put returns are derived with similar arguments.

The next example illustrates the option return distribution and the expected option returns, and how both the distribution and the expectation depend on the exercise price.

Example 15.10

Table 15.7 applies the above formulas to illustrate the returns on options in the Black-Scholes setting. The underlying asset price is currently $S(0) = 100$ and evolves in line with the Black-Scholes assumptions with $\mu = 0.1$ being larger than the riskfree rate $r = 0.01$. We consider European call and put options with one year to maturity and with three different exercise prices, $X \in \{80, 100, 120\}$. For call options, the price is naturally decreasing in X . For a high X , the probability that the call expires worthless and the return is -100% is large. On the other hand, if the stock price exceeds X , the option will often provide a high percentage return due to the low option price. Hence, the distribution of returns above -100% is relatively flat with a long right tail. For example, in the case of $X = 120$, the call expires worthless with a 69.6% probability, but there is a probability of 14.3% of getting an option return larger than 500%. The expected return on the call is 122.3%. With a lower X , there is a bigger chance that the call ends up in the money, but the price is larger so the option return is more likely to be fairly modest. For example, with $X = 80$, there is only a 6.5% probability that the call expires worthless, but also only a 25% chance that the return will exceed 100%. In this case, the expected return on the call is 41.5%.

Let us turn to put options. With a low X , there is a large probability that the put ends up worthless, but on the other hand the price is so low that you will get a significant rate of return in the relatively unlikely event that the price of the underlying asset ends up even lower than X . With $X = 80$, the put is worthless with a 93.5% probability, but there is still a 2.7% probability of a rate of return above 500%. The expected return is -60.7% . With a high X , such as $X = 120$ in the table, the put is less likely to end up worthless but, since the price is higher, the rate of return is typically modest. The expected put return with $X = 120$ is -30.6% .

In the example above, the expected option return is higher than the riskfree rate for the call options and lower than the riskfree rate for the put options. This is no coincidence and will always be the case when $\mu > r$, i.e. the expected return on the underlying exceeds the riskfree rate. Note that this is consistent with the conclusions in the one-period binomial model. The way the options' expected returns vary with X also holds very generally, see Coval and Shumway (2001).

Why would anyone be interested in purchasing a put option when it has such a low expected return? The primary motivation for purchasing an option is to obtain protection against a certain risk, not to get a high expected return. If you have a long position in the underlying asset, you might fear that its price will drop significantly, and you can protect yourself against that situation by purchasing a put option. If the price of the underlying asset falls below the exercise price, the payoff of the put will compensate you for your loss on the underlying asset. The put acts as an insurance. In general, people are willing to pay for insurance against bad things even though the insurance contract itself has a negative expected return. Finally, note that the expected large negative return on investing in a put option suggests that institutions selling such options can expect to make large profits. Of course, the sellers of put options have to be ready to accept big losses in the case of a large drop in the price of the underlying asset.

	Call options			Put options		
	$X = 80$	$X = 100$	$X = 120$	$X = 80$	$X = 100$	$X = 120$
BS price	21.86	8.43	2.34	1.07	7.44	21.15
Expected return	41.5%	73.9%	122.3%	-60.7%	-44.2%	-30.6%
<i>Return distribution</i>						
-100%	6.5%	34.5%	69.6%	93.5%	65.5%	30.4%
-99.9% to -50%	12.6%	7.9%	1.7%	0.4%	6.7%	17.5%
-50% to 0%	18.8%	7.9%	1.6%	0.4%	6.2%	19.7%
0% to 50%	20.1%	7.6%	1.6%	0.4%	5.5%	17.1%
50% to 100%	16.7%	7.0%	1.5%	0.4%	4.7%	10.5%
100% to 150%	11.5%	6.3%	1.4%	0.3%	3.7%	4.0%
150% to 200%	6.9%	5.6%	1.4%	0.3%	2.8%	0.8%
200% to 250%	3.7%	4.8%	1.3%	0.3%	2.0%	0.1%
250% to 300%	1.8%	4.0%	1.2%	0.3%	1.3%	0.0%
300% to 350%	0.8%	3.3%	1.2%	0.3%	0.8%	0.0%
350% to 400%	0.4%	2.6%	1.1%	0.3%	0.4%	0.0%
400% to 450%	0.2%	2.1%	1.1%	0.2%	0.2%	0.0%
450% to 500%	0.1%	1.6%	1.0%	0.2%	0.1%	0.0%
More than 500%	0.0%	5.0%	14.3%	2.7%	0.0%	0.0%

Table 15.7: Option returns in the Black-Scholes model.

For six different options, the table shows the option price, the expected lifetime return on the option, as well as the probability of the return being either equal to -100% or in each of the specified intervals. Both the expected return and the probabilities are based on the real-world probabilities, not the risk-neutral probabilities. The options are European and mature in one year. The underlying asset price is currently $S = 100$ and satisfy the Black-Scholes assumptions of Section 15.6.2 with $\mu = 0.1$ and $\sigma = 0.2$. The underlying asset is not paying dividends in the life of the options. The continuously compounded riskfree rate is $r = 0.01$.

15.8 Concluding remarks

This chapter has described the main types of financial options and introduced the basic models for the pricing of options on stocks and similar assets. As already suggested by the discussions in the chapter, these models are not perfect, especially for certain underlying assets. For example, the basic models assume constant interest rates which is not appropriate when pricing long-term options or options on assets having very interest rate sensitive prices such as bonds. For coverage of other models and more exotic types of options, the interested reader is referred to specialized textbooks on options and other derivative securities. The popular and accessible book by Hull (2021) is a good place to start. For the mathematically savvy and courageous readers, there are also numerous books developing advanced models for option pricing which, despite their complexity, are widely applied in today's option markets.

As shown in this chapter, options can be used to hedge certain risks, but we have not discussed in detail when or in which situations it is optimal for a given investor to do that. When should the investor buy the specific insurance that an option offers? Obviously, this depends on the preferences and the overall financial situation of the investor, so there is no general answer to this question. Some corporate finance textbooks like Hillier, Grinblatt, and Titman (2024) include a discussion of when a company may want to use options, for example, to reduce expected tax payments or to reduce the risk of bankruptcy.

15.9 Exercises

Exercise 15.1. You are asked to price a European put option based on the following information:

Current price of underlying asset	$S = 100$
Riskfree rate per year	$r = 0.10$
Exercise price	$X = 110$
Time to expiration, in years	$T = 2$

You can assume that the underlying asset does not pay any dividends during the next two years.

- (a) Construct a two-period binomial tree. In each period the underlying asset grows by 20% or drops by 10%.
- (b) Determine the possible payoffs of the put option at expiration.
- (c) Calculate the price of the put today.
- (d) Determine a portfolio of the underlying asset and the riskfree asset so that the portfolio replicates the put option. Explain how the portfolio is rebalanced through the tree.
- (e) Determine the option price according to the Black-Scholes formula with the volatility set to $\sigma = 0.15$.
- (f) What is the initial replicating portfolio according to the Black-Scholes model?
- (g) Based on the binomial model, would it be advantageous if the option could also be exercised after one year? What is then the current option price?
- (h) Would it be advantageous if the option could also be exercised right now? Explain!

Exercise 15.2. The stocks in the company SOS are traded at \$480 per share. The stock price volatility is 30% per year, and the riskfree rate is 8% per year (continuously compounded). The company is expected not to pay any dividends in the near future. Assume that the conditions for applying the Black-Scholes formula are satisfied.

- (a) A European call on an SOS stock has an exercise price of \$475 and expires in 75 days. What is the price of the option?
- (b) An American call on an SOS stock has an exercise price of \$475 and expires in 75 days. What is the price of the option?
- (c) Find a portfolio of SOS stocks and the riskfree asset that replicates the European call option considered in question (a).
- (d) If SOS stocks trade at \$500 per share tomorrow, what is then the value of the portfolio found above? What is then the price of the European call option? Compare and explain!
- (e) What is the price today of a European put on an SOS stock with an exercise price of \$475 and expiry in 75 days? If the put trades at a lower price than this, whereas the call trades at the price found in (a), how can you lock in a riskfree profit?

Exercise 15.3. A European call option matures in 6 months and has an exercise price of \$100. The underlying stock is traded at \$98 per share and does not pay dividends in the next year. The riskfree rate is 10% per year.

- (a) Determine the no-arbitrage price bounds for the call option.
- (b) Is there an arbitrage if the call trades at \$2? Or at \$4? If so, explain the arbitrage strategy.
- (c) A European put on the stock with the same maturity and exercise price trades at \$0.9. Can you now determine a unique price for the call option? If so, what is the price? Explain!
- (d) The volatility of the underlying stock is 0.25. Compute the price of both the call and the put according to the Black-Scholes formula. Does the call price satisfy the bounds found in question (a)?

Exercise 15.4. The stocks of a listed company currently trade at \$100 per share. The company does not pay dividends. The current riskfree rate is 5% per year, continuously compounded. Both forwards and call and put options on the stock are traded.

- (a) Determine the no-arbitrage forward price on the stock for delivery in three months. How could you profit if you could buy or sell forwards with a lower delivery price?

Based on a fundamental analysis of the company, you expect the stock price to drop by at least 10% over the next three months.

- (b) Which transaction using forwards on the stocks would be profitable to you if your prediction comes true?
- (c) Which transaction using options on the stocks would be profitable to you if your prediction comes true? Briefly compare with the strategy involving forwards detected in the previous question.

The stock price is assumed to behave in line with the Black-Scholes model with a volatility of $\sigma = 0.6$.

- (d) What is the price of a 3-month European call on the stock with an exercise price of \$90?
- (e) What is the price of a 3-month European put on the stock with an exercise price of \$90? If the market price of the put is lower than this, but the market price of the call is identical to the price found in the previous question, how can you lock in a riskfree profit?
- (f) What can you say about the prices of 3-month American call and put options with the same exercise price compared to the prices of their European counterparts?

Exercise 15.5. A Danish company expects to receive 1.200.000 USD in six months. The management is concerned about the developments in the DKK/USD exchange rate over the period and considers hedging the position. The CFO has gathered the following information:

Spot DKK/USD exchange rate:	5.5547
6-month interest rate on DKK:	0.55% per year
6-month interest rate on USD:	0.33% per year
Volatility of DKK/USD rate:	4.5% per year

- (a) Describe the factors that are relevant for the decision whether to hedge the given position or not.
- (b) The CFO suggests to use a forward contract as a hedging instrument. Determine the 6-month forward DKK/USD exchange rate. What are the pros and cons of using forwards as the hedging instrument?
- (c) As an ambitious new employee of the company you suggest to hedge the position using options. Which type of options would be appropriate in this case? What are the pros and cons of using options as the hedging instrument?
- (d) Determine the price of the relevant option if the exercise price is 5.50 DKK/USD. How costly is it to hedge the entire position using options?

Exercise 15.6. You want to price options on a stock and have the following information:

Time to expiration	6 months
Standard deviation	50% per year
Exercise price	\$50
Stock price	\$50
Interest rate	3% per year

The stock pays no dividends within the next 6 months.

- (a) Use the Black-Scholes formula to find the value of a call option.
- (b) Use the Black-Scholes formula to find the value of a put option.
- (c) Recalculate the value of the call option, successively substituting one of the changes below while keeping the other parameters as listed above:
 - i. Time to expiration = 3 months.
 - ii. Standard deviation = 25% per year.
 - iii. Exercise price = \$55.
 - iv. Stock price = \$55.

- v. Interest rate = 5% per year.

Consider each scenario independently. Confirm that the option value changes in accordance with the general predictions in Table 15.5.

Exercise 15.7. Consider a two-year European put option with a strike price of \$104 on a stock whose current price is \$100. The riskfree interest rate is 5% per year. The stock is not expected to pay dividends in the next two years.

- Compute the current price of the European put option using a two-period binomial tree with up- and down-factors $u = 1.2$ and $d = 0.8$. What is the current replicating portfolio and how does that change over time?
- Use the same two-period binomial tree to value an American put option with the same characteristics as the European put considered above. Is it possible that early exercise will be profitable?

Exercise 15.8. Let F denote the forward price for a forward on a stock with delivery in T years from now. Consider a European call option and a European put option written on the same stock and having the same time-to-maturity T . The strike price of each option is equal to the forward price F .

- Show that a portfolio of a long call option and a short put option replicates the payoff from the forward contract. Explain why this implies that the call option and the put option should currently trade at identical prices. For this conclusion to hold, do you need to assume that the stock does not pay dividends before the delivery date?
- Suppose that the stock price is currently \$50 and that the forward price for delivery in one year is \$54. For \$6, you can buy a one-year European call option on the stock with an exercise price of \$54, whereas a similar European put option costs \$4. Explain how you can make a riskfree profit by trading in these securities. How big a profit can you make?

Exercise 15.9. Suppose you have \$100,000 to invest over the next six months and that you consider the following three strategies:

- P1: Invest \$100,000 in stocks of the company ABC. The unit price of the stock now is \$100.
- P2: Invest \$100,000 in at-the-money call options written on an ABC stock. The call options are of the European style and expire in six months. The price per option is \$10.
- P3: Invest \$10,000 in the before-mentioned call options and deposit \$90,000 in a bank account paying a riskfree interest rate of 2% over the next six months. (Note: the interest rate is not annualized but applies to the six month period.)

You want to compare the rate of return on each strategy over the following six months. The rate of return will depend on the unit price $S(T)$ of ABC stocks six months from now. The stock pays no dividends in this six-month period.

- Compute the rate of return on the three strategies for each of the following values of $S(T)$: 80, 90, 100, 110, 120. Illustrate in a diagram how the rate of return on the three strategies depends on $S(T)$ (focus on $S(T)$ in the range between \$80 and \$120).
- How does the choice between the three strategies depend on your expectations about $S(T)$? Which strategy would you say is the most risky? How does the choice between the three strategies depend on your risk aversion?
- What is the price of an at-the-money six-month European put option on an ABC stock?

Exercise 15.10. The stocks of the company Hypothetics Inc. currently trade at \$40 per share. The company has announced that it will not pay out any dividends within the next 12 months. The following questions consider forwards and options on a single share/stock of Hypothetics. All the forwards and options considered expire in six months from now. Suppose that the assumptions behind the Black-Scholes model are satisfied and that the stock price volatility is $\sigma = 0.5$. A default-free zero-coupon bond maturing in six months and having a face value of \$1,000 is currently traded at \$990 corresponding to a continuously compounded interest rate of 2.0101% per year.

- (a) What is the theoretically fair (no-arbitrage) forward price?
- (b) Suppose you are offered to take a position in a forward contract with a forward price of \$41. Find and describe an arbitrage strategy.
- (c) What is the price of a European call option with an exercise price of \$38?
- (d) What is the price of a European put option with an exercise price of \$38?
- (e) What can you say about the price of an American call option and the price of an American put option, both having an exercise price of \$38? (You are not supposed to do any extensive computations to answer this question.)

Exercise 15.11. This problem considers options on a stock in the company *Made-Up Inc.* The options mature in one year and have a strike price of \$110. *Made-Up* stocks currently trade at \$100 per share. The company does not pay dividends in the next year.

- (a) Compute the current price of a European call option using a two-period binomial tree with up- and down-factors $u = 1.25$ and $d = 0.8$. Assume that the 6-month riskfree rate is 1% in both periods.
- (b) Use the same tree to compute the current price of a European put option.
- (c) Use the same tree to compute the current price of an American put option. Is it possible that early exercise will be profitable?
- (d) What is the current price of an American call option? Explain.
- (e) Verify that the prices of the European options are consistent with the put-call parity.
- (f) Apply the Black-Scholes formula to value the European call option if the volatility is $\sigma = 0.3$ and if the volatility is $\sigma = 0.35$. What is the value of σ so that the call price according to the Black-Scholes formula agrees with the call price computed with the binomial tree above?

