Portfolio Management

Lecture Notes by Johannes Muhle-Karbe*

Spring 2021

Contents

*Imperial College London, Department of Mathematics, email j.muhle-karbe@imperial.co.ak.

1 Introduction: "Efficiently Inefficient Markets"

Efficient Market Hypothesis A very influential paradigm of classical financial theory is that markets are *efficient*. More specifically, the so-called *efficient market hypothesis* stipulates that prices reflect *all* relevant information at all times. This means that there is no point in trying to "beat the market" because this is simply not possible. As a consequence, it makes no sense to pay the high management fees charged by actively managed investment funds since, on average, these will underperform the market precisely by the fees charged. Instead, one should invest in a suitable mix between (broadly diversified) passive index funds and (almost) riskless bonds according to one's risk preferences.

However, the question remains why markets should be efficient in the first place. The usual argument for this goes as follows. Imagine there was an investment opportunity that yields high profits with little risk. Then, some attentive investor would immediately exploit it to the point that it disappears. But, if everyone believes in efficient markets and gives up on trying to exploit mispricings, then who should be making the market efficient in the first place?[1]

Inefficiencies On the other end of the spectrum, it has been argued that markets are *inefficient*, in that they are significantly affected by the whims of irrational investors and systematic behavioral biases. If this were true, and market prices would bounce around erratically with little link to fundamentals, then beating the market should be easy. As a consequence, there should be an abundance of highly profitable active investment strategies.

However, financial markets are extremely competitive, and even investment professionals who dedicate their entire careers to this often fail to consistently beat the market.

Efficient Inefficiency Altogether, the discussion above suggests that the behavior of real markets lies in between these two extreme views, in that markets are *efficiently inefficient*. This means that *some* inefficiencies exist and can be exploited for a profit, but only to an *efficient* extent. To wit, competition between professional investors makes markets almost efficient, but they remain inefficient enough so that investors are still compensated for their costs and risks.

If markets are efficiently inefficient, then a limited amount of capital can be invested profitably by active investors with a comparative advantage. In this course, we will discuss some of the basic tools that play important roles in this context. For example:

1. How should the performance of a proposed trading strategy be evaluated?

2. How can one come up with potentially profitable trading strategies?

[1] This intuitive argument is formalized in Grossman and Stiglitz (1980).

3. How should a promising strategy be optimised and how should one manage the associated risk?

4. Once one has decided on a basic trading strategy, how should it be implemented?

This Course Parts of this lecture closely follow the textbook *Efficiently Inefficient* (Pedersen, 2015) by Lasse H. Pedersen, a leading finance academic and a principal at AQR, a large quantitative asset manager. This book provides an excellent overview over the theory and practice of asset management, including many institutional details that we will not be able to cover in this short elective. However, since the book is tailored to MBA students, it is deliberately light on theory, both regarding methods for portfolio construction and implementation, and asset pricing theory. Good accessible textbook references on this for readers with a background in Mathematics are not easy to find,[2] so these are developed in a self-contained manner here.

This course does *not* require prior knowledge of stochastic optimal control, and therefore presents most results in the simplest one-period settings that can be analyzed without such sophisticated tools. The asset-pricing lecture notes of Pietro Veronesi available online at `http://pietroveronesi.org/teaching/BUS35907.htm` are a great exposition of multi-period and continuous-time versions of some of the results touched upon in this course.

A crucial complement to the contents covered in the course are the problem sets and the empirical courseworks. Some of the problem sets explore variations and extensions of the theory from the lectures; working these out by yourself is key to really understanding the concepts and tools involved. The other problem sets and the coursework focus on practical implementations with real data. This is crucial both to understand the applications and the limitations of the theoretical material, and a great practice for a future career in the financial industry.

2 Evaluating Trading Strategies

Suppose you work for a hedge fund and one of your colleagues has come up with a new trading strategy which would have produced handsome profits on historical data. How do you evaluate whether it might indeed make sense to invest money according to this strategy?

Alternatively, suppose you work for a pension fund that is investing some of its capital through hedge funds. How do you measure which funds are performing well, and which ones are doing poorly and should potentially be replaced?

2.1 Alpha and Beta

The most basic measure of trading performance is the *return* R_t in a given period t, that is the amount earned per dollar invested.[3] Since (almost) risk-free bonds typically pay nonzero interest rates and stock prices rise even faster on average, just asking for a positive return is

[2] One classical reference is the textbook *Dynamic Asset Pricing Theory* (Duffie, 2001), but this is rather light on concrete examples.

[3] That is, if the value of the portfolio changes from X_{t-1} at time $t-1$ to X_t at time t, then the return in the period $[t-1, t]$ is $R_t = (X_t - X_{t-1})/X_{t-1}$.

not a reasonable benchmark for an actively managed trading strategy. Instead, one typically decomposes returns in "alpha" and "beta". Here, beta is the strategy's market exposure, and alpha is the *excess* return that the strategy achieves on top of the corresponding market returns. These parameters are computed by running a linear regression of the strategy's excess returns over the safe interest rate,

$$R_t^e = R_t - R^f$$

against the market excess return $R_t^{M,e} = R_t^M - R^f$:

$$R_t^e = \alpha + \beta R_t^{M,e} + \varepsilon_t. \tag{2.1}$$

The slope β in this representation measures the strategy's exposure to market movements. For example, if $\beta = 0.5$ and the market drops by 10%, then the strategy can be expected to drop by 5% everything else being equal. In addition to the market risk that is taken this way, the strategy may also be exposed to other idiosyncratic risks collected in the noise term ε_t. For example, if the strategy overweights certain industries (e.g., FAANG) relative to the whole market, then this term captures the relative outperformance of this portfolio relative to the market. This can be positive or negative; by virtue of the linear regression it will be zero on average and independent of the market moves (in sample!).

What is the key difference between these two exposures to risk? Idiosyncratic risks can be diversified away to a certain extent be investing into many different strategies that are independent enough. In contrast, exposure to market risk is a common factor that does not average out. Moreover, exposure to market risk can be obtained very cheaply by buying index funds with very low management fees, so high fees cannot be justified by just buying and holding the market. As a consequence, many hedge funds in fact strive to be *market neutral*, in that $\beta = 0$. Then, the strategy should generate profits independent of the performance of the market as a whole. However, since stock price do rise faster than the risk-free rate on average, many funds also include some market exposure to boost their gross returns.

In any case, if the above linear regression does a good job of describing the returns, then the strategy can be made market neutral by hedging out the market exposure. To wit, for every dollar invested in the strategy, one shorts β dollars of market exposure, leading to a total excess return of

$$R_t^e - \beta R_t^{M,e} = \alpha + \varepsilon_t.$$

Since the idiosyncratic risk ε_t is zero on average, this implies that α is the *market-neutral excess return*. This is the holy grail of active investing, since it measures the value added by investment skills on top of just buying market exposure. As α is estimated from a historical sample, it could also be nonzero simply due to luck or, even worse, data snooping as we discuss later.

Since only limited data is available and the relationship (2.1) cannot be expected to be stable over long time horizons, the alpha and beta of a trading strategy can typically only be estimated with considerable error. To address this, it is standard practice to look at the *t-statistic*, that is, the estimate for alpha divided by the standard error of the estimate (which is reported by all common regression analysis tools). A large t-statistic suggests that the value of α is large and estimated reliably. More specifically, it is customary to consider t-statistics above 2 as statistically significant.

2.2 Capital Asset Pricing Model

Why should one run the regression above on market returns and not on something else, or include additional explanatory variables? The motivation for this comes from the classic *Capital Asset Pricing Model (CAPM)* pioneered by Sharpe (1964); Lintner (1965). We illustrate the main idea in the simplest one-period setting.