Exercise 15.12. Stocks in the company *Imaginary Inc.* currently sell at \$50 per share. The company has explicitly stated that it is not going to pay any dividends for the next two years. You are considering buying some put options on the stock. The options have an exercise price of \$50 and expire in one year from now. You always evaluate options using a two-period binomial tree and assume that the stock price will either increase or decrease by 25% each period. The riskfree rate is 1% for any six-month period (not annualized).

- (a) Compute the price of a European put option on the stock using the two-period binomial tree.
- (b) Compute the price of an American put option on the stock using the two-period binomial tree. What is the early exercise premium of the American option?
- (c) A one-year European call option on the same stock is currently selling at \$7.25. If you trust the put price you computed in (a), is there an arbitrage opportunity? Explain!
- (d) You are offered to purchase a “Power Put” on the stock. It provides a payoff one year from now equal to the square of the payoff of the European put considered in (a). What is the current value of such a security according to your two-period binomial tree?

Exercise 15.13. You are considering purchasing a put option on a stock. The current stock price is 100, the exercise price is 105, the option matures in exactly one year from now, and the volatility of the stock price is $\sigma = 0.32$. There are no dividend payments from the stock during the next year. The continuously compounded riskfree rate is 2% per year.

- (a) Compute the price of a European put option on the stock using the Black-Scholes model.
- (b) What is the initial replicating portfolio according to the Black-Scholes model? Briefly explain what the term “replicating portfolio” means.
- (c) Suppose that in a month from now, the stock price has changed to 90. What is then the Black-Scholes price of the option? And what is then the value of the initial replicating portfolio? Compare and explain.

You also want to price the option using a two-period binomial model with a period length of $\Delta t = 0.5$ years. In an attempt to construct a tree that leads to a similar option price as the Black-Scholes model you use $u = 1.25$ as this is close to $e^{\sigma\sqrt{\Delta t}} \approx 1.2539$, and you put $d = 1/u$. Assume that the non-annualized simple riskfree rate is 1% over each 6-month period.

- (d) Compute the price of the European put option using the two-period binomial model and compare with the Black-Scholes price.
- (e) Compute the price of the corresponding American put option using the two-period binomial model and compare with the price of the European put.

Exercise 15.14. You want to value some options on an ETF tracking the stock market index. The options mature in one year and have a strike price of \$90. The ETF is currently trading at \$100 and does not pay dividends during the next year. You intend to use a two-period binomial tree to value the options. You assume that in every six-month period the value of the ETF is either going up by 20% or down by 15%. You estimate that the true or real-world probability of the up-movement in the ETF price in any six-month period is 60%. The risk-free rate is 0.5% in every six-month period.

- (a) For each of the end-nodes in the binomial tree calculate the rate of return on the ETF. What is the expected rate of return on the ETF over the next year if you use the real-world up-probability?
- (b) Calculate the risk-neutral probability of an up-movement in the ETF price in every six-month period. What is the expected rate of return on the ETF over the next year if you use the risk-neutral up-probability? Compare this expected return with the riskfree return over the one-year period.
- (c) Compute the current price of a European put option using the two-period binomial tree. What is the expected rate of return on the put option if you use the real-world up-probability? You should find the expected return on the put is negative. Why would anyone then want to purchase the put option?
- (d) Use the put-call parity to compute the arbitrage-free price C of a European call option. If the call is traded at a price \$2 lower than C , which trading strategy would give you a riskfree profit?

Exercise 15.15. You want to value put options on a stock in the company *Monsters, Inc.* The options mature in nine months and have a strike price of \$220. *Monsters* stocks currently trade at \$200 per share. The company does not pay dividends during the next nine months.

- (a) Construct a three-period binomial tree with up- and down-factors $u = 1.2$ and $d = 0.8$ that shows how the stock price can evolve over the next nine months.
- (b) Assume that the three-month riskfree rate is 1.0% in each three-month period. Compute the current price of a European put option using the three-period binomial tree. Also compute the current price of an American put option using the three-period binomial tree. Is it possible that early exercise will be profitable? If so, in which nodes of the binomial tree should you exercise the option?
- (c) Apply the Black-Scholes formula to compute the price of the European put option assuming a volatility of $\sigma = 0.35$. What is the value of σ so that the put price according to the Black-Scholes formula agrees with the European put price computed with the binomial tree in Question (b)?

Exercise 15.16. Your boss has asked to you to price some options on the stocks of the company AiryFairy. The stock is currently traded at a price of \$100 per share. The company does not pay dividends over the next year. You intend to use a binomial model in which the stock price will either increase by 10% or decrease by 10% over any 6-month period. The riskfree rate of return over any 6-month period is 1%.

- (a) Using a two-period binomial tree, calculate the value of a European call option and of a European put option, both written on the AiryFairy stock and having an exercise price of \$100 and a time-to-maturity of one year.
- (b) A so-called *chooser option* is also traded. This contract gives you the right after 6 months to choose either a European call option or a European put option, both having an exercise price of \$100 and maturing 6 months later, i.e. one year from the current date. Based on the two-period binomial tree, what is the value of the chooser option today?

- (c) Suppose that the European call is traded at the price found in Question (a), whereas the European put option is traded at a higher price than found in Question (a). Can you set up an arbitrage strategy? If so, describe this strategy. If not, explain why no arbitrage strategy is possible.

Exercise 15.17. Consider European call options having the same underlying asset and expiration date, but different exercise prices. Let $C(X)$ denote the price of such an option for an exercise price of X . Let X_0 and X_1 be positive numbers with $X_0 < X_1$. For any $\lambda \in [0,1]$, define the convex combination $X_\lambda = (1 - \lambda)X_0 + \lambda X_1$. Show how you can construct an arbitrage strategy if

$$C(X_\lambda) > (1 - \lambda)C(X_0) + \lambda C(X_1).$$

Conclude that the opposite has to hold and thus that $C(X)$ is a convex function of the exercise price.

Exercise 15.18. (Mathematically challenging!) Consider the Black-Scholes option pricing formula (15.37). Show that

$$\frac{\partial C}{\partial S} = N(d_1)$$

by going through the following steps:

1. Find $N'(x)$.
2. Show that $SN'(d_1) = Xe^{-r[T-t]}N'(d_2)$.
3. Compute $\frac{\partial d_1}{\partial S}$ and $\frac{\partial d_2}{\partial S}$.
4. Now verify that $\frac{\partial C}{\partial S} = N(d_1)$.

Exercise 15.19. (Mathematically challenging!) Consider the Black-Scholes option pricing formula (15.37). Compute the following derivatives and determine, if possible, whether they are positive or negative:

- (i) $\frac{\partial C}{\partial X}$.
- (ii) $\Gamma := \frac{\partial^2 C}{\partial S^2}$ (the *gamma* of the option).
- (iii) $\rho := \frac{\partial C}{\partial r}$ (the *rho* of the option).
- (iv) $\Theta := -\frac{\partial C}{\partial T}$ (the *theta* of the option).
- (v) $\frac{\partial C}{\partial \sigma}$ (the *vega* of the option).

APPENDIX A

The lognormal distribution

A random variable Y is said to be lognormally distributed if the random variable $X = \ln Y$ is normally distributed. In the following we let m be the mean of X and s^2 be the variance of X so that

$$X = \ln Y \sim N(m, s^2).$$

The probability density function for X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{(x-m)^2}{2s^2}\right\}, \quad x \in \mathbb{R}.$$

The next theorem states the probability density function, the cumulative distribution function, and the percentiles of a lognormally distributed random variable.

Theorem A.1

Assume $X = \ln Y \sim N(m, s^2)$. Then, for any $y > 0$, the probability density function and the cumulative distribution functions for the rate of return Y are

$$f_Y(y) = \frac{1}{y\sqrt{2\pi s^2}} \exp\left\{-\frac{(\ln y - m)^2}{2s^2}\right\}, \quad (\text{A.1})$$

$$F_Y(y) = N\left(\frac{\ln y - m}{s}\right), \quad (\text{A.2})$$

and $f_Y(y) = F_Y(y) = 0$ for $y \leq 0$. The percentiles in the distribution of Y are given by

$$p = \text{Prob}(Y < y_p) \Leftrightarrow y_p = \exp\left\{sN^{-1}(p) + m\right\}. \quad (\text{A.3})$$

In particular, the median is

$$\text{Med}[Y] = e^m. \quad (\text{A.4})$$

Proof

Note that $Y \leq y$ if and only if $X \leq \ln y$. Hence, the cumulative distribution function for Y is given by

$$F_Y(y) = \text{Prob}(Y \leq y) = \text{Prob}(X \leq \ln y) = N\left(\frac{\ln y - m}{s}\right).$$

The probability density function f_Y follows from differentiation of F_Y :

$$\begin{aligned} f_Y(y) &= F'_Y(y) = N'\left(\frac{\ln y - m}{s}\right) \frac{1}{ys} \\ &= \frac{1}{ys} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\ln y - m)^2}{2s^2}\right\} \\ &= \frac{1}{y\sqrt{2\pi s^2}} \exp\left\{-\frac{(\ln y - m)^2}{2s^2}\right\}. \end{aligned}$$

The percentiles follows from

$$\begin{aligned} p &= \text{Prob}(Y < y_p) = F_Y(y_p) = N\left(\frac{\ln y_p - m}{s}\right) \\ \Leftrightarrow \quad N^{-1}(p) &= \frac{\ln y_p - m}{s} \\ \Leftrightarrow \quad y_p &= \exp\left\{sN^{-1}(p) + m\right\}. \end{aligned}$$

Since $N^{-1}(0.5) = 0$, the median is $\text{Med}[Y] = y_{0.5} = e^m$.

The next theorem turns out to be useful in calculating the key moments of a lognormal random variable.

Theorem A.2

Assume that $X = \ln Y \sim N(m, s^2)$. For any $k \in \mathbb{R}$ we have

$$\mathbb{E}[Y^k] = \mathbb{E}[e^{kX}] = \exp\left\{km + \frac{1}{2}k^2s^2\right\}. \quad (\text{A.5})$$

Proof

Since $Y = e^X$, it is clear that $Y^k = e^{kX}$. Per definition we have

$$\mathbb{E}[e^{kX}] = \int_{-\infty}^{+\infty} e^{kx} f_X(x) dx = \int_{-\infty}^{+\infty} e^{kx} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}} dx.$$

Manipulating the exponent we get

$$\begin{aligned} \mathbb{E}[e^{kX}] &= e^{km + \frac{1}{2}k^2s^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}s^2} e^{-\frac{1}{2s^2}[(x-m)^2 - 2k(x-m)s^2 + k^2s^4]} dx \\ &= e^{km + \frac{1}{2}k^2s^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}s^2} e^{-\frac{(x-[m+ks^2])^2}{2s^2}} dx \\ &= e^{km + \frac{1}{2}k^2s^2}, \end{aligned}$$

where the last equality is due to the fact that the function

$$x \mapsto \frac{1}{\sqrt{2\pi}s^2} e^{-\frac{(x-[m+ks^2])^2}{2s^2}}$$

is a probability density function, namely the density function for an $N(m + ks^2, s^2)$ distributed random variable.

Using this theorem, we can compute the key moments of the lognormally distributed random variable $Y = e^X$.

Theorem A.3

Assume $X = \ln Y \sim N(m, s^2)$. Then the key moments for $Y = e^X$ are

$$\mathbb{E}[Y] = e^{m + \frac{1}{2}s^2}, \quad (\text{A.6})$$

$$\text{Var}[Y] = e^{2m+s^2} (e^{s^2} - 1), \quad (\text{A.7})$$

$$\text{Std}[Y] = e^{m + \frac{1}{2}s^2} \sqrt{e^{s^2} - 1}, \quad (\text{A.8})$$

$$\text{Skew}[Y] = (e^{s^2} + 2) \sqrt{e^{s^2} - 1}, \quad (\text{A.9})$$

$$\text{Kurt}[Y] = e^{4s^2} + 2e^{3s^2} + 3e^{2s^2} - 6. \quad (\text{A.10})$$

Proof

The expectation follows directly from an application of Theorem A.2 with $k = 1$.

Applying $k = 2$ in the same theorem, we get $\mathbb{E}[Y^2] = e^{2m+2s^2}$. Hence,

$$\text{Var}[Y] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = e^{2m+2s^2} - (e^{m + \frac{1}{2}s^2})^2 = e^{2m+2s^2} - e^{2m+s^2} = e^{2m+s^2} (e^{s^2} - 1),$$

which shows (A.7). The standard deviation follows as the square root of the variance.

For the calculation of the skewness, first note that $(Y - a)^3 = Y^3 - 3Y^2a + 3Ya^2 - a^3$,

so

$$\begin{aligned}
E[(Y - E[Y])^3] &= E[Y^3] - 3E[Y^2]E[Y] + 3E[Y](E[Y])^2 - (E[Y])^3 \\
&= E[Y^3] - 3E[Y^2]E[Y] + 2(E[Y])^3 \\
&= e^{3m+\frac{9}{2}s^2} - 3e^{3m+\frac{5}{2}s^2} + 2e^{3m+\frac{3}{2}s^2} \\
&= e^{3m+\frac{3}{2}s^2} (e^{3s^2} - 3e^{s^2} + 2) \\
&= e^{3m+\frac{3}{2}s^2} (e^{s^2} + 2) (e^{s^2} - 1)^2,
\end{aligned}$$

where the third equality follows from Theorem A.2. The skewness is

$$\text{Skew}[Y] = \frac{E[(Y - E[Y])^3]}{(Std[Y])^3} = \frac{e^{3m+\frac{3}{2}s^2} (e^{s^2} + 2) (e^{s^2} - 1)^2}{e^{3m+\frac{3}{2}s^2} (e^{s^2} - 1)^{3/2}} = (e^{s^2} + 2) (e^{s^2} - 1)^{1/2},$$

which confirms (A.9).

For the calculation of the kurtosis, recall that $(Y - a)^4 = Y^4 - 4Y^3a + 6Y^2a^2 - 4Ya^3 + a^4$, so

$$\begin{aligned}
E[(Y - E[Y])^4] &= E[Y^4] - 4E[Y^3]E[Y] + 6E[Y^2](E[Y])^2 - 4E[Y](E[Y])^3 + (E[Y])^4 \\
&= E[Y^4] - 4E[Y^3]E[Y] + 6E[Y^2](E[Y])^2 - 3(E[Y])^4 \\
&= e^{4m+8s^2} - 4e^{4m+5s^2} + 6e^{4m+3s^2} - 3e^{4m+2s^2} \\
&= e^{4m+2s^2} (e^{6s^2} - 4e^{3s^2} + 6e^{s^2} - 3) \\
&= e^{4m+2s^2} (e^{s^2} - 1)^2 (e^{4s^2} + 2e^{3s^2} + 3e^{2s^2} - 3),
\end{aligned}$$

where the third equality follows from Theorem A.2. The kurtosis is

$$\begin{aligned}
\text{Kurt}[Y] &= \frac{E[(Y - E[Y])^4]}{(Std[Y])^4} - 3 \\
&= \frac{e^{4m+2s^2} (e^{s^2} - 1)^2 (e^{4s^2} + 2e^{3s^2} + 3e^{2s^2} - 3)}{e^{4m+2s^2} (e^{s^2} - 1)^2} - 3 \\
&= e^{4s^2} + 2e^{3s^2} + 3e^{2s^2} - 6,
\end{aligned}$$

which shows (A.10).

The next theorem provides an expression of the truncated mean of a lognormally distributed random variable, that is the mean of the part of the distribution that lies above some level. This result turns out to be useful in option pricing as the subsequent theorem will indicate. We define the indicator variable $1_{\{Y>K\}}$ to be equal to 1 if the outcome of

the random variable Y is greater than the constant K and equal to 0 otherwise.