Mean-Variance Optimization As a first step, we discuss how the trading decisions of an individual investor can be optimized. Consider a market consisting of a risk-free asset and N risky assets. The risk-free asset earns the risk-free return R^f, whereas the returns $R_t^n = R^f + R_t^{n,e}$ of the risky assets $n = 1, \ldots, N$ are uncertain. If you start with W_{t-1} dollars at time $t-1$ and invest x_{t-1}^n dollars into asset n then, after the trades are implemented and price changes are realized, your wealth at time t will be

$$\begin{aligned} W_t^{x_{t-1}} &= x_{t-1}^1(1+R_t^1) + \ldots + x_{t-1}^N(1+R_t^N) + (W_{t-1} - x_{t-1}^1 - \ldots - x_{t-1}^N)(1+R^f) \\ &= W_{t-1}(1+R^f) + x_{t-1}^\top R^e. \end{aligned}$$

The simplest goal functional one can try to optimize in this context is to maximize expected wealth, penalized for a multiple of its variance:

$$\begin{aligned} J_{t-1}(x_{t-1}) &= \mathrm{E}_{t-1}\left[W_t^{x_{t-1}}\right] - \frac{\gamma}{2}\mathrm{Var}_{t-1}\left[W_t^{x_{t-1}}\right] \\ &= W_{t-1}(1+R^f) + x_{t-1}^\top \mathrm{E}_{t-1}[R_t^e] - \frac{\gamma}{2} x_{t-1}^\top \mathrm{Cov}_{t-1}[R_t^e, R_t^e] x_{t-1}. \end{aligned}$$

The first-order condition for this strictly convex goal functional is

$$0 = \nabla J_{t-1}(x_{t-1}) = \mathrm{E}_{t-1}[R_t^e] - \gamma \mathrm{Cov}_{t-1}[R_t^e, R_t^e] x_{t-1}.$$

Whence, given that the covariance matrix of the risky returns is invertible, the optimal investments in the risky assets are

$$\hat{x}_{t-1} = (\gamma \mathrm{Cov}_{t-1}[R_t^e, R_t^e])^{-1} \mathrm{E}_{t-1}[R_t^e]. \tag{2.2}$$

Representative Agent Equilibrium So far, we have considered the optimization problem of a single agent, who chooses how to invest in a set of assets. The capital asset pricing model takes this as the starting point and *assumes* that the trading behaviour of the *entire* market can be subsumed by a single "representative investor" who invests according to a mean-variance goal functional of the above form. By matching the demand of this representative agent to a zero net supply of the risk-free asset (individual agents borrow from each other) and fixed supplies s^n, $n = 1, \ldots, N$ of the risky assets, one then draws conclusions about the relationships between the assets' expected returns.

To clear the market, the currently observable "market capitalizations" (outstanding shares times market prices)

$$\mathrm{Cap}_{t-1} = (s^1 P_{t-1}^1, , \ldots, s^N P_{t-1}^N)^\top$$

of the risky assets need to match the monetary investments $\hat{x}_t$ from (2.2) that the representative agent wants to make:

$$\mathrm{Cap}_{t-1} = (\gamma \mathrm{Cov}_{t-1}[R_t^e, R_t^e])^{-1} \mathrm{E}_{t-1}[R_t^e].$$

Equivalently:

$$\mathrm{E}_{t-1}[R_t^e] = \gamma \mathrm{Cov}_{t-1}[R_t^e, R_t^e] \mathrm{Cap}_{t-1},$$

so that the expected excess return of asset n in turn has to be[4]

$$\begin{aligned} \mathrm{E}_{t-1}[R_t^{n,e}] &= \gamma \sum_{m=1}^{N} \mathrm{Cov}_{t-1}[R_t^{n,e}, R_t^{m,e}] \mathrm{Cap}_{t-1}^m \\ &= \gamma \mathrm{Cov}_{t-1} \left[R_t^{n,e}, \sum_{m=1}^{N} \mathrm{Cap}_{t-1}^m R_t^{m,e} \right] \\ &= \gamma \mathrm{Cov}_{t-1} \left[R_t^{n,e}, R_t^{M,e} \right] \sum_{m=1}^{N} \mathrm{Cap}_{t-1}^m. \end{aligned} \tag{2.3}$$

This representation has the disadvantage of depending on the risk aversion of the representative agent, which is not observable. However, by multiplying each return with the market capitalization of the respective asset, then summing across assets, and finally dividing by the total market capitalization, we obtain an analogous representation for the market portfolio:

$$\mathrm{E}_{t-1}[R_t^{M,e}] = \gamma \mathrm{Var}_{t-1} \left[R_t^{M,e} \right] \sum_{m=1}^{M} \mathrm{Cap}_{t-1}^m. \tag{2.4}$$

By combining (2.3) and (2.4), we then obtain a linear relationship between individual and market returns as postulated in (2.1):

$$\mathrm{E}_{t-1}[R_t^{n,e}] = \beta_{t-1}^n \mathrm{E}_{t-1}[R_t^{M,e}], \quad \text{where } \beta_{t-1}^n = \frac{\mathrm{Cov}_{t-1}[R_t^{n,e}, R_t^{M,e}]}{\mathrm{Var}_{t-1}[R_t^{M,e}]}.$$

If all returns appearing here a iid, then the coefficient β^n becomes constant. If one replaces variances and covariances by their standard estimators, then this is exactly the slope coefficient appearing in a linear regression of the individual returns against the market returns. Thus, if (i) excess returns are approximately iid and (ii) there is a representative agent with one-period mean-variance preferences, then the expected excess returns of the individual assets and the market are linked by a linear relationship of exactly the same form as in (2.1) – with the market portfolio as the only explanatory variable. Crucially, we have $\alpha = 0$

[4]Here, we use in the third step that the value of the market portfolio is $\sum_{m=1}^{N} \mathrm{Cap}_{t-1}^m$ so that its excess return is

$$R_t^M = \frac{\sum_{m=1}^{N} (\mathrm{Cap}_t^m - \mathrm{Cap}_{t-1}^m)}{\sum_{k=1}^{N} \mathrm{Cap}_{t-1}^k} = \sum_{m=1}^{N} \frac{\mathrm{Cap}_{t-1}^m}{\sum_{k=1}^{N} \mathrm{Cap}_{t-1}^k} \frac{s^m P_t^m - s^m P_{t-1}^m}{s^m P_{t-1}^m} = \sum_{m=1}^{N} \frac{\mathrm{Cap}_{t-1}^m}{\sum_{k=1}^{N} \mathrm{Cap}_{t-1}^k} R_t^m.$$

As the weights in the sum add to one, it follows that the excess return of the market portfolio indeed is $R_t^{M,e} = \sum_{m=1}^{N} \mathrm{Cap}_{t-1}^m R_t^{m,e} / \sum_{k=1}^{N} \mathrm{Cap}_{t-1}^k$.

here – there is no alpha for each individual asset and accordingly also not for any portfolio formed of them. Therefore, hedge funds' search for alpha is a quest to defy the CAPM and earn higher returns than simple compensation for market risk.

Is the CAPM a reasonable approximation of real data? These has been debated heatedly. Figure 1 from Pedersen (2015) plots the performance of ten portfolios of US stocks sorted by their betas, estimated using data from 1926 to 2010. The horizontal axis shows the CAPM predicted returns (beta times average market excess return), whereas the vertical axis plots the average returns actually realized. Whence, the dashed 45-degree line is the hypothetical "security market line" implied by the CAPM on which all the dots should lie if the model were correct.

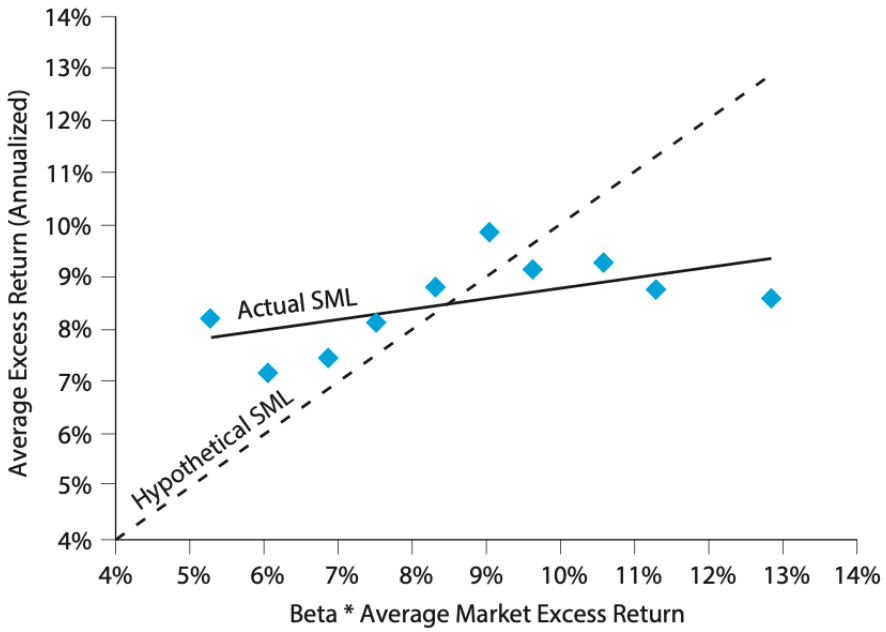


Figure 1: The "security market line" is too flat empirically compared to the CAPM.

Clearly, the empirical security market line is much flatter than predicted by the model, in that the returns of stocks with much higher betas are not all that much higher than the returns of stocks with lower betas. We will see later how (market and funding) liquidity risk can explain (part of) this observation.

Dividends and Equilibrium Prices Under certain assumptions, the discussion above has linked the *relative* returns of the market portfolio and the individual assets comprising it. However, we have said nothing so far about where the *absolute* returns or price levels come from or, even more generally, why asset prices are random in the first place.

In the context of stocks, the classical argument for this is the following. A stock is a claim to all the dividends the underlying company generates in the future. There clearly is substantial uncertainty about the future profitability of the company, and this is revealed

only gradually over time. This in turn should translate into random prices that move to reflect the arrival of new information.

To illustrate this point, let us consider a simple one-period model where all uncertainty comes from a liquidating dividend D_t paid at the terminal time t. As the stock pays no more dividends after time t, it's time-t-price (before paying the dividend) has to be $P_t = D_t$ to avoid arbitrage. Its return from time $t-1$ to time t in turn is

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{D_t - P_{t-1}}{P_{t-1}}.$$

If we consider a market with a single risky asset for simplicity, then it follows from the CAPM relationship (2.4) that

$$\frac{\mathrm{E}_{t-1}[D_t] - P_{t-1}}{P_{t-1}} - R^f = \gamma \mathrm{Cap}_{t-1} \frac{\mathrm{Var}_{t-1}[D_t]}{P_{t-1}^2} = \gamma s \frac{\mathrm{Var}_{t-1}[D_t]}{P_{t-1}}.$$

As a consequence, the equilibrium price *level* at time $t-1$ is

$$P_{t-1} = \frac{1}{1+R^f}\Big(\mathrm{E}_{t-1}[D_t] - \gamma s \mathrm{Var}_{t-1}[D_t]\Big). \tag{2.5}$$

This makes sense intuitively: if there is no uncertainty about the dividend, then its time $t-1$ price is just this sure payoff, discounted at the risk-free rate. That is, the stock price is not random in this case. In contrast, with uncertainty, the stock price at time $t-1$ is given by the *expected* discounted dividend, minus a penalty for risk. This "discount" for risky assets with a given payoff is exactly what translates into the risk *premium* in expected returns.[5]

However, the above formula clearly displays one of the deficiencies of the mean-variance criterion considered here: prices can become negative for assets with positive payoffs if the representative agent's risk aversion is large enough. This paradoxical feature can be avoided by instead considering goal functionals that are increasing in wealth, see Section 4.2 below.

2.3 More General Linear Factor Models

Despite the theoretical underpinnings the CAPM provides for the simplest linear one-factor model (2.1), it is widely acknowledged that a number of other systematic factors other than market risk also affect expected returns. For example, stocks with small market capitalizations tend to outfperform large-cap stocks. To make exposure to this factor tradeable, one can sort a range of stocks by their market capitalizations, and then go long in small stocks and short in large ones. (Such long-short portfolios are used to reduce the exposure of the factor portfolio to market fluctuations.) One can then run a joint regression of the returns of a strategy against both the market portfolio and the "small-minus-big" (SMB) portfolio":

$$R_t^e = \alpha + \beta^M R_t^{M,e} + \beta^{SMB} R_t^{SMB,e} + \varepsilon_t.$$

[5] Of course, the current stock price can be observed today. Thus, the relationship (2.5) induces a constraint the distribution of dividends have to satisfy to match this. The use of endogenizing prices is that this allows to study what happens if fundamentals change. For example, if new information arrives that suggests expected dividends will be 10% higher than previously expected (with the same variance), then (2.5) predicts how the current market price will change to reflect this new information.

This in turn sheds light on how much of the strategy's excess returns over the market portfolio can be explained by overexposure to small-cap stocks. Another commonly used factor are "high-minus-low" (HML) portfolios, which long stocks with high book-to-market ratios and short stocks with low ones.[6] The joint regression on these three factors is known as the "Fama-French 3-factor model" (Fama and French, 1993), and is a common benchmark:

$$R_t^e = \alpha + \beta^M R_t^{M,e} + \beta^{SMB} R_t^{SMB,e} + \beta^{HML} R_t^{HML,e} + \varepsilon_t.$$

In the so-called "Carhart model" (Carhart, 1997), one adds a fourth factor that focuses on "momentum", i.e., goes long in stocks that have recently performed well, and short on stocks that have recently performed poorly.

In addition to these market and accounting variables, there is also a vast amount of research that tries to link asset returns to other "characteristics" such a index membership, industry membership, liquidity (see Section 3.2 below), option implied volatilities, bond-yield spreads, or "ESG". With widespread access to powerful computing technology, there is of course also no need to restrict to linear models of the above form. For example, machine-learning methods are compared to regression analyses in Cont and Sirignano (2019); Gu et al. (2020).

2.4 Risk Reward Ratios

In the above models for return attribution, positive alphas are good and negative alphas are bad. But is a larger alpha always better? Not necessarily – this depends on the amount of idiosyncratic risk one needs to take on to realize the additional profits. Moreover, the alpha generated by a strategy can also be scaled up by leveraging the strategy, i.e., borrowing extra money to invest into multiples of it. *Risk-reward ratios* address these issues. The basic idea is to compare the expected excess return $\mathrm{E}_{t-1}[R_t - R_t^f]$ generated by the strategy to the corresponding risk.

Sharpe Ratio The most well-known risk-reward measure is the *Sharpe Ratio (SR)*, which compares the expected excess returns of the strategy to to the corresponding standard deviations:

$$\mathrm{SR}_{t-1} = \frac{\mathrm{E}_{t-1}[R_t - R_t^f]}{\mathrm{Std}_{t-1}(R_t - R_t^f)}.$$

How does the Sharpe ratio change when a strategy is leveraged? Suppose we are choosing what part x_{t-1} of our wealth W_{t-1} to invest in a strategy with return $R_t = R_t^f + R_t^e$; the remaining funds $W_{t-1} - x_{t-1}$ then are invested at the risk-free rate R^f. Here, *leverage* corresponds to investing more than the available funds into the strategy, $x_{t-1} > W_{t-1}$. For example, $x_{t-1} = 2W_{t-1}$ means that we take out a loan of the same size as our funds. (Of course, this incurs extra borrowing and margin costs that we will discuss later.) The return on the combination between the strategy and the safe investment then is

$$\frac{x_{t-1}(1+R_t) + (W_{t-1} - x_{t-1})(1+R^f) - W_{t-1}}{W_{t-1}} = R^f + \frac{x_{t-1}}{W_{t-1}} R_t^e.$$

[6] The intuition for this is that "value stocks" with high book-to-market ratios should have higher returns going forward. This has worked well in the past, but strategies that bet on this have performed very poorly in recent years.

The corresponding Sharpe ratio thus is

$$\mathrm{SR}_{t-1}^{x_{t-1}} = \mathrm{sgn}\left(\frac{x_{t-1}}{W_{t-1}}\right)\mathrm{SR}_{t-1}^{1}.$$

Whence, as long as one goes long in the strategy ($x_{t-1} > 0$), the Sharpe ratio does not depend on the leverage involved at all. (If one shorts the strategy, the sign of the Sharpe ratio is flipped.) This is in stark contrast to the expected excess return, which can be scaled up by leveraging the strategy.

Information Ratio The Sharpe ratio gives the investor credit for all the returns the strategy generates on top of the risk-free rate. However, as we have discussed in the previous section, there is a difference between excess returns earned by just taking on exposure to market risk and other idiosyncratic alphas. This is addressed by the so-called *information ratio (IR)*, which focuses on the expected *abnormal return*:

$$\mathrm{IR} = \frac{\alpha}{\mathrm{Std}(\varepsilon_t)}.$$

Here, the alpha α and the idiosyncratic risk ε_t come from a regression of the strategy's excess return on some benchmark with excess return $R_t^{b,e}$:

$$R_t^e = \alpha + \beta R_t^{b,e} + \varepsilon_t.$$

Many investment funds have a mandate to beat a specific benchmark (e.g., the S&P500 index). Then, the IR is often computed without running a regression, by simply considering returns in excess of the benchmark:

$$\mathrm{IR} = \frac{\mathrm{E}[R - R^b]}{\mathrm{Std}[R - R^b]}.$$

The IR then measures how much the benchmark is outperformed per unit of tracking error risk. In the special case where the benchmark is the safe investment, we recover the Sharpe ratio.

2.5 Estimating Performance Measures

Expected returns are estimated as sample averages of realized returns. Some people use *geometric* averages

$$\text{geometric average } = [(1 + R_1) \times (1 + R_2) \times \ldots \times (1 + R_T)]^{1/T} - 1.$$

This corresponds to the experience of a buy-and-hold investor who neither adds nor removes capital over extended time periods. Others use the *arithmetic* average

$$\text{arithmetic average } = [R_1 + R_2 + \ldots + R_T]/T.$$

This is the optimal statistical estimator for the expected return; it corresponds to the experience of an investor who adds and removes capital so as to keep a constant dollar exposure to the strategy.