Theorem A.4

If $X = \ln Y \sim N(m, s^2)$ and $K > 0$, then we have

$$\begin{aligned} E[Y \mathbf{1}_{\{Y>K\}}] &= e^{m+\frac{1}{2}s^2} N\left(\frac{m - \ln K}{s} + s\right) \\ &= E[Y] N\left(\frac{m - \ln K}{s} + s\right). \end{aligned}$$

Proof

Because $Y > K \Leftrightarrow X > \ln K$, it follows from the definition of the expectation of a random variable that

$$\begin{aligned} E[Y \mathbf{1}_{\{Y>K\}}] &= E\left[e^X \mathbf{1}_{\{X>\ln K\}}\right] \\ &= \int_{\ln K}^{+\infty} e^x \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}} dx \\ &= \int_{\ln K}^{+\infty} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m+s^2])^2}{2s^2}} e^{\frac{2ms^2+s^4}{2s^2}} dx \\ &= e^{m+\frac{1}{2}s^2} \int_{\ln K}^{+\infty} f_{\bar{X}}(x) dx, \end{aligned}$$

where

$$f_{\bar{X}}(x) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m+s^2])^2}{2s^2}}$$

is the probability density function for an $N(m + s^2, s^2)$ distributed random variable. The calculations

$$\begin{aligned} \int_{\ln K}^{+\infty} f_{\bar{X}}(x) dx &= \text{Prob}(\bar{X} > \ln K) \\ &= \text{Prob}\left(\frac{\bar{X} - [m + s^2]}{s} > \frac{\ln K - [m + s^2]}{s}\right) \\ &= \text{Prob}\left(\frac{\bar{X} - [m + s^2]}{s} < -\frac{\ln K - [m + s^2]}{s}\right) \\ &= N\left(-\frac{\ln K - [m + s^2]}{s}\right) \\ &= N\left(\frac{m - \ln K}{s} + s\right) \end{aligned}$$

complete the proof.

A call option has a payoff of the form $\max(0, Y - K)$, where K is the exercise price (or strike price) and Y is the value of the underlying asset at the exercise date of the option. In some option pricing models, this value Y is assumed to be lognormally distributed, and then the following theorem is crucial for determining the option price. In particular, this applies to the Black-Scholes model described in Section 15.6.

Theorem A.5

If $X = \ln Y \sim N(m, s^2)$ and $K > 0$, we have

$$\begin{aligned} \mathbb{E} [\max (0, Y - K)] &= e^{m + \frac{1}{2}s^2} N \left(\frac{m - \ln K}{s} + s \right) - KN \left(\frac{m - \ln K}{s} \right) \\ &= \mathbb{E} [Y] N \left(\frac{m - \ln K}{s} + s \right) - KN \left(\frac{m - \ln K}{s} \right). \end{aligned}$$

Proof

Note that

$$\begin{aligned} \mathbb{E} [\max (0, Y - K)] &= \mathbb{E} [(Y - K) \mathbf{1}_{\{Y>K\}}] \\ &= \mathbb{E} [Y \mathbf{1}_{\{Y>K\}}] - K \text{Prob} (Y > K). \end{aligned}$$

The first term is known from Theorem A.4. The second term can be rewritten as

$$\begin{aligned} \text{Prob} (Y > K) &= \text{Prob} (X > \ln K) \\ &= \text{Prob} \left(\frac{X - m}{s} > \frac{\ln K - m}{s} \right) \\ &= \text{Prob} \left(\frac{X - m}{s} < -\frac{\ln K - m}{s} \right) \\ &= N \left(-\frac{\ln K - m}{s} \right) \\ &= N \left(\frac{m - \ln K}{s} \right). \end{aligned}$$

The claim now follows immediately.

APPENDIX B

The Greek alphabet

Lower case	Upper case	Name	Sound/pronunciation
α	A	Alpha	al-fah
β	B	Beta	bay-tah
γ	Γ	Gamma	gam-ah
δ	Δ	Delta	del-tah
ε	E	Epsilon	ep-si-lon
ζ	Z	Zeta	zay-tah
η	H	Eta	ay-tah
θ	Θ	Theta	thay-tah
ι	I	Iota	eye-o-tah
κ	K	Kappa	cap-ah
λ	Λ	Lambda	lamb-dah
μ	M	Mu	mew
ν	N	Nu	new
ξ	Ξ	Xi	zzEye
\o	O	Omicron	om-ah-cron
π	Π	Pi	pie
ρ	R	Rho	row
σ	Σ	Sigma	sig-ma
τ	T	Tau	tawh
υ	Υ	Upsilon	oop-si-lon
φ	Φ	Phi	figh or fie
χ	X	Chi	kigh
ψ	Ψ	Psi	sigh
ω	Ω	Omega	o-may-gah

The suggested sound/pronunciation was taken from <http://www.keyway.ca/htm2002/greekal.htm> but the link does not appear to be active at the time of writing (July 2022).

APPENDIX C

Short answers to selected exercises

Chapter 2

- 2.1 (a) 15.82%. Note: Adjusted closing prices are sometimes revised, so the returns you calculate might depend somewhat on when you downloaded the data.
- 2.2 (a) 3.
(b) Ranging from -255% to 385% .
- 2.3 (a) 20%, 30%, 50%.
(b) Returns of -40% , -20% , 8% on individual stocks, -10% on portfolio. New weights are 13%, 27%, 60%.
(c) Assuming you can trade fractions of stocks, you should buy 40 A and 7.5 B and sell -8.3333 C without adding or withdrawing cash.
- 2.4 Returns: -12.2% , -0.5% , -40.7% .

Chapter 3

- 3.1 (a) 8.75%, 7%.
(b) $79.6875(\%)^2$, $57(\%)^2$.
(c) $156.25(\%)^2$, $106(\%)^2$.
(d) Approx. 8.927%, 7.550%.
(e) $16.25(\%)^2$, approx. 0.2411.
- 3.2 (b) 30.85%, 30.85%, 19.74%.
- 3.3 (a) Normal with mean 0 and variance $2(1 - \rho)\sigma^2$.
(b) 0.5 independent of μ , σ , ρ .
(c) Approx. 46.06%, 42.17%.
- 3.7 Approx. 6.733%.
- 3.10 (a) Approx. 36.94%, -7.87% , -11.96% .
(c) Approx. -15.94% , -48.21% .
- 3.12 (a) 0.2611
(b) 0.6306
(c) 0.1140
(d) 61.05%
(e) 0.4886

Chapter 4

- 4.2 (a-c) Approx. 40.13%, -55.79%, -83.05%.
 (d) For $\rho = -0.8$: Expectation 10%, standard deviation 12.65%, probability 21.46% of negative returns, Value-at-Risk -10.81% and -19.43%.
- 4.3 (a) Approx. 36.94%, -78.69%.
 (b) For $\rho = 0$: 31.87%, -49.79%. For $\rho = 0.5$: 35.02%, -65.47%.
 (c) For $\rho = 0$: 14.59%, -11.21%. For $\rho = 0.5$: 32.65%, -53.19%.
- 4.8 (c) Largest lower bound is 0.2. You need 7 stocks or more.
- 4.9 (b) -38.80%.
 (d) For $\rho = 0.5$: -36.53%. For $\rho = 0.2$: -19.43%.
- 4.11 (a) 4.65%.
 (b) 5.25% and 22.91%.
 (c) 50% in each stock. Standard deviation 21.65%.
 (d) Stocks' Sharpe ratio: 0.40 and 0.24. Portfolio has weights 0.875 and 0.125 and Sharpe ratio 0.4027.
 (e) 56.36%.

Chapter 5

- 5.1 Yes. For example, short 2 units of the 3% bond, and buy one unit of the zero-coupon bond and one unit of the 6% bond.
- 5.2 (b) 10%, 8%.
 (c) 10% and approx. 6.036%.
 (d) Approx 89.27.
 (e) Buy the cheap bond, sell replicating portfolio consisting of 0.99586 units of the bullet bond and -0.08285 units of the serial bond.
- 5.3 (a) Approx. 995.02, 980.30, 942.32, 888.49, 821.93.
 (b) Approx. 1044.78, 1079.06, 1088.20, 1078.79, 1053.33.
 (c) Approx. 0.5000%, 0.9882%, 1.9448%, 2.8860%, 3.8085%.
 (d) Approx. 1044.78, 1062.34, 1071.38, 1073.37, 1068.96.
 (e) Approx. 0.5000%, 0.8318%, 1.4048%, 2.0173%, 2.6333%.
- 5.4 (a) Portfolio: -36 and 35 units of the bonds. Yield 4%.
 (b) Portfolio: approx. 2.6667 and -1.6667 units of the bonds. Yield 5%.
 (c) 4% and approx. 6.01%.
- 5.5 (a-c) 3%, 18.97%, -10.30%.
- 5.6 (a) One-year zero-coupon bond: 0.95238 units of 5% bullet, price 985.2190.
 Two-year zero-coupon bond: -0.03663 units of 5% bullet and 0.96154 units of 4% bullet, price 942.5973.
 (b) 1.5% and 3%.
 (c) 1.5% and 4.52%.
 (d) 1051.50 and 2.49%. Replicating portfolio: 0.49948 units of 5% bullet and 0.52446 units of 4% bullet.
- 5.7 (a) The 5% coupon bond.
 (b) 0.866%.
- 5.9 (a) 3,450,435.14.
 (b) 4-year bond: 1037.17 and 3.7797. 8-year bond: 1140.39 and 6.8751.
 (c) 2015.23 units of 4-year bond and 1192.83 units of 8-year bond.
 (d) Rebalancing is necessary.
- 5.10 (a) Approx. 444,498.18.
 (b) 3-year zero-coupon bonds.
 (c) 2-year bond: 103.77 and 1.9444. 4-year bond: 103.63 and 3.7287.
 (d) 1749.34 units of 2-year bond and 2537.54 units of 4-year bond.

- (e) Rebalancing is necessary.
- 5.11 (a) Price 1243.33, duration 7.7977, convexity 78.1112.
 (b) Ranging from 0.1655 to -0.1365.
 (c) Ranging from 0.1500 to -0.1500.
 (d) Ranging from 0.1644 to -0.1355.
- 5.12 (a) Liability: PV 4,439,856.11, duration 6. 3-year bond: price 1028.84, duration 2.9148.
 10-year bond: price 1269.48, duration 8.3502.
 (b) 1865.92 units of 3-year bond and 1985.17 units of 10-year bond.
 (c) Liability: 4,313,043.92. Portfolio: 4,315,458.40. Successful. Must rebalance.
 (d) Liability: 4,313,043.92. Portfolio: 4,079,855.60. Unsuccessful.
- 5.13 (a) Price 1019.8025, yield 0.9951%, duration 1.9806.
 (b) Price 1073.4278, yield 2.4230%, duration 4.6447.
 (c) 2-year bond: drop 5.65%, yield 4.0050%. 5-year bond: drop 0.68%, yield 2.5745%.
- 5.14 (a) Price 1000.23, yield 0.9884%, duration 1.9901.
 (b) Price 1068.94, yield 1.7784%, duration 5.5959.
 (c) PV 921,431.56. Portfolio: 152.24 units of 2-year bond and 719.55 units of 6-year bond.
 (d) Drop of 1.96% and 1.21%.
 (e) Liability: 905,730.81. Portfolio: 909,140.66. Successful.
- 5.16 (a) Price 1093.71, yield 1.0653%, duration 4.7314, convexity 27.8544.
 (b) PV 915,141.66, duration 3, convexity 12.
 (c) 575.86 units of 2-year bond and 310.11 units of 5-year bond.
 (d) Liability: 970,590.15. Portfolio: 886,875.06. Unsuccessful.
 (e) Units of bonds needed: 204.29, -336.86, 1003.52.
 (f) Portfolio: 946,981.53. Better, but still unsuccessful.
- 5.17 (a) \$931,126.93 and 4 years.
 (b) Prices \$1051.0060 and \$990.1694. Durations 1.9625 and 5.7118.
 (c) 404.4805 units of bond 1 and 511.0394 units of bond 2.
 (d) Almost successful: liability worth \$942,706.85, portfolio worth \$942,699.28.
 (e) Unsuccessful due to steepening of yield curve: liability worth \$898,658.47, portfolio worth \$886,103.48.
- 5.18 (a) Yields 1.4996%, 1.8944%, 2.1835%, 2.3999%. Durations 1.0000, 1.9712, 2.8892, 4.0000.
 (b) Units -0.0366, -0.0373, 0.9615 of the 1-year, 2-year, and 3-year bond, respectively. Price \$936.7970. Yield 2.2001%.
 (c) Price \$1040.9584. Arbitrage: purchase new bond and sell replicating portfolio, i.e. 0.2270, 0.2315, 0.2885, and 0.3 units of bond 1, 2, 3, and 4, respectively.
 (d) Expected: \$905.9559. Confidence interval: [\$895.7058, \$923.5608].

Chapter 6

- 6.1 (a) 10.
 (b) 5.30.
 (c) 53.
 (d) 3.
 (e) 17.67.
 (f) 53.
 (g) 11.33.
- 6.2 (a) 6, 8, and 10.
 (b) 127.50.
 (c) 115.4.
 (d) 10%.
- 6.3 (a) 26.25.
 (b) With 2% growth rate: 18.55. With 15% growth rate: ∞ .
 (c) With 8% discount rate: 70.00. With 18% discount rate: 16.15.

Chapter 7

- 7.1 (b) Min-var portfolio for $\rho = -0.5$: 59.46% in asset 1 and 40.54% in asset 2.
 (c) 52.63% in asset 1 and 47.37% in asset 2. Expected return 7.42% and standard deviation 24.66%.
 (d) 47.06% in asset 1 and 52.94% in asset 2. Expected return 7.59% and standard deviation 25.45%.
- 7.2 (a) No.
 (b) P1 cannot be on the frontier.
- 7.7 (a) Variance-covariance matrix is
- $$\underline{\Sigma} = \begin{pmatrix} 196 & 56 & 168 \\ 56 & 400 & 420 \\ 168 & 420 & 900 \end{pmatrix}.$$
- (b) With equal weights of 1/3: $E[r_p] \approx 13.33\%$ and $\sigma_p \approx 17.59\%$.
 (c) Minimum-variance portfolio is $(0.7472, 0.4458, -0.1931)^\top$.
 (d) Tangency portfolio is $(0.5272, 0.2048, 0.2681)^\top$.
 (e) Weights: 0.9068, 0.3523, 0.4611 in the stocks and -0.720 in the riskfree asset.
 (f) Weights: 0.1543, -0.2037, 1.0498.
 (g) 21.4% and 22.4%.
 (h) 10.05% and 18.09%.
- 7.10 (a) The minimum-variance portfolio is $(-0.3457, -0.0860, 0.7120, 0.5769, 0.1428)^\top$ and has $\mu_{\min} \approx 0.843\%$, $\sigma_{\min} \approx 2.072\%$.
 (b) The tangency portfolio is $(-0.4295, 0.1055, 0.9377, 0.5040, -0.1178)^\top$ and has $\mu_{\tan} \approx 1.324\%$, $\sigma_{\tan} \approx 2.602\%$.
 (d) The constrained tangency portfolio is $(0.0000, 0.2784, 0.1422, 0.5795, 0.0000)^\top$ with mean 1.131% and standard deviation 2.678%.
 (e) Without constraints, the portfolio is $(-0.9902, 0.2432, 2.1620, 1.1621, -0.2716)^\top$ with expected return 3.040%.
 With constraints, the portfolio is $(0.0000, 0.9852, 0.0000, 0.0148, 0.0000)^\top$ with expected return 1.833%.
- 7.11 (a) The minimum-variance portfolio is $(0.3110, -0.1411, 0.2739, -0.1787, 0.7349)^\top$ with expected return 1.121% and standard deviation 2.864%.
 (b) The tangency portfolio is $(0.4884, -0.4182, -0.1197, 0.1584, 0.8910)^\top$ with expected return 1.447% and standard deviation 3.261%.
 (d) With $\gamma = 10$, the optimal portfolio consists of -0.3424 in the riskfree asset (i.e., borrowing) and then the weights $(0.6557, -0.5613, -0.1606, 0.2126, 1.1961)^\top$ in the ETFs. The expected return is 1.936% and the standard deviation is 0.04377. With $\gamma = 20$, the optimal portfolio consists of 0.3288 in the riskfree asset (i.e., a deposit) and then the weights $(0.3278, -0.2807, -0.0803, 0.1063, 0.5980)^\top$ in the ETFs. The expected return is 0.978% and the standard deviation is 0.02188.
 (e) With $\gamma = 10$, the optimal portfolio consists of a zero position in the riskfree asset and the weights $(0.3530, 0, 0, 0, 0.6470)^\top$ in the ETFs. The expected return is 1.331% and the standard deviation is 0.03479. With $\gamma = 20$, the optimal portfolio consists of a weight of 0.4537 in the riskfree asset and $(0.1784, 0, 0.0019, 0, 0.3660)^\top$ in the ETFs. The expected return is 0.730% and the standard deviation is 0.01884.
- 7.13 (c) 3%, 9%, 15%.
- 7.15 (b) Indifferent for $\gamma = 3$.
 (c) 11%, 33%.