No matter which estimate one uses, it is very important to keep in mind that estimates for expected returns are *extremely noisy*. The quality of the estimation improves with the length of the sample period (however, not with the frequency at which returns are observed!). Then again, it is unrealistic to assume that expected returns remain constant over very long sample periods.

The standard deviation of expected returns can be estimated with much more precision. The usual estimator is the square root of the sample variance, i.e., the squared deviations from the arithmetic mean $\bar{R}$:

$$\text{variance estimate } = [(R_1 - \bar{R}^2 + (R_2 - \bar{R})^2 + \ldots + (R_T - \bar{R})^2]/(T-1).$$

Covariances between assets can be estimated analogously from the sample covariances. This works reasonably well for small numbers of assets, but also causes problems when many assets are considered. For example, the sample covariance matrix need not even be positive semidefinite in general.

2.6 Time Horizons for Performance Measures

All of the above performance measures depend on the time horizon over which they are measured. That is, they change when one switches from daily, to monthly, or yearly returns. For example, the same strategy typically does *not* have the same Sharpe ratio if measured over a day or a year. To illustrate this, suppose that the daily excess returns $R_1^e, R_2^e, \ldots$ are iid with some mean μ and variance σ^2. Then, the daily Sharpe ratio is

$$\text{SR}^{\text{daily}} = \frac{\text{E}[R^e]}{\text{Std}[R^e]} = \frac{\mu}{\sigma}.$$

However, since we have assumed returns to be iid, the yearly Sharpe ratio is (assuming 252 trading days)

$$\text{SR}^{\text{yearly}} = \frac{\text{E}[\sum_{t=1}^{252} R_t^e]}{\sqrt{\text{Var}[\sum_{t=1}^{252} R_t^e]}} = \frac{252\text{E}[R_t^e]}{\sqrt{252\text{Var}[R_t^e]}} = \frac{\mu}{\sigma}\sqrt{252}.$$

Hence, the Sharpe ratio is higher when measured over longer time intervals. When talking about performance measures, it is therefore important to clarify what is the time horizon that is referred to. To this end, measurements are often *annualized*, that is, either measured from yearly data or, more commonly, converted into annual units. Since real returns are typically close to independent, the Sharpe ratio is commonly adjusted as above, for example.

Another important effect regarding time horizons is that they crucially affect how often one experiences profits and losses. If you observe the P&L of your strategy more frequently (e.g., if you are a hedge fund manager with a live screen), then you have a lower Sharpe ratio between each time you look at the P&L and the risk is perceived as larger.

For example, if returns are approximately normally distributed (a reasonable first-order approximation, even through they are much more heavy tailed in reality), then the loss

probability (relative to the risk-free rate) can be written as follows, for a standard normal random variable Z:

$$\mathbb{P}[R^e < 0] = \mathbb{P}\Big[\mathrm{E}[R^e] + \mathrm{Std}[R^e]Z < 0\Big] = \mathbb{P}[Z < -\mathrm{SR}].$$

Whence, the probability of losing money only depends on the Sharpe ratio in this case. More specifically, suppose returns are iid on all time scales (e.g., they follow Brownian motion as in the Black-Scholes model), and consider a strategy with a fixed yearly Sharpe ratio. Then, the Sharpe ratio at frequency Δt years is $\sqrt{\Delta t}$-times the yearly Sharpe ratio. For short time intervals, it this tends to zero and the probability of losses goes to one half – for short time intervals, the random fluctuations completelty drown out even a very large mean. Whence, observed at sufficiently high frequencies, losses occur almost every other minute, even when you are having a great year!

2.7 High-Water Marks and Drawdowns

Another important characteristic of strategies, in particular for hedge funds, is their performance relative to their past maximum. Indeed, many hedge funds only charge performance fees on when the fund performance exceeds the *high water mark*,[7] the highest value the fund has achieved in the past:[8]

$$\mathrm{HWM}_t = \max_{s \le t} P_s.$$

Since investors often withdraw their money when the fund drops too far below the high-water mark, an important risk measure in this context is the (relative) *drawdown*

$$\mathrm{DD}_t = \frac{\mathrm{HWM}_t - P_t}{\mathrm{HWM}_t}.$$

These concepts are illustrated in Figure 2, taken from Pedersen (2015).

3 Finding and Backtesting Strategies

In the previous section, we have discussed how to assess the performance of a given strategy. Now we turn to the problem of coming up with potentially profitable strategies in the first place. Clearly, in a competitive market, one will very rarely be able to find a literal *arbitrage* delivers a profit without taking any risk. However, some strategies have delivered profits more often than losses over extensive time periods. Why is that? One reason is of course luck. But the goal is to find strategies that will continue to make money going forward and for that, it is important to understand the underlying mechanisms.

In particular, it is crucial to understand who is taking the other side of the trades involved, and why. If your strategy is making money more often than not, then who is trading with you and why will they continue to do so in the future? This is particularly

[7] The typical compensation scheme is "two and twenty" – a management fee of 2%, combined with a performance fee of 20% on gains relative to the high-water mark.

[8] Here, P_t is the cumulative return computed via $P_t = (1 + R_t)P_{t-1}$.

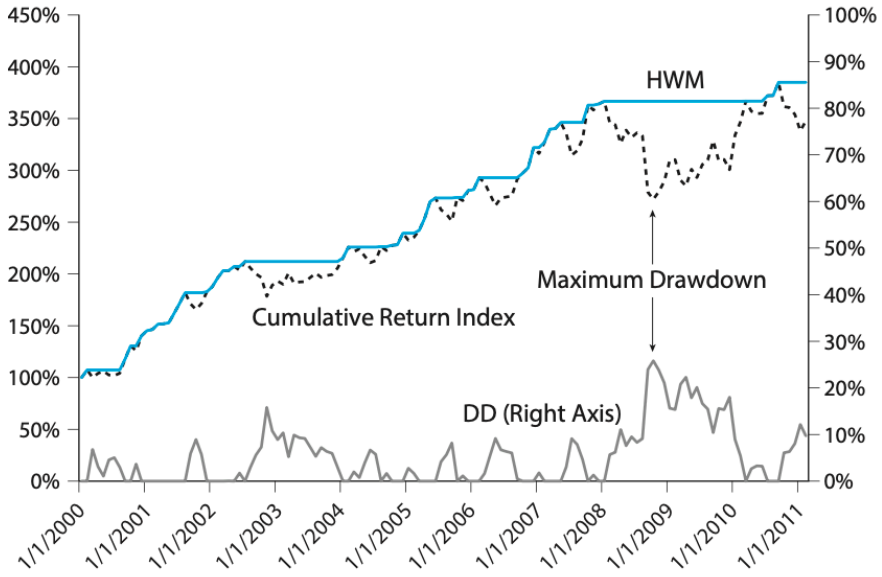


Figure 2: A hedge-fund strategy's high-water mark (HWM) and drawdown (DD).

important for (often market-neutral) active strategies, where every buyer needs to find a seller. In contrast, since stocks are in positive net supply, everyone can buy and hold a passive index fund and profit from the general growth of the market.

Two basic sources for repeatable trading profits are compensation for liquidity risk and informational advantages. We now discuss these in more detail.

3.1 Efficiently Inefficient Levels of Information

In order for market prices to convey accurate information about the fundamental values of the traded assets, someone must collect the relevant information and trade on it. This role is often played by hedge funds.

Why can information production lead to a repeatable source of profits? To see this, first recall that markets cannot be perfectly efficient and always display all information. Otherwise, no one would have an incentive to actually collect the information in the first place and implement the trades that make sure it is reflected in prices. Conversely, markets also cannot be very inefficient, because hedge funds and other active investors would then trade more and more to exploit the available profit opportunities. Instead, markets must be inefficiently efficient, in that it is difficult to make money, but not so difficult that no one collects information to try to do this.

Production of Information One classic way in which hedge funds and other investors produce information is through *fundamental analysis*. This means that an extensive analysis is carried out of companies and their future profit opportunities, e.g., by dissecting their balance sheets, researching consumer dynamics, studying competition in the firm's sector, etc.

Access to Information Another source of profits is access to superior information. The most drastic (and illegal!) example is insider trading, e.g., information about earnings that has not been made public yet, or the briefings US senators received about the COVID epidemic before the public. But there are also many legal ways to acquire informational edges ranging from analyzing satellite pictures of parking lots, calling doctors about what medicines they are prescribing, to purchasing ultra-fast transatlantic cable connections to have access to price changes at exchanges abroad before the general public.

Behavioral Finance and Limits of Arbitrage A third source of profits is that news and other public information often do not get fully reflected in prices right away. For example, after an unexpectedly positive earnings announcement prices move up but, on average, continue to drift up for several weeks afterwards. This is an example of a general tendency for initial underreaction and eventual overreaction to new information, which creates trends and momentum that form the basis of many quantitative investment strategies.

Why do these effects arise? One explanation is that (some) investors suffer from behavioral biases and make systematic mistakes that push prices away from fundamentals. If other, more sophisticated, investors are constrained in the amount of capital they can deploy or the risk they can take, then they may not be able to "arbitrage away" such trading opportunities completely.

3.2 Efficiently Inefficient Compensation for Liquidity Risk

Another basic reason for positive expected returns is that they are a compensation for risk. The simplest example is the market risk that takes center stage in the CAPM – if this would not be compensated, then no investor would buy stocks rather than holding government bonds.

Of course, no active investment is needed to gain exposure to market risk – one can just buy a passive index fund. However, active investors can also earn compensation for other risks; a prime example is *liquidity risk*. This consists of the risk of rising transaction costs ("market liquidity risk"), the risk of running out of cash, in particular for leveraged funds ("funding liquidity risk"), and the risk of accommodating "demand pressure".

Liquidity risk is another important limit of arbitrage, in that it limits traders' ability to exploit and thereby eliminate mispricings that they observe. However, liquidity risk not just limits the reversal of prices to fundamentals if something else has caused this deviation. Instead, liquidity risk generates systematic *liquidity premia*, in analogy to risk premia in the CAPM.

Market Liquidity Risk Some securities are much less liquidly traded than others, and the differences tend to be exacerbated in times of crisis. For example, during the financial crisis of 2008, bid-ask spreads of convertible bonds went from less than 1% to more than 5% of the midprices, and many assets even had no quoted bid prices at all anymore.

It seems natural that – like for market risk that cannot be diversified away – investors should demand a premium for holding assets that are more difficult to trade, in particular, when one may need to do this quickly in times of crisis. How can this be incorporated in a simple *liquidity-adjusted CAPM*?

Liquidity-Adjusted CAPM Of course, trading costs play no role if we only consider a representative agent who just holds the market portfolio forever and never trades. Instead, we assume as in Acharya and Pedersen (2005) that at each time $t-1$ a new (representative) agent is born with some cash endowment (think of labor income earned and invested into a retirement account while young), that she uses to purchase a safe asset with return R_f and risky assets with returns $R^f + R_t^{n,e}$, $n = 1, \ldots, N$ to be determined. At time t, the time-$(t-1)$ agent reaches "retirement age" and sells her holdings in the risky asset to the time-t agent at current market prices P_t^n, minus a transaction cost C_t^n. (For simplicity, we assume that the full cost is always paid by the seller.) This reduces the returns of the risky assets from

$$R_t^n = \frac{P_t^n - P_{t-1}^n}{P_{t-1}^n} \quad \text{to} \quad R_t^{n,\text{net}} = \frac{P_t^n - P_{t-1}^n - C_t^n}{P_{t-1}^n} = R_t^n - c_t^n,$$

where the "relative transaction costs" c_t^n are defined as

$$c_t^n = \frac{C_t^n}{P_{t-1}^n}.$$

Likewise, if s^n denotes the fixed number of outstanding shares of risky asset n, then transaction costs reduce the return of the market portfolio from

$$R_t^M = \frac{\sum_n s^n P_t^n - \sum_n s^n P_{t-1}^n}{\sum_n s^n P_{t-1}^n}$$

to

$$R_t^{M,\text{net}} = \frac{\sum_n s^n P_t^n - \sum_n s^n P_{t-1}^n - \sum^n s^n C_t^n}{\sum^n s^n P_{t-1}^n} = R_t^M - c^M,$$

where the "relative market transaction cost" is defined as

$$c_t^M = \frac{\sum_n s^n C_t^n}{\sum_n s^n P_{t-1}^n}.$$

Whence, this simple liquidity-adjusted CAPM is equivalent to a standard CAPM, where the returns of all risky assets and the market are reduced by appropriate (relative) transaction costs. In particular, the standard CAPM argument yields that the equilibrium return (that makes the representative agents to hold the market portfolio) *net of transaction costs* are related to the risk premium earned by the market portfolio by the standard CAPM relationship:

$$\mathrm{E}_{t-1}[R_t^{\text{net}}] = R^f + \beta_{t-1}^n \mathrm{E}_{t-1}[R_t^{M,\text{net}} - R^f], \quad \text{where } \beta_{t-1}^n = \frac{\mathrm{Cov}_{t-1}[R_t^{M,\text{net}}, R_t^{n,\text{net}}]}{\mathrm{Var}_{t-1}[R_t^{M,\text{net}}]}.$$

We now want to rewrite this in a form that highlights the contributions of the *gross* returns and transaction costs. With the relative transaction costs of the individual assets and the market portfolio introduced above, we have

$$\mathrm{E}_{t-1}[R_t^n] = R^f + \mathrm{E}_{t-1}[c_t^n] + \frac{\mathrm{Cov}_{t-1}[R_t^M - c_t^M, R_t^n - c_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]} \mathrm{E}_{t-1}[R_t^M - R^f - c_t^M].$$

Whence, the equilibrium return of each asset is composed of three parts. The first is the risk-free rate R^f, like in the standard CAPM. The second is the expected trading cost for the respective asset – ceteris paribus, less liquid assets need to earn higher returns. The last term is the product of the risk premium of the market portfolio (net of the market transaction cost), times a liquidity-adjusted beta. The (net) market risk premium is the same for each asset, but the multiplier of this term depends on the interplay between transaction costs and gross returns. To wit, we can expand the covariance term to obtain a decomposition into four distinct betas:

$$\begin{aligned}\frac{\mathrm{Cov}_{t-1}[R_t^M - c_t^M, R_t^n - c_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]} =& \frac{\mathrm{Cov}_{t-1}[R_t^M, R_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]} + \frac{\mathrm{Cov}_{t-1}[c_t^M, c_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]} \\ &- \frac{\mathrm{Cov}_{t-1}[R_t^M, c_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]} - \frac{\mathrm{Cov}_{t-1}[c_t^M, R_t^n]}{\mathrm{Var}_{t-1}[R_t^M - c_t^M]}.\end{aligned}$$

The first term is analogous to the beta from the standard CAPM, which suggests that the expected returns are high for assets that have substantial exposure to market risk. The second term implies that this risk premium is increased for assets whose transaction costs are high when market-wide transaction costs are high – that is, assets that are particularly difficult to trade when market-wide liquidity dries up. The third term suggests that investors are willing to accept a lower return for securities which are easy to trade in down markets. The fourth term implies that investors accept lower returns for securities that offer high returns when the market as a whole is illiquid.

In summary, traders only want to buy illiquid assets only when they get compensated for this – by a low price or, equivalently, large expected returns. Thus, hedge funds (or the endowment funds of American universities, for example) willing to be exposed to substantial liquidity risk can earn the corresponding liquidity premia. This may be a reasonable strategy for funds with a long planning horizon. However, this involves risks – such as forced liquidations during crises – which is exactly what the funds are compensated for.

The empirical performance of the liquidity-adjusted CAPM (for a specific proxy of liquidity) is compared to the standard CAPM in Figure 3 from Acharya and Pedersen (2005). We see that the liquidity adjustment improves the fit especially for the illiquid portfolios (with high returns), consistent with what intuition would suggest. Also note that the number of free parameters is the same in both models (the market risk premium and, potentially, an alpha term), so the improvement in fit is not a consequence of more degrees of freedom. Liquidity risk therefore is one reason why the empirical security market line is flatter than predicted by the CAPM: high returns are not just a compensation for large market exposure, but also for liquidity risk.

The liquidity-adjusted CAPM also provides some insight into what happens in a *liquidity crisis*. When transaction costs and liquidity risk suddenly increase, the required return increases, forcing prices to drop sharply.