Chapter 8

- 8.1 (a) 50% in each.
 (b) Answer (ii).

- (c) 41.67% in domestic stock index, 8.33% in emerging markets fund, and still 50% riskfree.
 (d) For $\rho = 0.75$: 50% domestic index, 50% riskfree, nothing in emerging markets fund.
- 8.2 (a) 50% in each independent of horizon.
 (b) 37.5% stocks, 62.5% riskfree. Loss of 1.24% for $T = 10$ and 3.68% for $T = 30$.
 (c) 62.5% stocks, 37.5% riskfree. Same loss as in (b).

Chapter 9

- 9.1 (a) \$335,301 at age 60, \$445,527 at age 65, and \$581,552 at age 70.
 (c) \$24,672 at age 60, \$40,071 at age 65, and \$71,700 at age 70.
- 9.2 (a) 30% stocks, 70% riskfree.
 (e) 1% and \$1,614,663.
 (f) Borrow \$449,399 at the riskfree rate and invest all that plus the \$50,000 she has in financial wealth in stocks.
- 9.3 (a) For $T = 5$: 9.24%. For $T = 40$: 53.37%.
 (b) 20% stocks, 80% riskfree.
 (c) \$194,109 with 5 years and \$880,380 with 25 years to go.
 (d) With 5 years: borrow \$22,822 and invest \$42,822 in stocks. Weights of -114.1% and 214.1%, respectively.
- 9.4 (a) \$200,000 in stocks and \$200,000 in riskfree asset (i.e., 50% in each), independent of horizon.
 (b) Sell for \$10,083.37 of riskfree asset and invest that amount in stocks.
 (c) Human capital is \$422,325.08. Borrow \$11,162.54 and invest \$411,162.54 in stocks.
 (d) Human capital is \$888,146.16. Borrow \$244,073.08 and invest \$644,073.08 in stocks.
- 9.6 (a) Human capital is 900,707.75 with $r = 4\%$ and 457,058.68 with $r = 10\%$.
 (b) With 4%: \$324,664.63 in stocks, loan of \$224,664.63. With 10%: \$186,640.48 in stocks, loan of \$86,640.48.
 (c) \$223,968.57 in stocks, total financial wealth of \$135,595.29, human capital of \$461,964.55.
 (d) Sell stocks for \$19,981.70 and use the proceeds to reduce loan.

Chapter 10

- 10.1 (a) 10.84% and 0.7872.
 (b) 10.7232% so CAPM does not hold.
- 10.3 Consistent with CAPM if $r_f = -2\%$.
- 10.4 2/3 and 1/3.
- 10.5 (a) Weight 75% in stocks. Standard deviation 50%.
 (c) 2.
 (d) 15%.
- 10.6 (a) 1.5 and 60%.
 (b) 7.75%.
 (c) 40%.
- 10.7 (a) Beta 2, correlation 0.8.
 (b) 11%.
 (c) Can be anything from -6.5% to 8.5%.
- 10.8 Yes, no, yes.
- 10.9 (a) 6%.
 (b) 1.6.
 (c) 0.8.
 (d) 7.4%.
 (e) 80%.

Chapter 11

- 11.3 (a) Portfolio weights 1, -1.4, and 1.4. Expected return 7%.
 (b) For example, portfolio with weights 2, -2.8, and 1.8 has zero beta and 4% alpha.
- 11.4 (a) 2% and -1%.
 (b) Need $\pi_B = -1.4 \times \pi_A$.
 (c) Portfolio with weight 67.8% in A, -94.9% in B, and 127.1% in riskfree has standard deviation 5% and zero beta with an alpha of 2.305%.
 (d) Sharpe ratio 0.4612, information ratio 0.461.
 (e) Not correct.
- 11.6 (a) -2% and +2%.
 (b) Need $\pi_Y = -4 \times \pi_X$ with negative π_X to generate positive alpha.
 (c) Portfolio with weights -0.288675 in X, 1.15470 in Y, and 0.133975 in riskfree has alpha of 2.887% and expected return of 4.887%.
 (d) Correct.
- 11.7 (a) Correct for A and C, incorrect for B.
- 11.8 (a) 1.2191, not significantly different from 1.
 (b) -0.4667%, not significantly different from 0.
 (c) Total variance 87.53, systematic component 45.93, firm-specific component 41.60.
 (e) Inconsistent.
- 11.9 (a) 5% and 10%.
 (b) 11.5%. Portfolio: weights 0.2212 in A, 0.5457 in B, and 0.2332 in riskfree.
 (d) Consistent with CAMP if $E[r_m] = 9\%$.
 (e) Impossible.
- 11.10 (a) 1.25 and 11%.
 (b) 1.75 and 15%. Correlation 0.25.
 (d) Risk premium 7%, expected return on Generics now 16.6%.
- 11.11 (a) 0.4477, significantly different from one.
 (b) 0.4761, not statistically significant from zero.
 (c) Total variance of 26.3513, decomposed into a systematic risk of 6.1952 and a firm-specific risk of 20.1561.
 (d) Market beta is 0.5373, SMB beta is -0.4638, and HML beta is -0.1715.
 (e) 0.5305% and 0.6633%.
- 11.12 (a) 0.114, -0.026, 0.206, and 0.11.
 (b) Standard deviations: 0.1855, 0.4200, 0.2891, and 0.2720.
 (c) Portfolio $(0.7850, 0.3194, -0.0406, -0.0638)^\top$ with expected return 0.0658 and standard deviation of 0.1066.
 (d) Portfolio $(0.4699, -0.1534, 0.2230, 0.4604)^\top$ with expected return 0.1541 and standard deviation 0.1712.
 (f) Minimum-variance portfolio: 0.2772 and 0.3149. Tangency portfolio: 0.7505 and 0.1554. Note that the tangency portfolio has a much higher loading on the first factor that carries a high expected value relative to its standard deviation.
 (g) With expected return 0.05, portfolio has weights $(-0.0550, 0.6548, 0.0821, 0.5319)$ in the stocks and -0.2138 in the riskfree asset and a standard deviation of 0.4092. Not efficient.
- 11.13 (b) 0.025 and 7.
 (e) Betas are 1.5612 and 0.8292. Standard deviation 0.2397.
- 11.14 (a) 0.04 and 0.10.
 (b) Weights 2 and -1.
 (c) 0.03 and 0.6.
 (d) Weight 0.8333 in A, -1 in B, and 1.1667 in H. Not an arbitrage.
- 11.15 (a) 1.5 and -1.5%.
 (b) 0.225.
 (c) Need $\pi_I = -2\pi_H$. Alpha equals $6\% \times \pi_H$.

- (e) Weight 1 in Hypo, -2 in Illu, and 2 in riskfree. Standard deviation 0.8426.
(f) 2.7802%.
- 11.16 (a) For AA: standard deviation 0.6708, Sharpe ratio 0.2087.
(c) Portfolio: $(-0.0059, 0.2116, 0.3147, 0.2427, 0.2368)^\top$ with expected return 0.1055, standard deviation 0.3699, and factor betas 1.3179 and 1.4804.
(d) Portfolio: $(0.1221, 0.5225, 0.0185, 0.0613, 0.2757)^\top$ with expected return 0.1345, standard deviation 0.4223, and factor betas 1.9733 and 1.2916.
(f) With expected return of 10%: weights $(0.0919, 0.3075, -0.1004, 0.0604, 0.2719)^\top$ in stocks and 0.3688 in riskfree asset. Not efficient.
- 11.17 (a) Expected return 0.07, standard deviation 0.1942, Sharpe ratio 0.3090.
(b) Not an arbitrage, still firm-specific risk.
(c) With correlation 0.5, the Sharpe ratio is 0.27588. With correlation -0.5, the Sharpe ratio is 0.35793.
(d) The 5% Value at Risk is -24.9373% or -\$249,373. Correlation must be at most 0.3947.
- 11.18 (a) Note $\underline{\Sigma}_F = \begin{pmatrix} 0.04 & 0.005 \\ 0.005 & 0.01 \end{pmatrix}$. Then $\text{Var}[r_X] = \begin{pmatrix} 1.2 \\ -0.2 \end{pmatrix} \cdot \begin{pmatrix} 0.04 & 0.005 \\ 0.005 & 0.01 \end{pmatrix} \begin{pmatrix} 1.2 \\ -0.2 \end{pmatrix} + (0.2)^2 = 0.0956$ and $\text{Cov}[r_X, r_Y] = \begin{pmatrix} 1.2 \\ -0.2 \end{pmatrix} \cdot \begin{pmatrix} 0.04 & 0.005 \\ 0.005 & 0.01 \end{pmatrix} \begin{pmatrix} 0.9 \\ 1.4 \end{pmatrix} = 0.0479$. The other elements of $\underline{\Sigma}$ are calculated in the same way.
(b) Expected returns: 0.0840, 0.1300, 0.0740. Standard deviations: 0.3092, 0.5609, 0.3265.
(c) Portfolio weights 0.4113, 0.3908, 0.1978. Standard deviation 0.3051.
(d) Portfolio weights 0.4479, 0.2405, 0.3116. Expected return 0.0919, standard deviation 0.2649, Sharpe ratio 0.2716.
(e) Stock weights 0.4981, 0.2674, 0.3465, riskfree weight -0.1119. Standard deviation 0.2945.
(f) Weight 1.3333 on market portfolio and -0.3333 on riskfree. Standard deviation 0.2667.
(g) Weight $w_m = 0.4211$ on market, 0.5789 on riskfree, and 1.3684 on long-short HML. Standard deviation only 0.1777.
- 11.19 (a) Without constraints: invest \$210,390.63 in market-ETF and borrow \$160,390.63.
(b) Without constraints: \$41,640.63 in the market-ETF and \$8,359.37 in the riskfree asset.
(c) $E[r_{Hi}] = 0.085$, $\underline{\Sigma} = \begin{pmatrix} 0.0256 & 0.0384 \\ 0.0384 & 0.0976 \end{pmatrix}$
(d) Invest \$241.065,63 in the market-ETF, short for \$20,450.00 of the HiRisk-ETF, and borrow \$170,615,63.
(e) Invest \$23,383.62 in the market-ETF, \$26,616.38 in the HiRisk-ETF, and nothing in the riskfree asset.
- 11.20 (a) $E[r_i] = k_i$. Standard deviations: 0.2773, 0.2478, 0.2010, 0.2926.
(b) $RP_1 = 0.08$ and $RP_2 = 0.10$.
(d) Weights 0.0429, 0.5700, 0.2439, 0.1432. Exp. return 0.1245, std dev 0.2109, Sharpe ratio 0.5428, factor betas 0.5602 and 0.6964.
(f) Weights 0.4586, 0.0433, 0.5568, 0.0586. Exp. return 0.0904, std dev 0.2201, Sharpe ratio 0.3655, factor betas 1.0056 and 0.
(g) Straight lines, i.e. wedge generated by riskfree asset and portfolio from (f).
(h) 0.1186 and 0.0831.
- 11.21 (a) Sharpe ratios 0.1323, 0.1565, 0.0871.
Variance-covariance matrix: $\underline{\Sigma} = \begin{pmatrix} 19.8203 & 19.2407 & 22.2001 \\ 19.2407 & 19.4481 & 20.3703 \\ 22.2001 & 20.3703 & 27.5520 \end{pmatrix}$
(b) Weights -1.5356, 3.2156, -0.6801. Sharpe ratio 0.1902.
(c) 0.233%, 2.5019%, $-2.9595\%^2$, -0.2657.
(d) Weight 1 on market and "weight" 1.4535 on RmW, i.e. weight 1.4535 on robust and -1.4535 on weak portfolio. Hence, the market portfolio must have a weight of 1. Sharpe ratio 0.1876, lower than 0.1902 from (b).
(e) 0.9411%, 10.0135%, 0.0940.
(f) Zero in Giggle (due to zero alpha), so identical to (d).

Chapter 12

- 12.2 (b) 3.125% and 0.09766%.
 (c) 2500 and 78.

12.4 Event (c).

12.7 Not necessarily violating Efficient Market Hypothesis as it could reflect variations in investors' risk aversion or in total market risk.

Chapter 13

- 13.1 (a) 0.375.
 (b) For Hypothetics: expected return 0.22, variance 0.25, information ratio 0.16, Sharpe ratio 0.3369.
 (c) Weights 0.3970, 0.4652, and 0.1378. Alpha 0.0652, beta 1.4245, expected return 0.1707, standard deviation 0.3644, Sharpe ratio 0.4134.
 (d) Weight 0.4027 on active portfolio.

13.3 (a) For SuperDuper: Sharpe ratio 0.56, M-squared 0.1208, Treynor ratio 0.1167, alpha 0.02, information ratio 0.1589.

- 13.4 (a) AA has expected return 0.14 and variance 0.41. BB has beta 1 and residual standard deviation 0.3.
 (b) Weights 0.7059, -2.3529, 1.7647, and 0.8824. Sharpe ratio 0.3710.
 (c) Weight 0.5152 on active portfolio. Sharpe ratio 0.3742.
 (d) Only market portfolio and riskfree asset: weight 0.5 in each.

With mispriced assets: weight 0.1417 in active, 0.1333 in market, and 0.725 in riskfree.

- 13.5 (a) For GG: expected return 0.14 and standard deviation 0.6403.
 (b) Weights 0.9664, 2.1477, -1.5101, and -0.6040. Sharpe ratio 0.2562.
 (c) weight 0.2200 on active portfolio. Sharpe ratio 0.2931.
 (e) For $\gamma = 1$: Without mispriced assets, invest everything in market portfolio. With mispriced assets, invest with weights 0.2069 in active portfolio, 0.7337 in market portfolio, and 0.0594 in riskfree asset.

13.6 (a) For FineFunds: Sharpe ratio 0.3571, M-squared 0.0671, Treynor ratio 0.1111, alpha 0.046, information ratio 0.1916.

Chapter 14

- 14.1 (a) 51.25.
 (b) -8.2451.

14.2 0.6811 EUR/CAD.

- 14.3 (a) 1760.72.
 (b) Buy futures, short index and place proceeds in riskfree asset.

14.4 Forward price 1300.69 for delivery in three months.

14.5 12.8%.

- 14.6 (a) 35,152.97.
 (b) 29,552.76.
 (c) 5,600.21.
 (d) 0.5947%.

Chapter 15

- 15.1 (c) 3.3976.
 (d) Initially the replicating portfolio consists of -0.3131 stocks and a riskfree position of 34.7107.
 (e) 4.3143.
 (f) -0.2893 stocks and a riskfree position of 33.2466.