Market Making Trading illiquid securities is one way to earn market liquidity premia. Another is *market making*. To wit, many investors want to trade immediately, but buyers and sellers are not always present in the market at the same time. Market makers (*liquidity providers*) step in and smooth out supply-demand imbalances. As a compensation for these

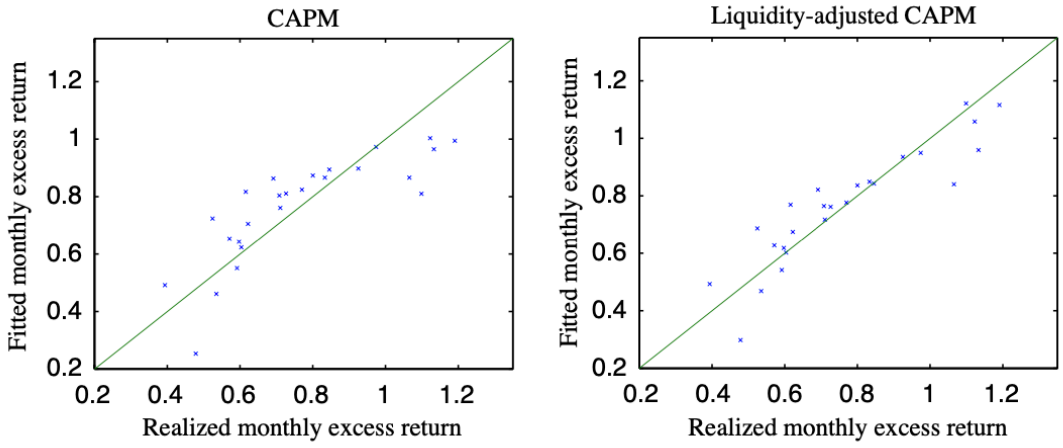


Figure 3: Left panel: fitted CAPM returns vs. realized returns using monthly data from 1964–1999 for value-weighted illiquidity portfolios. Right panel: the same for the liquidity-adjusted CAPM.

liquidity services, they earn higher "ask" prices for buying than the "bid" prices they pay for selling the same securities.

Let us discuss in a bit more detail how this works in a simple stylized model, where we look for the equilibrium price at which a liquidity provider absorbs a given incoming order flow by investors who want to trade immediately. We consider a liquidity provider who, before trading at time $t-1$, holds a cash position C_{t-1} and φ_{t-1} shares of a risky asset. Suppose the liquidity provider is small,[9] and takes the market price P_{t-1} quoted at time $t-1$ as given and simply decides how many shares $\Delta\varphi_{t-1}$ of the risk asset to buy or sell at this price. After trading, the liquidity provider holds $\varphi_t = \varphi_{t-1} + \Delta\varphi_{t-1}$ risky shares and has a cash position of $C_t = C_{t-1} - \Delta\varphi_{t-1}P_{t-1}$. For simplicity, let us suppose the risk-free rate is zero and the fundamental value P_t of the asset is fully revealed at time t.[10] Then, the liquidity provider's final wealth is

$$W_t^{\Delta\varphi_{t-1}} = C_t + \varphi_t P_t = C_{t-1} - \Delta\varphi_{t-1}P_{t-1} + (\varphi_{t-1} + \Delta\varphi_{t-1})P_t.$$

Suppose the liquidity provider maximizes the following mean-variance functional of her wealth:

$$\begin{aligned} J_{t-1}(\Delta\varphi_{t-1}) &= \mathrm{E}_{t-1}\left[W_t^{\Delta\varphi_{t-1}}\right] - \frac{\gamma}{2}\mathrm{Var}_{t-1}\left[W_t^{\Delta\varphi_{t-1}}\right] \\ &= C_{t-1} - \Delta\varphi_{t-1}P_{t-1} + (\varphi_{t-1} + \Delta\varphi_{t-1})\mathrm{E}_{t-1}[P_t] - \frac{\gamma}{2}(\varphi_{t-1} + \Delta\varphi_{t-1})^2\mathrm{Var}_{t-1}[P_t]. \end{aligned}$$

[9] More specifically, suppose we have many small liquidity providers of the same form.

[10] One can also study multiperiod versions of this model, where uncertainty about the fundamental value is revealed gradually, but this is considerably more challenging. The main insights are similar to the one-period model, however.

(Here, we have used in the second step that the number $\Delta\varphi_{t-1}$ of shares traded at time $t-1$ is chosen based only on information available then.) The maximum of this quadratic function is characterized by the following first-order condition:

$$0 = J'_{t-1}(\Delta\varphi_{t-1}) = -P_{t-1} + \mathrm{E}_{t-1}[P_t] - \gamma(\varphi_{t-1} + \Delta\varphi_{t-1})\mathrm{Var}_{t-1}[P_t].$$

As a consequence, the liquidity provider's optimal trade at time $t-1$ is

$$\Delta\hat{\varphi}_{t-1} = \frac{\mathrm{E}_{t-1}[P_t] - P_{t-1}}{\gamma\mathrm{Var}_{t-1}[P_t]} - \varphi_{t-1}. \tag{3.1}$$

To wit, if the risky asset offers no expected return ($\mathrm{E}_{t-1}[P_t] = P_{t-1}$), then the risk-averse liquidity providers simply liquidate their entire risky position. In general, their optimal trade is a combination between trying to offset the risk of keeping incoming positions and earning the corresponding risk premia.

Now suppose another investor comes to the market at time $t-1$ who needs to trade δ_{t-1} shares of the risky asset immediately, irrespective of the quoted market price and the fundamental value of the asset.[11] What is the market price at which this "demand pressure" is absorbed by the liquidity providers? In view of the formula (3.1) for the liquidity provider's optimal trade, we need

$$\delta_{t-1} = \Delta\hat{\varphi}_{t-1} \quad \rightsquigarrow \quad P_{t-1} = E_{t-1}[P_t] - \gamma\mathrm{Var}_{t-1}[P_t]\varphi_{t-1} - \gamma\mathrm{Var}_{t-1}[P_t]\delta_{t-1}.$$

We see that prices for buying and selling are symmetric around the *midprice*

$$M_t = E_{t-1}[P_t] - \gamma\mathrm{Var}_{t-1}[P_t]\varphi_{t-1},$$

which depends on the liquidity provider's forecast of the fundamental value, shifted by a risk correction depending on the liquidity provider's incoming position. The intuition for this is that, if they already have a large risky positions, liquidity providers want to sell. The price in turn needs to decrease to entice them to continue to hold their positions. In particular, if liquidity providers never trade with liquidity takers but simply hold the entire supply s of the risky asset over time, then we recover the CAPM, where the (absolute) return has expectation $\mathrm{E}_{t-1}[P_t - P_{t-1}] = \gamma s\mathrm{Var}_{t-1}[P_t - P_{t-1}]$.

For trades with liquidity takers, the "bid-ask spread" between the selling and buying prices for one share of the risky asset is

$$S_t = 2\gamma\mathrm{Var}_{t-1}[P_t].$$

In the model considered here, the only risk incurred by providing liquidity is the danger of adverse price changes, and the bid-ask spread compensates liquidity providers for this. For larger trades, the price charged per share increases linearly in trade size, so that the terms of trade become worse and worse for larger trades. This "price impact" of large trades is a crucial concern for large hedge funds; we will come back to this later in the course.

[11] E.g., because they are forced to liquidate their portfolio or have to sell a stock because it has dropped out of the S&P500 index. More complex versions of the model where liquidity takers are react to market prices and fundamental values are of course an important extension of the model.

Funding Liquidity Risk Another important risk hedge funds take on is *funding liquidity risk*, i.e., the risk that they become unable to fund a leveraged position throughout the lifetime of a trade. Put differently, this is the risk of being forced to unwind a position because the fund hits its "margin constraint" or gets uncomfortably close.

What exactly does this mean? To this end, we need to briefly look into how leverage is implemented in reality, i.e., how investors buy assets worth more than the cash they have at hand. (The mechanisms are similar for short positions that correspond to owning a negative amount of an asset.)

Suppose you have no cash at hand but want to buy 1 million bonds at a price of \$100 each. To obtain a loan for this transaction, you try to use the bonds as a collateral, just like when taking out a mortgage when buying a house. However, just like a fraction of the total price is typically required as a down payments when purchasing a house, you can also only borrow a fraction of the total value against the bond collateral, say \$90 per bond. The difference, \$10 per bond in this example, is called the *haircut* or the *margin requirement*. The haircut gives the lender an extra margin of safety in case the value of the bond suddenly drops and you do not want to pay the lender back. In this case, the lender can sell the bond and recover the loan, as long as the bond value is at least \$90, i.e., as long as the price drop is smaller than the haircut. (Again, this is exactly the same for mortgages.) In summary, to buy \$100 million worth of bonds, you do not need the full cash amount, but you do need \$10 million for the margin requirement. Thus, in this example, you can build up a maximal leverage of 10.

Now suppose you instead want to *shortsell* 1 million bonds. In this case, you need to borrow the *securities*, and then sell them. Later, say the next day, you can then buy back the securities and deliver them back to the lender. Of course, you hope that the price will drop, so that you can buy back the bonds for less than you sold them. This way, you effectively have a negative position, because you profit from price drops and lose money when prices increase. When a hedge fund shortsells securities, its broker keeps the sale proceeds and, in addition, asks for an additional margin requirement.