- (g) Exercise in down-state. Current option price 6.4279.
 (h) Exercise immediately. Current option value 10.
- 15.2 (a) 32.5479.
 (b) Also 32.5479.
 (c) 0.6048 stocks and loan of 257.77.
 (d) Portfolio value 44.5880, option price now 45.5468.
 (e) 19.8035. Buy put and stock, sell call, and borrow present value of exercise price at the riskfree rate.
- 15.4 (a) 101.26.
 (d) 17.6701.
 (e) 6.5521.
 (f) American call price is 17.6701, American put price is at least 6.5521.
- 15.6 (a) 7.3420.
 (b) 6.5976.
- 15.7 (a) Price 8.50. Initial replicating portfolio consists of -0.4048 stocks and a riskfree investment of 48.9794.
 (b) Price 10.26. Initial replicating portfolio consists of -0.5286 stocks and a riskfree investment of 63.1292.
- 15.9 (c) 8.04.
- 15.10 (a) 40.4040.
 (b) Short forward, buy stock, and borrow.
 (c) 6.7166.
 (d) 4.3366.
 (e) American call price is 6.7166, American put price is at least 4.3366.
- 15.11 (a) 9.874.
 (b) 17.706.
 (c) 18.281.
 (d) 9.874.
 (f) Call price 8.8605 for volatility of 0.3. Implied volatility 0.3255.
- 15.12 (a) 6.4700.
 (b) Price 6.7052, early exercise premium 0.2352.
 (c) Buy call, sell put, sell stock, invest in riskfree asset.
 (d) 112.8566.
- 15.13 (a) 14.41.
 (b) -0.47208 stocks and \$61.61815 in riskfree asset.
 (c) Option price now 19.43, portfolio value 19.23.
 (d) 13.87.
 (e) 14.42.
- 15.14 (a) Return is 44%, 2%, or -27.75% . Expected return is 12.36%.
 (b) $q \approx 0.4429$. Risk-neutral expected return of 1.0025% equals riskfree return.
 (c) Put price 5.46. Expected return is -47.94% .
 (d) Buy call, sell put, sell ETF, invest in riskfree asset.
- 15.15 (b) Put prices: 35.1347 for European, 35.8518 for American.
 (c) Black-Scholes price 32.2461; $\sigma \approx 0.39185$.
- 15.16 (a) Price 6.2273 for call and 4.2569 for put.
 (b) 10.2416.
 (c) Sell the put and buy the replicating portfolio, which means: buy the call option, sell the underlying stock, and invest the present value of X in the riskfree asset.

Bibliography

- Abell, J. D. and T. M. Krueger (1989). Macroeconomic Influences on Beta. *Journal of Economics and Business* 41(2), 185–193.
- Acharya, V. V. and L. H. Pedersen (2005). Asset Pricing with Liquidity Risk. *Journal of Financial Economics* 77(2), 375–410.
- Aghassi, M., C. Asness, C. Fattouche, and T. J. Moskowitz (2023). Fact, Fiction, and Factor Investing. *Journal of Portfolio Management* 49(2), 57–94.
- Alan, S. (2006). Entry Costs and Stock Market Participation over the Life Cycle. *Review of Economic Dynamics* 9(4), 588–611.
- Albuquerque, R. (2012). Skewness in Stock Returns: Reconciling the Evidence on Firm versus Aggregate Returns. *Review of Financial Studies* 25(5), 1630–1673.
- Alexander, G. J. (1993). Short Selling and Efficient Sets. *Journal of Finance* 48(4), 1497–1506.
- Alexander, G. J., A. Baptista, and S. Yan (2007). Mean-Variance Portfolio Selection with “at-risk” Constraints and Discrete Distributions. *Journal of Banking & Finance* 31(12), 3761–3781.
- Almeida, J. and R. M. Gaspar (2021). Accuracy of European Stock Target Prices. *Journal of Risk and Financial Management* 14(9), 443.
- Amihud, Y., H. Mendelson, and L. H. Pedersen (2005). Liquidity and Asset Prices. *Foundations and Trends in Finance* 1(4), 269–364.
- Anarkulova, A., S. Cederburg, and M. S. O’Doherty (2022). Stocks for the Long Run? Evidence from a Broad Sample of Developed Markets. *Journal of Financial Economics* 143(1), 409–433.
- Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2014). Discounting Behavior: A Reconsideration. *European Economic Review* 71, 15–33.
- Andersson, M., E. Krylova, and S. Vähämaa (2008). Why Does the Correlation between Stock and Bond Returns Vary over Time? *Applied Financial Economics* 18(2), 139–151.
- Andrews, D. and A. C. Sánchez (2011). The Evolution of Homeownership Rates in Selected OECD Countries: Demographic and Public Policy Influences. *OECD Journal: Economic Studies* 2011(1), 207–243.
- Ang, A. (2014). *Asset Management*. Oxford University Press.
- Ang, A. and G. Bekaert (2007). Stock Return Predictability: Is It There? *Review of Financial Studies* 20(3), 651–707.
- Ang, A. and J. Chen (2007). CAPM over the Long Run: 1926–2001. *Journal of Empirical Finance* 14(1), 1–40.
- Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The Cross-Section of Volatility and Expected Returns. *Journal of Finance* 61(1), 259–299.

- Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2009). High Idiosyncratic Volatility and Low Returns: International and Further US Evidence. *Journal of Financial Economics* 91(1), 1–23.
- Ardia, D., K. Bluteau, K. Boudt, and K. Inghelbrecht (2023). Climate Change Concerns and the Performance of Green vs. Brown Stocks. *Management Science* 69(12), 7607–7632.
- Arditti, F. D. and H. Levy (1975). Portfolio Efficiency Analysis in Three Moments: The Multi-period Case. *Journal of Finance* 30(3), 797–809.
- Arnott, R., C. R. Harvey, V. Kalesnik, and J. Linnainmaa (2019). Alice's Adventures in Factor-land: Three Blunders That Plague Factor Investing. *Journal of Portfolio Management* 45(4), 18–36.
- Arnott, R. D., C. R. Harvey, V. Kalesnik, and J. T. Linnainmaa (2021). Reports of Value's Death May Be Greatly Exaggerated. *Financial Analysts Journal* 77(1), 44–67.
- Asness, C., S. Chandra, A. Ilmanen, and R. Israel (2017). Contrarian Factor Timing is Deceptively Difficult. *Journal of Portfolio Management* 43(5), 72–87.
- Asness, C., A. Frazzini, R. Israel, T. J. Moskowitz, and L. H. Pedersen (2018). Size Matters, If You Control Your Junk. *Journal of Financial Economics* 129(3), 479–509.
- Asness, C. S., A. Frazzini, and L. H. Pedersen (2019). Quality Minus Junk. *Review of Accounting Studies* 24(1), 34–112.
- Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013). Value and Momentum Everywhere. *Journal of Finance* 68(3), 929–985.
- Attanasio, O. P., R. Bottazzi, H. W. Low, L. Nesheim, and M. Wakefield (2012). Modelling the Demand for Housing over the Life Cycle. *Review of Economic Dynamics* 15(1), 1–18.
- Attanasio, O. P. and G. Weber (1995). Is Consumption Growth Consistent with Intertemporal Optimization? Evidence from the Consumer Expenditure Survey. *Journal of Political Economy* 103(6), 1121–1157.
- Avramov, D., T. Chordia, G. Jostova, and A. Philipov (2007). Momentum and Credit Rating. *Journal of Finance* 62(5), 2503–2520.
- Azevedo, V. (2023). Analysts' Underreaction and Momentum Strategies. *Journal of Economic Dynamics and Control* 146, 104560.
- Ball, R., J. Gerakos, J. T. Linnainmaa, and V. Nikolaev (2016). Accruals, Cash Flows, and Operating Profitability in the Cross Section of Stock Returns. *Journal of Financial Economics* 121(1), 28–45.
- Banz, R. W. (1981). The Relationship between Return and Market Value of Common Stocks. *Journal of Financial Economics* 9(1), 3–18.
- Barbee, Jr., W. C., S. Mukherji, and G. A. Raines (1996). Do Sales-Price and Debt-Equity Explain Stock Returns Better than Book-Market and Firm Size? *Financial Analysts Journal* 52(2), 56–60.
- Barber, B., R. Lehavy, M. McNichols, and B. Trueman (2001). Can Investors Profit from the Prophets? Security Analyst Recommendations and Stock Returns. *Journal of Finance* 56(2), 531–563.
- Barber, B. M., E. T. D. George, R. Lehavy, and B. Trueman (2013). The Earnings Announcement Premium Around the Globe. *Journal of Financial Economics* 108(1), 118–138.
- Barber, B. M. and T. Odean (2001). Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment. *Journal of Political Economy* 116(1), 261–292.
- Barber, B. M. and T. Odean (2013). The Behavior of Individual Investors. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B, Chapter 22, pp. 1533–1570. Elsevier.
- Barber, J. R. (1995). A Note on Approximating Bond Price Sensitivity using Duration and Convexity. *Journal of Fixed Income* 4(4), 95–98.
- Barberis, N. (2000). Investing for the Long Run when Returns are Predictable. *Journal of Finance* 55(1), 225–264.

- Barberis, N., A. Schleifer, and R. Vishny (1998). A Model of Investor Sentiment. *Journal of Financial Economics* 49, 307–343.
- Barberis, N. and R. Thaler (2003). A Survey of Behavioral Finance. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 1B, Chapter 18. Elsevier.
- Basu, S. (1983). The Relationship between Earnings' Yield, Market Value and Return for NYSE Common Stocks: Further Evidence. *Journal of Financial Economics* 12(1), 129–156.
- Batten, J. A., T. A. Fetherston, and P. G. Szilagyi (Eds.) (2004). *European Fixed Income Markets*. Wiley.
- Beber, A. and M. Pagano (2013). Short-Selling Bans Around the World: Evidence from the 2007–09 Crisis. *Journal of Finance* 68(1), 343–381.
- van der Beck, P. (2021, September). Flow-driven ESG returns. Available at SSRN: <http://ssrn.com/abstract=3929359>.
- Benartzi, S. and R. H. Thaler (2007). Heuristics and Biases in Retirement Savings Behavior. *Journal of Economic Perspectives* 21(3), 81–104.
- Berg, F., J. F. Kölbel, and R. Rigobon (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance* 26(6), 1315–1344.
- Berk, J., R. Green, and V. Naik (1999). Optimal Investment, Growth Options, and Security Returns. *Journal of Finance* 54(5), 1553–1608.
- Berk, J. and J. H. van Binsbergen (2024, January). The Impact of Impact Investing. Available at SSRN: <http://ssrn.com/abstract=3909166>.
- Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica* 22(1), 23–36. Translation of the 1738 version.
- Bessembinder, H. (2018). Do Stocks Outperform Treasury Bills? *Journal of Financial Economics* 129(3), 440–457.
- Bessembinder, H., T.-F. Chen, G. Choi, and K. C. J. Wei (2019, July). Do Global Stocks Outperform US Treasury Bills? Available at SSRN: <http://ssrn.com/abstract=3415739>.
- Best, M. J. and R. R. Grauer (1991). On the Sensitivity of Mean-Variance-Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results. *Review of Financial Studies* 4(2), 315–342.
- Bilinski, P., D. Lyssimachou, and M. Walker (2013). Target Price Accuracy: International Evidence. *The Accounting Review* 88(3), 825–851.
- van Binsbergen, J. H., M. W. Brandt, and R. S. J. Koijen (2012). On the Timing and Pricing of Dividends. *American Economic Review* 102(4), 1596–1618.
- van Binsbergen, J. H., W. Hueskes, R. S. J. Koijen, and E. Vrugt (2013). Equity Yields. *Journal of Financial Economics* 110(3), 503–519.
- van Binsbergen, J. H. and R. S. J. Koijen (2010). Predictive Regressions: A Present-Value Approach. *Journal of Finance* 65(4), 1439–1471.
- van Binsbergen, J. H. and R. S. J. Koijen (2017). The Term Structure of Returns: Facts and Theory. *Journal of Financial Economics* 124(1), 1–21.
- Black, F. (1972). Capital Market Equilibrium with Restricted Borrowing. *Journal of Business* 45(3), 444–454.
- Black, F. (1976). Studies of Stock Price Volatility Changes. In *Proceedings of the 1976 Meetings of the Business and Economics Statistics Section, American Statistical Association*, pp. 177–181.
- Black, F., M. C. Jensen, and M. Scholes (1972). The Capital Asset Pricing Model: Some Empirical Tests. In M. C. Jensen (Ed.), *Studies in the Theory of Capital Markets*, pp. 79–121. Praeger, New York.
- Black, F. and P. Karasinski (1992). Global Portfolio Optimization. *Financial Analysts Journal* 48(5), 28–43.

- Black, F. and M. Scholes (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81(3), 637–654.
- Blitz, D. (2022). Factor Investing: The Best Is Yet to Come. *Journal of Portfolio Management* 49(2), 10–18.
- Blitz, D. and M. X. Hanauer (2021). Settling the Size Matter. *Journal of Portfolio Management* 47(2), 99–112.
- Blitz, D., M. X. Hanauer, and M. Vidojevic (2020). The Idiosyncratic Momentum Anomaly. *International Review of Economics and Finance* 69, 932–957.
- Blitz, D., J. Huij, and M. Martens (2011). Residual Momentum. *Journal of Empirical Finance* 18(3), 506–521.
- Bloom, N. (2009). The Impact of Uncertainty Shocks. *Econometrica* 77(3), 623–685.
- Blume, M. E. (1971). On the Assessment of Risk. *Journal of Finance* 26(3), 1–10.
- Bodie, Z., A. Kane, and A. J. Marcus (2021). *Investments* (12th ed.). McGraw-Hill Higher Education.
- Bodie, Z., R. C. Merton, and W. F. Samuelson (1992). Labor Supply Flexibility and Portfolio Choice in a Life Cycle Model. *Journal of Economic Dynamics and Control* 16(3-4), 427–449.
- Boudoukh, J., R. Michaely, M. Richardson, and M. R. Roberts (2007). On the Importance of Measuring Payout Yield: Implications for Empirical Asset Pricing. *Journal of Finance* 62(2), 877–916.
- Boudoukh, J., M. Richardson, and R. F. Whitelaw (2008). The Myth of Long-Horizon Predictability. *Review of Financial Studies* 21(4), 1577–1605.
- Bourassa, S. C., D. R. Haurin, J. L. Haurin, M. Hoesli, and J. Sun (2009). House Price Changes and Idiosyncratic Risk: The Impact of Property Characteristics. *Real Estate Economics* 37(2), 259 – 278.
- Branger, N., B. Breuer, and C. Schlag (2010). Discrete-time Implementation of Continuous-Time Portfolio Strategies. *European Journal of Finance* 16(2), 137–152.
- Branger, N., L. S. Larsen, and C. Munk (2019). Hedging Recessions. *Journal of Economic Dynamics and Control* 107, 103715.
- Brealey, R. A., S. C. Myers, and F. Allen (2009). *Principles of Corporate Finance* (First concise ed.). McGraw-Hill Higher Education.
- Breeden, D. T. (1979). An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities. *Journal of Financial Economics* 7(3), 265–296.
- Brennan, M. J., T. Chordia, A. Subrahmanyam, and Q. Tong (2012). Sell-Order Liquidity and the Cross-Section of Expected Stock Returns. *Journal of Financial Economics* 105(3), 523–541.
- Brennan, M. J. and Y. Xia (2000). Stochastic Interest Rates and the Bond-Stock Mix. *European Finance Review* 4(2), 197–210.
- Bricker, J., A. B. Kennickell, K. B. Moore, and J. Sabelhaus (2012). Changes in U.S. Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances. *Federal Reserve Bulletin* 98(2), 1–80.
- Brigo, D. and F. Mercurio (2006). *Interest Rate Models – Theory and Practice* (Second ed.). Springer-Verlag.
- Brinson, G. P., L. R. Hood, and G. L. Beebower (1986). Determinants of Portfolio Performance. *Financial Analysts Journal* 42(4), 39–44.
- Brinson, G. P., B. D. Singer, and G. L. Beebower (1991). Determinants of Portfolio Performance II: An Update. *Financial Analysts Journal* 47(3), 40–48.
- Brøgger, A. and A. Kronies (2022, November). Skills and Sentiment in Sustainable Investing. Available at SSRN: <http://ssrn.com/abstract=3531312>.
- Brown, S., W. Goetzmann, and S. A. Ross (1995). Survival. *Journal of Finance* 50(3), 853–873.
- Browning, M. and M. D. Collado (2007). Habits and Heterogeneity in Demands: A Panel Data Analysis. *Journal of Applied Econometrics* 22(3), 625–640.

- Browning, M. and T. Crossley (2001). The Life-Cycle Model of Consumption and Saving. *Journal of Economic Perspectives* 15, 3–22.
- Brunnermeier, M. K. and L. H. Pedersen (2009). Market Liquidity and Funding Liquidity. *Review of Financial Studies* 22(6), 2201–2238.
- Campbell, J. Y. (1996). Understanding Risk and Return. *Journal of Political Economy* 104(2), 298–345.
- Campbell, J. Y. (2000). Asset Pricing at the Millennium. *Journal of Finance* 55(4), 1515–1567.
- Campbell, J. Y. (2003). Consumption-Based Asset Pricing. In G. M. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 1B, Chapter 13, pp. 803–887. Elsevier.
- Campbell, J. Y. (2006). Household Finance. *Journal of Finance* 61(4), 1553–1604.
- Campbell, J. Y. and J. Ammer (1993). What Moves the Stock and Bond Markets? A Variance Decomposition for Long-Term Asset Returns. *Journal of Finance* 48(1), 3–37.
- Campbell, J. Y. and J. F. Cocco (2003). Household Risk Management and Optimal Mortgage Choice. *Quarterly Journal of Economics* 118(4), 1449–1494.
- Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- Campbell, J. Y. and R. J. Shiller (1988a). The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors. *Review of Financial Studies* 1(3), 195–227.
- Campbell, J. Y. and R. J. Shiller (1988b). Stock Prices, Earnings, and Expected Dividends. *Journal of Finance* 43(3), 661–676.
- Campbell, J. Y. and R. J. Shiller (1991). Yield Spreads and Interest Rate Movements: A Bird's Eye View. *Review of Economic Studies* 58(3), 495–514.
- Campbell, J. Y. and S. B. Thompson (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies* 21(4), 1509–1531.
- Campbell, J. Y. and L. M. Viceira (1999). Consumption and Portfolio Decisions when Expected Returns are Time Varying. *Quarterly Journal of Economics* 114(2), 433–495.
- Campbell, J. Y. and L. M. Viceira (2001). Who Should Buy Long-Term Bonds? *American Economic Review* 91(1), 99–127.
- Campbell, J. Y. and L. M. Viceira (2002). *Strategic Asset Allocation*. New York: Oxford University Press.
- Cappiello, L., R. F. Engle, and K. Sheppard (2006). Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Journal of Financial Econometrics* 4(4), 537–572.
- Carhart, M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance* 52(1), 57–82.
- Carlson, M., A. Fisher, and R. Giammarino (2004). Corporate Investment and Asset Price Dynamics: Implications for the Cross-Section of Returns. *Journal of Finance* 59(6), 2577–2603.
- Carrasco, R., J. M. Labeaga, and J. D. Lopez-Salido (2005). Consumption and Habits: Evidence from Panel Data. *The Economic Journal* 115(500), 144–165.
- Case, K. E. and R. J. Shiller (1989). The Efficiency of the Market for Single-Family Homes. *American Economic Review* 79(1), 125–137.
- Cerutti, E. M., M. Obstfeld, and H. Zhou (2021). Covered Interest Parity Deviations: Macroeconomic Determinants. *Journal of International Economics* 130, 103447.
- Chan, K. C. and N.-F. Chen (1991). Structural and Return Characteristics of Small and Large Firms. *Journal of Finance* 46(4), 1467–1484.
- Chan, K. C., G. A. Karolyi, F. A. Longstaff, and A. Sanders (1992). An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *Journal of Finance* 47(3), 1209–1227.
- Chan, Y. L. and L. Kogan (2002). Catching Up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices. *Journal of Political Economy* 110(6), 1255–1285.

- Chen, L. (2009). On the Reversal of Return and Dividend Growth Predictability: A Tale of Two Periods. *Journal of Financial Economics* 92(1), 128–151.
- Chen, N.-F. (1991). Financial Investment Opportunities and the Macroeconomy. *Journal of Finance* 46(2), 529–554.
- Chen, N.-f. and J. E. Ingersoll, Jr. (1983). Exact Pricing in Linear Factor Models with Finitely Many Assets: A Note. *Journal of Finance* 38(3), 985–988.
- Chen, N.-f., R. Roll, and S. A. Ross (1986). Economic Forces and the Stock Market. *Journal of Business* 59(3), 383–403.
- Choi, H.-S., H. G. Hong, J. D. Kubik, and J. P. Thompson (2016, April). Sand States and the US Housing Crisis. Available at SSRN: <http://ssrn.com/abstract=2373179>.
- Chopra, V. K. and W. T. Ziemba (1993). The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice. *Journal of Portfolio Management* 19(2), 6–11.
- Christensen, P. O., K. Larsen, and C. Munk (2012). Equilibrium in Securities Markets with Heterogeneous Investors and Unspanned Income Risk. *Journal of Economic Theory* 147(3), 1035–1063.
- Cicchetti, C. J. and J. A. Dubin (1994). A Microeconomic Analysis of Risk Aversion and the Decision to Self-Insure. *Journal of Political Economy* 102(1), 169–186.
- Cocco, J. F. (2005). Portfolio Choice in the Presence of Housing. *Review of Financial Studies* 18(2), 535–567.
- Cocco, J. F., F. J. Gomes, and P. J. Maenhout (2005). Consumption and Portfolio Choice over the Life Cycle. *Review of Financial Studies* 18(2), 491–533.
- Cochrane, J. H. (2005). *Asset Pricing* (Revised ed.). Princeton University Press.
- Cochrane, J. H. (2008). The Dog That Did Not Bark: A Defense of Return Predictability. *Review of Financial Studies* 21(4), 1533–1575.
- Cochrane, J. H. (2011). Discount Rates. *Journal of Finance* 66(4), 1047–1108.
- Cochrane, J. H. and M. Piazzesi (2005). Bond Risk Premia. *American Economic Review* 95(1), 138–160.
- Connor, G. (1984). A Unified Beta Pricing Theory. *Journal of Economic Theory* 34(1), 13–31.
- Constantinides, G. M. (1979). Multiperiod Consumption and Investment Behavior with Convex Transactions Costs. *Management Science* 25(11), 1127–1137.
- Constantinides, G. M. (1986). Capital Market Equilibrium with Transaction Costs. *Journal of Political Economy* 94(4), 842–862.
- Constantinides, G. M. (2002). Rational Asset Prices. *Journal of Finance* 57(4), 1567–1591.
- Cooper, I. (2006). Asset Pricing Implications of Nonconvex Adjustment Costs and Irreversibility of Investment. *Journal of Finance* 61(1), 139–170.
- Cooper, I. and R. Priestley (2009). Time-Varying Risk Premiums and the Output Gap. *Review of Financial Studies* 22(7), 2801–2833.
- Cooper, M. J., O. Dimitrov, and P. R. Rau (2001). A Rose.com by Any Other Name. *Journal of Finance* 56(6), 2371–2388.
- Corradin, S., J. L. Fillat, and C. Vergara-Alert (2014). Optimal Portfolio Choice with Predictability in House Prices and Transaction Costs. *Review of Financial Studies* 27(4), 823–880.
- Cotter, J. and R. Roll (2015). A Comparative Anatomy of Residential REITs and Private Real Estate Markets: Returns, Risks and Distributional Characteristics. *Real Estate Economics* 43(1), 209–240.
- Coval, J. D. and T. Shumway (2001). Expected Option Returns. *Journal of Finance* 56(3), 983–1009.
- Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1981a). A Re-examination of Traditional Hypotheses about the Term Structure of Interest Rates. *Journal of Finance* 36(4), 769–799.

- Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1981b). The Relation between Forward Prices and Futures Prices. *Journal of Financial Economics* 9(4), 321–346.
- Cox, J. C., S. A. Ross, and M. Rubinstein (1979). Option Pricing: A Simplified Approach. *Journal of Financial Economics* 7(3), 229–263.
- Cremers, M., J. A. Fulkerson, and T. B. Riley (2019). Challenging the Conventional Wisdom on Active Management: A Review of the Past 20 Years of Academic Literature on Actively Managed Mutual Funds. *Financial Analysts Journal* 75(4), 8–35.
- Culbertson, J. M. (1957). The Term Structure of Interest Rates. *Quarterly Journal of Economics* 71(4), 489–504.
- Da, Z. and E. Schaumburg (2011). Relative Valuation and Analyst Target Price Forecasts. *Journal of Financial Markets* 14(1), 161–192.
- Dahlquist, M., J. V. Martinez, and P. Söderlind (2017). Individual Investor Activity and Performance. *Review of Financial Studies* 30(3), 866–899.
- Dahlquist, M., O. Setty, and R. Vestman (2018). On the Asset Allocation of a Default Pension Fund. *Journal of Finance* 73(4), 1893–1936.
- Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1998). Investor Psychology and Security Market Under- and Overreactions. *Journal of Finance* 53(6), 1839–1885.
- Daniel, K. D. and T. J. Moskowitz (2016). Momentum Crashes. *Journal of Financial Economics* 122(2), 221–247.
- Davis, M. H. A. and A. R. Norman (1990). Portfolio Selection with Transaction Costs. *Mathematics of Operations Research* 15(4), 676–713.
- Davis, S. J. and P. Willen (2000, March). Using Financial Assets to Hedge Labor Income Risks: Estimating the Benefits. Working paper, University of Chicago and Princeton University.
- DeBondt, W. F. M. and R. Thaler (1985). Does the Stock Market Overreact? *Journal of Finance* 40, 793–805.
- Dechow, P. M., R. D. Erhard, R. G. Sloan, and M. T. Soliman (2021). Implied Equity Duration: A Measure of Pandemic Shutdown Risk. *Journal of Accounting Research* 59(1), 243–281.
- Dechow, P. M., R. G. Sloan, and M. T. Soliman (2004). Implied Equity Duration: A New Measure of Equity Risk. *Review of Accounting Studies* 9(2/3), 197–228.
- Dechow, P. M. and H. You (2020). Understanding the Determinants of Analyst Target Price Implied Returns. *The Accounting Review* 95(6), 125–149.
- Deelstra, G., M. Grasselli, and P.-F. Koehl (2000). Optimal Investment Strategies in a CIR Framework. *Journal of Applied Probability* 37, 936–946.
- Detemple, J. B. and F. Zapatero (1992). Optimal Consumption-Portfolio Policies with Habit Formation. *Mathematical Finance* 2(4), 251–274.
- Dimson, E., P. Marsh, and M. Staunton (2002). *Triumph of the Optimists: 101 Years of Global Investment Returns*. Princeton, NJ: Princeton University Press.
- Dimson, E., P. Marsh, and M. Staunton (2019). *Credit Suisse Global Investment Returns Yearbook 2019*. Credit Suisse Research Institute.
- Du, W., A. Tepper, and A. Verdelhan (2018). Deviations from Covered Interest Rate Parity. *Journal of Finance* 73(3), 915–957.
- Dybvig, P. H. (1983). An Explicit Bound on Individual Assets' Deviations from APT Pricing in a Finite Economy. *Journal of Financial Economics* 12(4), 483–496.
- Eaton, G. W. and B. S. Paye (2017). Payout Yields and Stock Return Predictability: How Important Is the Measure of Cash Flow? *Journal of Financial & Quantitative Analysis* 52(4), 1639–1666.
- Edmans, A., X. Gabaix, and D. Jenter (2017). Executive Compensation: A Survey of Theory and Evidence. In B. E. Hermalin and M. S. Weisbach (Eds.), *The Handbook of the Economics of Corporate Governance*, Chapter 7, pp. 383–539. North-Holland.

- Elton, E. J. and M. J. Gruber (2013). Mutual Funds. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B, Chapter 15, pp. 1011–1061. Elsevier.
- Elton, E. J., M. J. Gruber, and M. D. Padberg (1976). Simple Criteria for Optimal Portfolio Selection. *Journal of Finance* 31(5), 1341–1357.
- Emery, K., S. Ou, J. Tennant, A. Matos, and R. Cantor (2009, February). Corporate Default and Recovery Rates, 1920–2008. Special comment, Moody's Investors Service.
- Engsted, T. and T. Q. Pedersen (2010). The Dividend-Price Ratio Does Predict Dividend Growth: International Evidence. *Journal of Empirical Finance* 17(4), 585–605.
- Estrella, A. and G. A. Hardouvelis (1991). The Term Structure as a Predictor of Real Economic Activity. *Journal of Finance* 46(2), 555–576.
- Fabozzi, F. J. (2010). *Bond Markets, Analysis and Strategies* (Seventh ed.). Pearson Prentice-Hall.
- Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25(2), 383–417.
- Fama, E. F. and R. R. Bliss (1987). The Information in Long-Maturity Forward Rates. *American Economic Review* 77(4), 680–692.
- Fama, E. F. and K. R. French (1988). Permanent and Temporary Components of Stock Prices. *Journal of Political Economy* 96(2), 246–273.
- Fama, E. F. and K. R. French (1989). Business Conditions and Expected Returns on Stocks and Bonds. *Journal of Financial Economics* 25(1), 23–49.
- Fama, E. F. and K. R. French (1992). The Cross-Section of Expected Stock Returns. *Journal of Finance* 47(2), 427–465.
- Fama, E. F. and K. R. French (1996). The CAPM is Wanted, Dead or Alive. *Journal of Finance* 51(5), 1947–1958.
- Fama, E. F. and K. R. French (2006). The Value Premium and the CAPM. *Journal of Finance* 61(5), 2163–2185.
- Fama, E. F. and K. R. French (2012). Size, Value, and Momentum in International Stock Returns. *Journal of Financial Economics* 105(3), 457–472.
- Fama, E. F. and K. R. French (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116(1), 1–22.
- Fama, E. F. and K. R. French (2016). Dissecting Anomalies with a Five-Factor Model. *Review of Financial Studies* 29(1), 69–103.
- Fama, E. F. and J. D. MacBeth (1973). Risk, Return, and Equilibrium: Some Empirical Tests. *Journal of Political Economy* 81(3), 607–636.
- Fang, J., B. R. Marshall, N. H. Nguyen, and N. Visaltanachoti (2021). Do Stocks Outperform Treasury Bills in International Markets? *Finance Research Letters* 40, 101710.
- Farago, A. and E. Hjalmarsson (2023). Long-Horizon Stock Returns are Positively Skewed. *Review of Finance* 27(2), 495–538.
- Ferson, W. E. (2013). Investment Performance: A Review and Synthesis. In G. M. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B, Chapter 14, pp. 969–1010. Elsevier.
- Fischer, M. and M. Stamos (2013). Optimal Life Cycle Portfolio Choice with Housing Market Cycles. *Review of Financial Studies* 26(9), 2311–2352.
- Fisher, I. (1896). Appreciation and Interest. *Publications of the American Economic Association*, 23–29 and 88–92.
- Fisher, L. and R. L. Weil (1971). Coping with the Risk of Interest Rate Fluctuations: Returns to Bondholders from Naive and Optimal Strategies. *Journal of Business* 44(4), 408–431.
- Flavin, M. and T. Yamashita (2002). Owner-Occupied Housing and the Composition of the Household Portfolio. *American Economic Review* 91(1), 345–362.