In each case, margin requirements are typically set so that some "worst-case" price move is covered with a certain confidence. Of course, expectations about this change over time. Whence, the margin requirements are updated dynamically. To wit, the positions of a hedge fund are "marked-to-market" each day using the current market prices. The margin accounts for each security are then credited or debited the respective price changes (as well as for interest rate payments). Whence, when implementing leverage or shortselling, you need to continuously monitor your positions and make sure your cash levels are above the minimum margin requirement. If a hedge fund has insufficient cash in its margin account (e.g., because of losses on its positions), then it receives a *margin call* from its broker. This means that either cash needs to be added to the account, or positions need to be reduced. If the fund does not do one or the other, the broker will liquidate the positions. Even if addressed successfully, repeated margin calls send out a negative signal to investors and can eventually lead brokers to increase margin requirements or even terminate the brokerage relationship. Hence, hedge funds naturally keep extra margin capital.

How are asset prices affected by how easy or difficult it is to fund a security? If it is difficult and expensive to fund investment in an asset, then it is natural to expect that investors

demand compensation for holding such "cash-intensive" securities. Therefore, required returns should increase with the margin requirement. In a variation of the CAPM (Garleanu and Pedersen, 2011), this leads to the linear relationship

$$\mathrm{E}_{t-1}[R_t^n] = R^f + \beta^n \mathrm{E}_{t-1}[R_t^M - R_t^f] + m^n \psi.$$

Here, m^n is the margin requirement of asset n and ψ is the compensation for tying up capital.

Another implication of funding and leverage constraints is that many investors prefer to buy risky securities over leveraging safer alternatives. This helps to explain why riskier securities tend to offer lower risk-adjusted returns than safer ones within in each asset, e.g., a portfolio of risky stocks tends to underperform a leveraged portfolio of safer stocks. This underlies what is called "risk-parity investment", where one does not invest an equal *fraction of wealth* in each asset, but instead equalizes *risk* across different investments.

Finally Illiquid assets with high transaction costs tend to also be difficult to finance, and vice versa.

3.3 How to Backtest a Trading Strategy

Once you have an idea for a trading strategy, it is important to *backtest* it, i.e., simulate how it would have performed historically. Of course, past performance does not necessarily predict future performance. But backtesting is nevertheless very useful to discard trading strategies that would have not even worked in the past.

Inputs To run a backtest, you need the following inputs:

- The universe of securities to be traded.
- The data to be used.
- The trading rule to be applied, which describes how frequently positions are reviewed and how they are rebalanced.
- Time lags: to make the strategy implementable, the data used in the trading rule must actually be available at the time when it is used. However, GDP or unemployment data is published with a substantial time lag, one cannot trade on closing prices, etc.

Data Mining and Biases Backtests typically looks a *lot* better than the actual performance of a trading strategy. This is to be expected for a number of reasons. First, unlike physical laws, financial markets keep changing so that a trading opportunity that existed in the past may disappear in the future. In particular, unlike physical systems, financial markets are influenced by the actions of the market participants. Whence, the discovery of a new inefficiency may in turn lead to its elimination as more and more people trade on it.

This nonstationarity and reflexivity of financial markets are unavoidable. A different and even more important bias suffered by *all* backtests is data mining. This means that whenever one is analyzing a dataset, one naturally tends to gravitate towards the hypotheses

supported by it (unless these are fixed before looking at the data, which is the – albeit rarely followed – gold standard in the experimental sciences). In the context of portfolio management, one (consciously or subconsciously) ends up favoring those trading ideas that have worked well in the past – even though one could of course not have known back then that this particular strategy would go on to do well going forward. The same bias occurs if you backtest a strategy because you heard that someone else made a lot of money with it – again, you would not have had access to this information in real time.

As a concrete example, you can try comparing the past performance of the *current* stocks in the S&P500 index to the historical performance of the index. Stocks tend to be included in this large-cap index after they perform particularly well, but one of course does not know this before hand. The same problem arises if one sorts stocks by characteristics averaged over the whole time series. For example, we have mentioned earlier that small-cap stocks tend to outperform stocks with larger market capitalizations. But this result is completely reversed if one sorts by markets capitalizations over the whole dataset (rather than a moving window), because the large-cap portfolio will then tend to include all the stocks whose prices increased a lot over the sample period. Another typical example for this is the use of regression analysis to obtain trading signals. Indeed, if the regression parameters are estimated over the whole dataset (rather than a backward-looking rolling window), than the performance is bound to look unrealistically good. This is because the parameter estimates are optimal for the *whole* dataset, including the future datapoints that one is trying to predict.

Even if one avoids these forward-looking biases, backtests typically make strategies look too good, because they make us focus our attention on those strategies that did well by chance. This is a particularly serious problem for strategies with many tuning parameters. Indeed, even if all of these have a purely random impact on performance, some combinations are bound to generate impressive results purely by chance.[12] This is a typical example for the difference between *in-sample* and *out-of-sample* tests. For the latter, one typically splits up the available dataset into two parts. The first is used to tune the parameters of the strategy, the second is then used to test its performance on a new set of data on which the strategy has not been optimized yet.

However, if one repeats this procedure too many times, the out-of-sample set becomes the new in-sample, as one just ends up data mining this new dataset as well. This highlights the importance of understanding the economic mechanisms underlying prospective trading strategies. As data analysis alone can lead to overfitting, given a finite data budget, one should use theory to guide the empirical questions to investigate. You want to spend your data budget intelligently!

Since the P&L of a hedge fund ultimately depends on robust out-of-sample performance, successful hedge funds keep careful track of how many different specifications of trading strategies they have already tested. However, data mining is often used to convince potential new customers to invest in a new fund (by trying many different trading strategies and only advertising the ones that ended up performing best, for example). Similarly, many academic

[12]The same problem appears in all empirical sciences, For example, neuroscientists refer to "voodoo correlations" (Vul et al., 2009), when talking about the high – but purely random – correlations that appear when studying the correlations between different parts of the human brain on finer and finer resolutions. This leads to exactly the same problems as backtesting more and more variations of a trading strategy.

studies suffer from the same problems, because highly significant positive results are much easier to publish. When tested out of sample, many relationships that appeared to be highly significantly in sample then disappear. See, e.g., Welch and Goyal (2008); Linnainmaa and Roberts (2018); Hou et al. (2020) for a detailed discussion.

Adjusting Backtests for Trading Costs Another issue that is often neglected and makes backtesting results look suspiciously impressive is the impact of trading costs. As discussed in the section on market making above, the prices liquidity takers pay for buying securities are systematically higher than the proceeds they receive for selling the same assets, and this wedge becomes wider the larger the traded positions are. As a consequence, the real performance of a trading strategy will tend to be worse than what would be estimated from "paper trades" at the quoted midprices, and the effect can be substantial if the strategy requires to frequently turn over large positions.

Backtests should therefore be adjusted for trading costs whenever possible. This requires accurate estimates for transaction costs and, ideally, should lead to adjustments of the trading process that take these into account. We will come back to these points in Section 5.

4 Portfolio Construction and Risk Management

Suppose you have identified a number of promising investment opportunities, either individual assets that look attractive or composite strategies that appear to produce a strong performance. The question then is how to combine these into an overall investment *portfolio*. This requires to (i) estimate the risk and rewards associated with each investment and the dependencies between them and (ii) choose how to size the different positions in order to achieve the best overall risk-reward tradeoff.

For active portfolios such as the holdings of most hedge funds, *risk management* also is of crucial importance. Indeed, risks vary substantially over time, as investment opportunities, price volatilities, and correlations between different trading strategies change. Controlling the resulting risks is particularly important for funds that employ leverage or shortselling, since these cannot simply "ride out" a prolonged drawdown.

Let us start with some general guidelines for portfolio construction used by most successful investors:

- The most important principle is *diversification* which, as the saying goes, is "the only free lunch in finance".
- A second related principle are position limits on each individual investment.
- A third paradigm is that correlations matter a *lot*, and are often underestimated. A classical example is the clustering of defaults of normally uncorrelated companies in times of crisis.