- Flor, C. R., H. Frimor, and C. Munk (2014). Options in Compensation: Promises and Pitfalls. *Journal of Accounting Research* 52(3), 703–732.
- Frankfurter, G. M., H. E. Phillips, and J. P. Seagle (1971). Portfolio Selection: The Effects of Uncertain Means, Variances, and Covariances. *Journal of Financial and Quantitative Analysis* 6(5), 1251–1262.
- Frazzini, A. (2006). The Disposition Effect and Underreaction to News. *Journal of Finance* 61(4), 2017–2046.
- Frazzini, A. and L. H. Pedersen (2014). Betting Against Beta. *Journal of Financial Economics* 111(1), 1–25.
- Fung, W. and D. A. Hsieh (2013). Hedge Funds. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B, Chapter 16, pp. 1063–1125. Elsevier.
- Garlappi, L., R. Uppal, and T. Wang (2007). Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach. *Review of Financial Studies* 20(1), 41–81.
- Garlappi, L. and H. Yan (2011). Financial Distress and the Cross-Section of Equity Returns. *Journal of Finance* 66(3), 789–822.
- Garleanu, N. and L. H. Pedersen (2013). Dynamic Trading with Predictable Returns and Transaction Costs. *Journal of Finance* 68(6), 2309–2340.
- Garleanu, N. and L. H. Pedersen (2018). Efficiently Inefficient Markets for Assets and Asset Management. *Journal of Finance* 73(4), 1663–1712.
- Garman, M. B. and S. W. Kohlhagen (1983). Foreign Currency Option Values. *Journal of International Money and Finance* 2(3), 231–237.
- Gibbons, M. R., S. A. Ross, and J. Shanken (1989). A Test of the Efficiency of a Given Portfolio. *Econometrica* 57, 1121–1152.
- Goetzmann, W. and M. Spiegel (2002). Policy Implications of Portfolio Choice in Underserved Mortgage Markets. In N. P. Retsinas and E. S. Belsky (Eds.), *Low-Income Homeownership: Examining the Unexamined Goal*, Chapter 9, pp. 257–266. Brookings Institution Press and Joint Center for Housing Studies at Harvard University.
- Gomes, J., L. Kogan, and L. Zhang (2003). Equilibrium Cross-Section of Returns. *Journal of Political Economy* 111(4), 693–732.
- Gomes, J. F., A. Yaron, and L. Zhang (2006). Asset Pricing Implications of Firms' Financing Constraints. *Review of Financial Studies* 19(4), 1321–1356.
- Gordon, M. (1962). *The Investment Financing and Valuation of the Corporation*. Richard D. Irwin.
- Gormsen, N. J. (2021). Time Variation of the Equity Term Structure. *Journal of Finance* 76(4), 1959–1999.
- Gormsen, N. J. and E. Lazarus (2023). Duration-Driven Returns. *Journal of Finance* 78(3), 1393–1447.
- Gourinchas, P.-O. and J. A. Parker (2002). Consumption Over the Life Cycle. *Econometrica* 70(1), 47–89.
- Goyal, A. (2012). Empirical Cross-Sectional Asset Pricing: A Survey. *Financial Markets and Portfolio Management* 26(1), 3–38.
- Goyal, A. and N. Jegadeesh (2018). Cross-Sectional and Time-Series Tests of Return Predictability: What Is the Difference? *Review of Financial Studies* 31(5), 1784–1824.
- Goyal, A. and P. Santa-Clara (2003). Idiosyncratic Risk Matters! *Journal of Finance* 58(3), 975–1007.
- Goyal, A. and I. Welch (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *Review of Financial Studies* 21(4), 1455–1508.
- Goyal, A., I. Welch, and A. Zafirov (2021). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction II. Research paper no. 21-85, Swiss Finance Institute. Available at SSRN: <http://ssrn.com/abstract=3929119>.

- Grinblatt, M. and B. Han (2005). Prospect Theory, Mental Accounting, and Momentum. *Journal of Financial Economics* 78(2), 311–339.
- Grinblatt, M. and S. Titman (1983). Factor Pricing in a Finite Economy. *Journal of Financial Economics* 12(4), 497–507.
- Grinblatt, M. and S. Titman (1985). Approximate Factor Structures: Interpretations and Implications for Empirical Tests. *Journal of Finance* 40(5), 1367–1373.
- Grinblatt, M. and S. Titman (1987). The Relation Between Mean-Variance Efficiency and Arbitrage Pricing. *Journal of Business* 60(1), 97–112.
- Grossman, S. J. and J. E. Stiglitz (1980). On the Impossibility of Informationally Efficient Markets. *American Economic Review* 70(3), 393–408.
- Guiso, L. and P. Sodini (2013). Household Finance: An Emerging Field. In G. M. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 2B, Chapter 21, pp. 1397–1532. Elsevier.
- Gutierrez, R. C. and C. A. Prinsky (2007). Momentum, Reversal, and the Trading Behaviors of Institutions. *Journal of Financial Markets* 10(1), 48–75.
- Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2021). What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics? *Econometrica* 89(5), 2303–2339.
- Hafner, W. and H. Zimmermann (Eds.) (2009). *Vinzenz Bronzin's Option Pricing Models: Exposition and Appraisal*. Springer-Verlag.
- Hakansson, N. H. (1970). Optimal Investment and Consumption Strategies Under Risk for a Class of Utility Functions. *Econometrica* 38(5), 587–607.
- Hall, B. J. and K. J. Murphy (2002). Stock Options for Undiversified Executives. *Journal of Accounting and Economics* 33, 3–42.
- Hartzmark, S. M. and D. H. Solomon (2021). Reconsidering Returns. *Review of Financial Studies* 35(1), 343–393.
- Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the Cross-Section of Expected Returns. *Review of Financial Studies* 29(1), 5–68.
- Hasbrouck, J. (2009). Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data. *Journal of Finance* 64(3), 1445–1477.
- Haugen, R. A. and A. J. Heins (1975). Risk and the Rate of Return on Financial Assets: Some Old Wine in New Bottles. *Journal of Financial and Quantitative Analysis* 10(5), 775–784.
- He, G. and R. Litterman (1999). The Intuition behind the Black-Litterman Model Portfolios. Note, Goldman Sachs Asset Management, New York.
- Heaton, J. and D. Lucas (1997). Market Frictions, Savings Behavior, and Portfolio Choice. *Macroeconomic Dynamics* 1(1), 76–101.
- Heaton, J. and D. Lucas (2000). Portfolio Choice and Asset Prices: The Importance of Entrepreneurial Risk. *Journal of Finance* 55(3), 1163–1198.
- van Hemert, O. (2010). Household Interest Rate Risk Management. *Real Estate Economics* 38(3), 467–505.
- Henderson, V. (2005). Explicit Solutions to an Optimal Portfolio Choice Problem with Stochastic Income. *Journal of Economic Dynamics and Control* 29(7), 1237–1266.
- Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies* 6(2), 327–343.
- Hicks, J. R. (1939). *Value and Capital*. Oxford: Clarendon Press.
- Hillier, D., M. Grinblatt, and S. Titman (2024). *Financial Markets and Corporate Strategy* (3nd European ed.). McGraw-Hill Higher Education.
- Hirshleifer, D. (2001). Investor Psychology and Asset Pricing. *Journal of Finance* 56(4), 1533–1597.
- Ho, T. S. Y. (1992). Key Rate Durations: A Measure of Interest Rate Risks. *Journal of Fixed Income* 2(2), 29–44.

- Hoepner, A. G. F., I. Oikonomou, Z. Sautner, L. T. Starks, and X. Y. Zhou (2024). ESG Shareholder Engagement and Downside Risk. *Review of Finance* 28(2), 483–510.
- Hoesli, M. and E. Oikarinen (2012). Are REITs Real Estate? Evidence from International Sector Level Data. *Journal of International Money and Finance* 31(7), 1823–1850.
- Hollstein, F. and M. Prokopczuk (2016). Estimating Beta. *Journal of Financial and Quantitative Analysis* 51(4), 1437–1466.
- Hong, H. and J. C. Stein (1999). A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *Journal of Finance* 54(6), 2143–2184.
- Hou, K., C. Xue, and L. Zhang (2020). Replicating Anomalies. *Review of Financial Studies* 33(5), 2019–2133.
- Huberman, G. (1982). A Simple Approach to Arbitrage Pricing Theory. *Journal of Economic Theory* 28(1), 183–191.
- Hull, J. and A. White (1987). The Pricing of Options on Assets with Stochastic Volatility. *Journal of Finance* 42(2), 281–300.
- Hull, J. C. (2021). *Options, Futures, and Other Derivatives* (11th ed.). Pearson Education.
- Hvidkjaer, S. (2006). A Trade-Based Analysis of Momentum. *Review of Financial Studies* 19(2), 457–491.
- Ilmanen, A. (2003). Stock-Bond Correlations. *Journal of Fixed Income* 13(2), 55–66.
- Ilmanen, A., R. Israel, T. J. Moskowitz, A. K. Thapar, and R. Lee (2021). How Do Factor Premia Vary over Time? A Century of Evidence. *Journal of Investment Management* 19(4), 15–57.
- Ingersoll, Jr., J. E. (1984). Some Results in the Theory of Arbitrage Pricing. *Journal of Finance* 39(4), 1021–1039.
- Jagannathan, R. and N. R. Kocherlakota (1996). Why Should Older People Invest Less in Stocks Than Younger People? *Federal Reserve Bank of Minneapolis Quarterly Review* 20(3), 11–23.
- Jagannathan, R. and Z. Wang (1996). The Conditional CAPM and the Cross-Section of Expected Returns. *Journal of Finance* 51(1), 3–53.
- Jegadeesh, N. and S. Titman (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance* 48(1), 65–91.
- Jegadeesh, N. and S. Titman (2001). Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *Journal of Finance* 56(2), 699–718.
- Jensen, T. I., B. Kelly, and L. H. Pedersen (2023). Is there a Replication Crisis in Finance? *Journal of Finance* 78(5), 2465–2518.
- Jobson, J. D. and B. Korkie (1980). Estimation for Markowitz Efficient Portfolios. *Journal of the American Statistical Association* 75(371), 544–554.
- Joenväärä, J., M. Kauppila, R. Kosowski, and P. Tolonen (2021). Hedge Fund Performance: Are Stylized Facts Sensitive to Which Database One Uses? *Critical Finance Review* 10(2), 271–327.
- Jondeau, E., Q. Zhang, and X. Zhu (2019). Average Skewness Matters. *Journal of Financial Economics* 134(1), 29–47.
- Jones, J. S. and B. Kincaid (2014). Can the Correlation among Dow 30 Stocks Predict Market Declines?: Evidence from 1950 to 2008. *Managerial Finance* 40(1), 33–50.
- Jorion, P. (1986). Bayes-Stein Estimation for Portfolio Analysis. *Journal of Financial and Quantitative Analysis* 21(3), 279–292.
- Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decisions under Risk. *Econometrica* 47(2), 263–291.
- Kan, R. and G. Zhou (2007). Optimal Portfolio Choice with Parameter Uncertainty. *Journal of Financial and Quantitative Analysis* 42(3), 621–656.
- Keown, A. J. and J. M. Pinkerton (1981). Merger Announcements and Insider Trading Activity: An Empirical Investigation. *Journal of Finance* 36(4), 855–869.

- Kim, T. S. and E. Omberg (1996). Dynamic Nonmyopic Portfolio Behavior. *Review of Financial Studies* 9(1), 141–161.
- Koijen, R. S. J., J. C. Rodriguez, and A. Sbuelz (2009). Momentum and Mean Reversion in Strategic Asset Allocation. *Management Science* 55(7), 1199–1213.
- Koijen, R. S. J. and S. van Nieuwerburgh (2011). Predictability of Returns and Cash Flows. *Annual Review of Financial Economics* 3, 467–491.
- Kothari, S. P. and J. Shanken (1997). Book-to-Market, Dividend Yield, and Expected Market Returns: A Time-Series Analysis. *Journal of Financial Economics* 44(2), 169–203.
- Kothari, S. P., J. Shanken, and R. G. Sloan (1995). Another Look at the Cross-Section of Expected Stock Returns. *Journal of Finance* 50(1), 185–224.
- Kraft, H. (2005). Optimal Portfolios and Heston's Stochastic Volatility Model. *Quantitative Finance* 5, 303–313.
- Kraft, H. and C. Munk (2011). Optimal Housing, Consumption, and Investment Decisions over the Life-Cycle. *Management Science* 57(6), 1025–1041.
- Kraft, H., C. Munk, F. T. Seifried, and S. Wagner (2017). Consumption Habits and Humps. *Economic Theory* 64(2), 305–330.
- Kraft, H., C. Munk, and S. Wagner (2018). Housing Habits and their Implications for Life-Cycle Consumption and Investment. *Review of Finance* 22(5), 1737–1762.
- Kroencke, T. A. (2017). Asset Pricing without Garbage. *Journal of Finance* 72(1), 47–98.
- Larsen, L. S. and C. Munk (2023). The Design and Welfare Implications of Mandatory Pension Plans. *Journal of Financial and Quantitative Analysis* 58(8), 3420–3449.
- Ledoit, O. and M. Wolf (2004). Honey, I Shrunk the Sample Covariance Matrix. *Journal of Portfolio Management* 30(4), 110–119.
- Ledoit, O. and M. Wolf (2017). Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *Review of Financial Studies* 30(12), 4349–4388.
- Lee, M.-L., M.-T. Lee, and K. Chiang (2008). Real Estate Risk Exposure of Equity Real Estate Investment Trusts. *Journal of Real Estate Finance & Economics* 36(2), 165–181.
- Lettau, M. and S. C. Ludvigson (2001). Consumption, Aggregate Wealth and Expected Stock Returns. *Journal of Finance* 56(3), 815–849.
- Lettau, M. and S. C. Ludvigson (2010). Measuring and Modeling Variation in the Risk-Return Tradeoff. In Y. Ait-Sahalia and L. P. Hansen (Eds.), *Handbook of Financial Econometrics*, Volume 1, pp. 618–682. North Holland.
- Lettau, M. and A. Madhavan (2018). Exchange-Traded Funds 101 for Economists. *Journal of Economic Perspectives* 32(1), 135–154.
- Lettau, M. and S. van Nieuwerburgh (2008). Reconciling the Return Predictability Evidence. *Review of Financial Studies* 21(4), 1607–1652.
- Li, X., R. N. Sullivan, and L. Garcia-Feijóo (2014). The Limits to Arbitrage and the Low-Volatility Anomaly. *Financial Analysts Journal* 70(1), 52–63.
- Liew, J. and M. Vassalou (2000). Can Book-to-Market, Size, and Momentum Be Risk Factors that Predict Economic Growth? *Journal of Financial Economics* 57(2), 221–245.
- Lin, Q. (2020). Idiosyncratic Momentum and the Cross-Section of Stock Returns: Further Evidence. *European Financial Management* 26(3), 579–627.
- Linnainmaa, J. T. and M. R. Roberts (2018). The History of the Cross Section of Stock Returns. *Review of Financial Studies* 31(7), 2606–2649.
- Lintner, J. (1965). The Valuation of Risky Assets and the Selection of Risky Investment in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics* 47(1), 13–37.
- Litzenberger, R. H. and K. Ramaswamy (1979). The Effects of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence. *Journal of Financial Economics* 7, 163–195.
- Liu, J. (1999, August). Portfolio Selection in Stochastic Environments. Working paper, Stanford University.

- Liu, J. (2007). Portfolio Selection in Stochastic Environments. *Review of Financial Studies* 20(1), 1–39.
- Liu, J. and J. Pan (2003). Dynamic Derivative Strategies. *Journal of Financial Economics* 69(3), 401–430.
- Liu, L. X. and L. Zhang (2008). Momentum Profits, Factor Pricing, and Macroeconomic Risk. *Review of Financial Studies* 21(6), 2417–2448.
- Lucas, R. E. (1978). Asset Prices in an Exchange Economy. *Econometrica* 46(6), 1429–1445.
- Lustig, H. N. and S. G. van Nieuwerburgh (2005). Housing Collateral, Consumption Insurance, and Risk Premia: An Empirical Perspective. *Journal of Finance* 60(3), 1167–1219.
- Lutz, F. (1940). The Structure of Interest Rates. *Quarterly Journal of Economics* 55(1), 36–63.
- Lynch, A. W. and S. Tan (2011). Labor Income Dynamics at Business-Cycle Frequencies: Implications for Portfolio Choice. *Journal of Financial Economics* 101(2), 333–359.
- Macaulay, F. R. (1938). *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States since 1856*. New York: Columbia University Press.
- MacKinlay, A. C. and L. Pastor (2000). Asset Pricing Models: Implications for Expected Returns and Portfolio Selection. *Review of Financial Studies* 13(4), 883–916.
- Madan, D. B., P. P. Carr, and E. C. Chang (1998). The Variance-Gamma Process and Option Pricing. *European Finance Review* 2(1), 79–105.
- Magill, M. J. P. and G. M. Constantinides (1976). Portfolio Selection with Transactions Costs. *Journal of Economic Theory* 13, 245–263.
- Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *Journal of Business* 36(4), 394–419.
- Markowitz, H. (1952). Portfolio Selection. *Journal of Finance* 7(1), 77–91.
- Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investment*. Wiley.
- Mayers, D. (1972). Nonmarketable Assets and Capital Market Equilibrium under Uncertainty. In M. C. Jensen (Ed.), *Studies in the Theory of Capital Markets*. Praeger Publishers.
- McLean, R. D. and J. Pontiff (2016). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance* 71(1), 5–32.
- Mehra, R. and E. C. Prescott (1985). The Equity Premium: A Puzzle. *Journal of Monetary Economics* 15(2), 145–162.
- Mehra, R. and E. C. Prescott (2003). The Equity Premium in Retrospect. In G. M. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 1B, Chapter 14, pp. 889–938. Elsevier.
- Merton, R. C. (1969). Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case. *Review of Economics and Statistics* 51(3), 247–257.
- Merton, R. C. (1971). Optimum Consumption and Portfolio Rules in a Continuous-Time Model. *Journal of Economic Theory* 3(4), 373–413.
- Merton, R. C. (1972). An Analytic Derivation of the Efficient Portfolio Frontier. *Journal of Financial and Quantitative Analysis* 7, 1851–1872.
- Merton, R. C. (1973a). An Intertemporal Capital Asset Pricing Model. *Econometrica* 41(5), 867–887.
- Merton, R. C. (1973b). The Relationship between Put and Call Option Prices: Comment. *Journal of Finance* 28(1), 183–184.
- Merton, R. C. (1973c). Theory of Rational Option Pricing. *Bell Journal of Economics and Management Science* 4, 141–183.
- Merton, R. C. (1976). Option Pricing When Underlying Stock Returns are Discontinuous. *Journal of Financial Economics* 3(1-2), 125–144.
- Merton, R. C. (1980). On Estimating the Expected Return on the Market: An Exploratory Investigation. *Journal of Financial Economics* 8(4), 323–361.

- Meyer, D. J. and J. Meyer (2005). Relative Risk Aversion: What Do We Know? *Journal of Risk and Uncertainty* 31(3), 243–262.
- Michaud, R. O. (1989). The Markowitz Optimization Enigma: Is ‘Optimized’ Optimal? *Financial Analysts Journal* 45(1), 31–42.
- Modigliani, F. and L. Modigliani (1997). Risk-Adjusted Performance. *Journal of Portfolio Management* 23(2), 45–54.
- Modigliani, F. and R. Sutch (1966). Innovations in Interest Rate Policy. *American Economic Review* 56(1-2), 178–197.
- Møller, S. V. and J. Rangvid (2015). End-of-the-year Economic Growth and Time-varying Expected Returns. *Journal of Financial Economics* 115(1), 136–154.
- Moskowitz, T., Y. H. Ooi, and L. H. Pedersen (2012). Time Series Momentum. *Journal of Financial Economics* 104(2), 228–250.
- Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica* 34(4), 768–783.
- Munk, C. (2000). Optimal Consumption-Investment Policies with Undiversifiable Income Risk and Liquidity Constraints. *Journal of Economic Dynamics and Control* 24(9), 1315–1343.
- Munk, C. (2008). Portfolio and Consumption Choice with Stochastic Investment Opportunities and Habit Formation in Preferences. *Journal of Economic Dynamics and Control* 32(11), 3560–3589.
- Munk, C. (2011). *Fixed Income Modelling*. Oxford University Press.
- Munk, C. (2013). *Financial Asset Pricing Theory*. Oxford University Press.
- Munk, C. (2017). Dynamic Asset Allocation. Lecture notes, Copenhagen Business School.
- Munk, C. (2020). A Mean-Variance Benchmark for Household Portfolios over the Life Cycle. *Journal of Banking & Finance* 116, 105833.
- Munk, C. and C. Sørensen (2004). Optimal Consumption and Investment Strategies with Stochastic Interest Rates. *Journal of Banking & Finance* 28(8), 1987–2013.
- Munk, C. and C. Sørensen (2010). Dynamic Asset Allocation with Stochastic Income and Interest Rates. *Journal of Financial Economics* 96(3), 433–462.
- von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. New Jersey: Princeton University Press.
- Nielsen, L. T. and M. Vassalou (2006). The Instantaneous Capital Market Line. *Economic Theory* 28(3), 651–664.
- Novy-Marx, R. (2011). Operating Leverage. *Review of Finance* 15(1), 103–134.
- Novy-Marx, R. (2013). The Other Side of Value: The Gross Profitability Premium. *Journal of Financial Economics* 108(1), 1–28.
- Novy-Marx, R. and M. Velikov (2016). A Taxonomy of Anomalies and Their Trading Costs. *Review of Financial Studies* 29(1), 104–147.
- Novy-Marx, R. and M. Velikov (2022). Betting Against Betting Against Beta. *Journal of Financial Economics* 143(1), 80–106.
- Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *Journal of Finance* 53(5), 1775–1798.
- Ogaki, M. and Q. Zhang (2001). Decreasing Relative Risk Aversion and Tests of Risk Sharing. *Econometrica* 69(2), 515–526.
- Ottaviani, M. and P. N. Sørensen (2015). Price Reaction to Information with Heterogeneous Beliefs and Wealth Effects: Underreaction, Momentum, and Reversal. *American Economic Review* 105(1), 1–34.
- Pagliari, J. L., K. A. Scherer, and R. T. Monopoli (2005). Public Versus Private Real Estate Equities: A More Refined, Long-Term Comparison. *Real Estate Economics* 33(1), 147–187.
- Pastor, L., R. F. Stambaugh, and L. A. Taylor (2021). Sustainable Investing in Equilibrium. *Journal of Financial Economics* 142(2), 550–571.

- Pastor, L., R. F. Stambaugh, and L. A. Taylor (2022). Dissecting Green Returns. *Journal of Financial Economics* 146(2), 403–424.
- Pastor, L., R. F. Stambaugh, and L. A. Taylor (2024, January). Green Tilts. Available at SSRN: <http://ssrn.com/abstract=4464537>.
- Patton, A. J. and B. M. Weller (2020). What You See Is Not What You Get: The Costs of Trading Market Anomalies. *Journal of Financial Economics* 137(2), 515–549.
- Pedersen, L. H. (2015). *Efficiently Inefficient*. Princeton University Press.
- Pedersen, L. H. (2018). Sharpening the Arithmetic of Active Management. *Financial Analysts Journal* 74(1), 21–36.
- Pedersen, L. H. (2022). Game On: Social Networks and Markets. *Journal of Financial Economics* 146(3), 1097–1119.
- Pedersen, L. H., S. Fitzgibbons, and L. Pomorski (2021). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics* 142(2), 572–597.
- Pelizzon, L. and G. Weber (2009). Efficient Portfolios when Housing Needs Change over the Life Cycle. *Journal of Banking & Finance* 33(11), 2110–2121.
- Penman, S. H. (2013). *Financial Statement Analysis and Security Valuation* (5 ed.). McGraw-Hill Education.
- Petersen, C. and T. Plenborg (2012). *Financial Statement Analysis: Valuation – Credit Analysis – Executive Compensation*. FT Prentice Hall.
- Petkova, R. and L. Zhang (2005). Is Value Riskier Than Growth? *Journal of Financial Economics* 78(1), 187–202.
- Poitras, G. (2009a). The Early History of Option Contracts. In W. Hafner and H. Zimmermann (Eds.), *Vinzenz Bronzin's Option Pricing Models: Exposition and Appraisal*, pp. 487–518. Springer-Verlag.
- Poitras, G. (2009b). From Antwerp to Chicago: The History of Exchange Traded Derivative Security Contracts. *Revue d'Historie des Sciences Humaines* 20, 11–50.
- Quinn, D. P. and H.-J. Voth (2008). A Century of Global Equity Market Correlations. *American Economic Review* 98(2), 535–540.
- Rangvid, J. (2006). Output and Expected Returns. *Journal of Financial Economics* 81(3), 595–624.
- Rangvid, J., M. Schmeling, and A. Schrimpf (2014). Dividend Predictability Around the World. *Journal of Financial and Quantitative Analysis* 49(5-6), 1255–1277.
- Ravina, E. (2019, April). Habit Formation and Keeping Up with the Joneses: Evidence from Micro Data. Available at SSRN: <http://ssrn.com/abstract=928248>.
- Reinganum, M. R. (1981). Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings' Yields and Market Values. *Journal of Financial Economics* 9(1), 19–46.
- Rendleman, R. and B. Bartter (1979). Two State Option Pricing. *Journal of Finance* 34(5), 1093–1110.
- Rendleman, R. J., C. P. Jones, and H. A. Latané (1982). Empirical Anomalies Based on Unexpected Earnings and the Importance of Risk Adjustments. *Journal of Financial Economics* 10(3), 269–287.
- Rogers, L. (2001). The Relaxed Investor and Parameter Uncertainty. *Finance and Stochastics* 5, 131–154.
- Roll, R. (1977). A Critique of the Asset Pricing Theory's Tests. *Journal of Financial Economics* 4(2), 129–176.
- Rosenberg, B. and J. Guy (1976a). Prediction of Beta from Investment Fundamentals: Part One. Prediction Criteria. *Financial Analysts Journal* 32(3), 60–72.
- Rosenberg, B. and J. Guy (1976b). Prediction of Beta from Investment Fundamentals: Part Two. Alternative Prediction Methods. *Financial Analysts Journal* 32(4), 62–70.
- Rosenberg, B., K. Reid, and R. Lanstein (1985). Persuasive Evidence of Market Inefficiency. *Journal of Portfolio Management* 11, 9–16.

- Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* 13(3), 341–360.
- Ross, S. A. (2005). *Neoclassical Finance*. Princeton University Press.
- Rouwenhorst, K. G. (1998). International Momentum Strategies. *Journal of Finance* 53(1), 267–284.
- Rubinstein, M. (1976). The Strong Case for the Generalized Logarithmic Utility Model as the Premier Model of Financial Markets. *Journal of Finance* 31(2), 551–571.
- Ryder, Harl E., J. and G. M. Heal (1973). Optimal Growth with Intertemporally Dependent Preferences. *Review of Economic Studies* 40(1), 1–31.
- Samuelson, P. A. (1965). Proof that Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review* 6(2), 41–49.
- Samuelson, P. A. (1969). Lifetime Portfolio Selection by Dynamic Stochastic Programming. *Review of Economics and Statistics* 51(3), 239–246.
- Sangvinatsoo, A. and J. A. Wachter (2005). Does the Failure of the Expectations Hypothesis Matter for Long-Term Investors? *Journal of Finance* 60(1), 179–230.
- Santos, T. and P. Veronesi (2006). Labor Income and Predictable Stock Returns. *Review of Financial Studies* 19(1), 1–44.
- Savor, P. G. and M. I. Wilson (2016). Earnings Announcements and Systematic Risk. *Journal of Finance* 71(1), 83–138.
- Savov, A. (2011). Asset Pricing with Garbage. *Journal of Finance* 66(1), 177–201.
- Schmelzing, P. (2020). Eight Centuries of Global Real Interest Rates, R-G, and the “Suprasecular” Decline, 1311–2018. Staff working paper no. 845, Bank of England.
- Schröder, D. and F. Esterer (2016). A New Measure of Equity and Cash Flow Duration: The Duration-Based Explanation of the Value Premium Revisited. *Journal of Money, Credit and Banking* 48(5), 857–900.
- Schwert, G. W. (2003). Anomalies and Market Efficiency. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 1B, Chapter 15, pp. 937–972. Elsevier.
- Sharpe, W. (1963). A Simplified Model of Portfolio Analysis. *Management Science* 9(2), 277–293.
- Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance* 19(3), 425–442.
- Shefrin, H. and M. Statman (1985). The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. *Journal of Finance* 40(3), 777–790.
- Shiller, R. J. (2000). *Irrational Exuberance*. Princeton, NJ: Princeton University Press.
- Shiller, R. J. (2005). *Irrational Exuberance* (Second ed.). Princeton, NJ: Princeton University Press.
- Shiller, R. J. and A. E. Beltratti (1992). Stock Prices and Bond Yields: Can Their Comovements be Explained in Terms of Present Value Models? *Journal of Monetary Economics* 30(1), 25–46.
- Siegel, J. J. (2016). The Shiller CAPE Ratio: A New Look. *Financial Analysts Journal* 72(3), 41–50.
- Sørensen, C. (1999). Dynamic Asset Allocation and Fixed Income Management. *Journal of Financial and Quantitative Analysis* 34(4), 513–531.
- Stambaugh, R. F. (1988). The Information in Forward Rates: Implications for Models of the Term Structure. *Journal of Financial Economics* 21(1), 41–70.
- Starks, L. T. (2023). Sustainable Finance and ESG Issues: Value versus Values. *Journal of Finance* 78(4), 1837–1872. Presidential Address to the American Finance Association.
- Stoll, H. R. (1969). The Relationship Between Put and Call Option Prices. *Journal of Finance* 24(5), 801–824.

- Svensson, L. E. O. and I. M. Werner (1993). Nontraded Assets in Incomplete Markets. *European Economic Review* 37(5), 1149–1168.
- Sydsæter, K., P. Hammond, A. Strøm, and A. Carvajal (2021). *Essential Mathematics for Economic Analysis* (6th ed.). Pearson.
- Thurow, L. (1969). The Optimum Lifetime Distribution of Consumption Expenditures. *American Economic Review* 59(3), 324–330.
- Tomz, M. and M. L. J. Wright (2007). Do Countries Default in “Bad Times”? *Journal of the European Economic Association* 5(2–3), 352–360.
- Treynor, J. L. (1961). Market Value, Time, and Risk. Originally unpublished. A slightly edited version is available at SSRN: <http://ssrn.com/abstract=2600356>.
- Treynor, J. L. and F. Black (1973). How to Use Security Analysis to Improve Portfolio Selection. *Journal of Business* 46(1), 66–86.
- Tu, J. and G. Zhou (2011). Markowitz Meets Talmud: A Combination of Sophisticated and Naïve Diversification Strategies. *Journal of Financial Economics* 99(1), 204–215.
- Vasicek, O. (1973). A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas. *Journal of Finance* 28(5), 1233–1239.
- Veronesi, P. (2010). *Fixed Income Securities*. Wiley.
- Viceira, L. M. (2001). Optimal Portfolio Choice for Long-Horizon Investors with Nontradable Labor Income. *Journal of Finance* 56(2), 433–470.
- Vissing-Jørgensen, A. (2002, March). Towards an Explanation of Household Portfolio Choice Heterogeneity: Nonfinancial Income and Participation Cost Structures. Available at SSRN: <http://ssrn.com/abstract=307121>.
- Wachter, J. A. (2002). Portfolio and Consumption Decisions under Mean-Reverting Returns: An Exact Solution for Complete Markets. *Journal of Financial and Quantitative Analysis* 37(1), 63–91.
- Weber, M. (2018). Cash Flow Duration and the Term Structure of Equity Returns. *Journal of Financial Economics* 128(3), 486–503.
- Welch, I. (2022). Simply Better Market Betas. *Critical Finance Review* 11(1), 37–64.
- Wilcox, D. W. (1992). The Construction of U.S. Consumption Data: Some Facts and Their Implications for Empirical Work. *American Economic Review* 82(4), 922–941.
- Yao, R. and H. H. Zhang (2005). Optimal Consumption and Portfolio Choices with Risky Housing and Borrowing Constraints. *Review of Financial Studies* 18(1), 197–239.
- Zeng, L. and P. Luk (2020). Examining Share Repurchasing and the S&P Buyback Indices in the U.S. Market. Research paper, S&P Dow Jones Indices.
- Zhang, L. (2005). The Value Premium. *Journal of Finance* 60(1), 67–103.