4.1 Mean-Variance Optimization

The simplest way to formalize these ideas is the mean-variance framework we have already looked at a number of times in the context of the CAPM. Starting from wealth W_{t-1} at

time $t-1$, we are then choosing what amounts $x_{t-1} = (x_{t-1}^1, \ldots, x_{t-1}^N)$ to invest in risky assets $n = 1, \ldots, N$. Once the returns $R_t = (R_t^1, \ldots, R_t^N) = R^f + R_t^e$ from time $t-1$ to time t are realized, the resulting time-t wealth then is

$$W_t^{x_{t-1}} = W_{t-1}(1 + R^f) + x_{t-1}^\top R_t^e.$$

The corresponding (one-period) mean-variance goal functionals in turn is

$$\begin{aligned} J_{t-1}(x_{t-1}) &= \mathrm{E}_{t-1}\left[W_t^{x_{t-1}}\right] - \frac{\gamma}{2}\mathrm{Var}_{t-1}\left[W_t^{x_{t-1}}\right] \\ &= W_{t-1}(1 + R^f) + x_{t-1}^\top \mu_{t-1} - \frac{\gamma}{2} x_{t-1}^\top \Sigma_{t-1} x_{t-1}, \end{aligned} \tag{4.1}$$

where

$$\mu_{t-1} = \mathrm{E}_{t-1}[R_t^e], \quad \Sigma_{t-1} = \mathrm{Cov}_{t-1}[R_t^e, R_t^e].$$

The optimal risky investments then are

$$\hat{x}_{t-1} = (\gamma \Sigma_{t-1})^{-1} \mu_{t-1}. \tag{4.2}$$

Two-Fund Separation and CAPM Revisited In view of (4.2), as long as mean-variance investors agree on expected excess returns $\mu_{t-1} = \mathrm{E}_{t-1}[R_t^e]$ (a big assumption) and their covariances $\Sigma_{t-1} = \mathrm{Cov}_{t-1}[R_t^e, R_t^e]$, then they all invest in the same risky portfolio, in that the relative proportions of funds allocated to the different assets are the same. The agents' individual risk aversions only influence how investments are split between this "mutual fund" and the risk-free asset (this is called "two-fund separation").

In particular, the aggregate risky investments made by investors with risk aversions γ^m, $m = 1, \ldots, M$ are

$$\sum_{m=1}^{M} (\gamma^m \Sigma_{t-1})^{-1} \mu_{t-1} = (\gamma^{\mathrm{rep}} \Sigma_{t-1})^{-1} \mu_{t-1},$$

where

$$\gamma^{\mathrm{rep}} = \left(\sum_{m=1}^{M} \frac{1}{\gamma^m} \right)^{-1}.$$

Whence, the aggregate investment of all individual agents can indeed be subsumed by a "representative agent" with risk aversion γ^{rep}.[13] Thus, if all investors have mean variance-preferences, the same probability beliefs, and invest in the same universe of assets, then the CAPM holds, and expected excess returns are of the form

$$\mu_{t-1} = \gamma^{\mathrm{rep}} \Sigma_{t-1} \mathrm{Cap}_{t-1}.$$

As a consequence, not just the representative agent but also each individual agent m only invests into the risky assets through the market portfolio:

$$\hat{x}_{t-1}^m = \frac{\gamma^m}{\gamma^{\mathrm{rep}}} \mathrm{Cap}_{t-1}.$$

[13] The "risk tolerance" $1/\gamma^{\mathrm{rep}}$ of the presentative agent then is the sum of the individual agents' risk tolerances $1/\gamma^m$.

Whence, in a CAPM world absent of trading and funding frictions, all agents should just mix safe assets and a broadly diversified market portfolio according to their risk preferences. Another crucial assumption for this result is that all agents share the same beliefs about expected returns and covariances. We will come back to this point later in Section 6.1 and discuss how heterogeneous beliefs about fundamentals affect these results.

It is crucial to observe that mean-variance optimization and the CAPM are *not* the same thing. Indeed, mean-variance optimization is one (of many) criteria individual investors may want to use to choose their portfolio allocations. In contrast, the CAPM makes equilibrium predictions about the behaviour of the entire market.

Diversification in Mean-Variance Optimization How is diversification reflected in mean-variance optimization? To illustrate this, consider a market with risk-free rate $R^f = 0$ (for simplicity) and two risky assets. The first risky asset has expected return, $\mu_1 = 5\%$ and low volatility $\sigma_1 = 20\%$, so that simply investing in this asset only yields a Sharpe ratio of $\mathrm{SR}_1 = \mu_1/\sigma_1 = 1/4$. The second risky asset has the same expected return $\mu_2 = 5\%$, but a higher volatility of $\sigma_2 = 30\%$, so that its Sharpe ratio is only $\mathrm{SR}_2 = \mu_2/\sigma_2 = 1/6$.

Since risky asset 1 is clearly "better" does that mean that we should disregard risky asset 2 completely or even short sell it to make a larger bet on risky asset 1? Not necessarily, and the answer depends on the dependence between the risky returns. Indeed, first suppose they are independent so that $\Sigma_{t-1} = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$. The optimal risky investments for an investor with risk aversion γ in turn are

$$x_{t-1}^1 = \frac{\mu_1}{\gamma\sigma_1^2} = \frac{0.05}{\gamma 0.04} = \frac{1}{\gamma} \times \frac{5}{4},$$

and

$$x_{t-1}^2 = \frac{\mu_2}{\gamma\sigma_2^2} = \frac{0.05}{\gamma 0.09} = \frac{1}{\gamma} \times \frac{5}{9}.$$

Whence, the optimal portfolio does not neglect the less attractive asset, but allocates a positive (albeit smaller) fraction of the available capital to it to diversify the risk across the two independent investments.

How does this result change if the returns of the risky assets are correlated? Suppose the correlation is $\rho \in [-1, 1]$, so that the covariance matrix becomes

$$\Sigma_{t-1} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \rightsquigarrow \quad \Sigma_{t-1}^{-1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}. \tag{4.3}$$

As a result, the optimal investments are

$$\begin{aligned} \hat{x}_{t-1}^1 &= \frac{1}{\gamma} \left(\frac{\sigma_2^2}{(1-\rho^2)\sigma_1^2\sigma_2^2}\mu_1 - \frac{\rho\sigma_1\sigma_2}{(1-\rho^2)\sigma_1^2\sigma_2^2}\mu_2 \right) \\ &= \frac{1}{\gamma\sigma_1} \left(\frac{1}{1-\rho^2}\mathrm{SR}_1 - \frac{\rho}{1-\rho^2}\mathrm{SR}_2 \right) \end{aligned}$$

and

$$\hat{x}_{t-1}^2 = \frac{1}{\gamma\sigma_2} \left(\frac{1}{1-\rho^2}\frac{\mu_2}{\sigma_2} - \frac{\rho}{1-\rho^2}\frac{\mu_1}{\sigma_1} \right).$$

How do these holdings depend on the correlation parameter ρ in the empirically relevant case where $\mu^1, \mu^2, \sigma_1, \sigma_2 > 0$ and $\rho \in (0,1)$? Let us first look at the risky asset 2 with the lower Sharpe ration $\mathrm{SR}_2 < \mathrm{SR}_1$. We have

$$\frac{d}{d\rho}\hat{x}^2_{t-1} = \frac{1}{(1-\rho^2)^2}\left(2\rho\mathrm{SR}_2 - (1+\rho^2)\mathrm{SR}_1\right) < -\frac{(1-\rho)^2}{(1-\rho^2)^2}\mathrm{SR}_1 < 0.$$

Whence, the investment in the less attractive asset decreases monotonically as the correlation rises and its value for diversifying the portfolio decreases. Once the correlation parameter crosses the threshold

$$\rho^* = \frac{\mathrm{SR}_2}{\mathrm{SR}_1} \in (0,1),$$

the less attractive risky asset is even sold short. Let us now turn to the risky asset with the higher Sharpe ratio, for which

$$\frac{d}{d\rho}\hat{x}^1_{t-1} = \frac{1}{(1-\rho^2)^2}\left(2\rho\mathrm{SR}_1 - (1+\rho^2)\mathrm{SR}_2\right).$$

We see that as the correlation increases from the uncorrelated case $\rho = 0$, the derivative is first negative so that the investment in the more attractive risky asset is also decreased because less risk can be diversified away. However, the derivative eventually becomes positive. In particular, for sufficiently strong correlations, increasingly large positions are built up in the attractive risky asset by shorting the less attractive one. For our example parameters from above and $\gamma = 1$ (which just is a scaling parameter), this is illustrated in Figure 4.

Risk-Reward Measures for Mean-Variance Optimization We have already seen that the Sharpe ratios of the underlying risky assets play an important role in mean-variance analysis. But what are the Sharpe ratios that can be achieved by mean-variance portfolios and how do these depend on the investor's risk aversion parameter?

In view of (4.2), the expected excess return of the mean-variance optimal portfolio is

$$\hat{x}^\top_{t-1}\mu_{t-1} = \frac{1}{\gamma}\mu^\top_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}.$$

The standard deviation of the excess return is

$$\sqrt{\hat{x}^\top_{t-1}\Sigma_{t-1}\hat{x}_{t-1}} = \frac{1}{\gamma}\sqrt{\mu^\top_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}}.$$

Whence, the Sharpe ratio for mean-variance optimal portfolios does *not* depend on the risk-aversion parameter at all:

$$\frac{\frac{1}{\gamma}\mu^\top_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}}{\frac{1}{\gamma}\sqrt{\mu^\top_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}}} = \sqrt{\mu^\top_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}}.$$

If the risky assets are uncorrelated, then the squared Sharpe ratio of the mean-variance optimal portfolio is the sum of the individual squared Sharpe ratios. The interpretation is

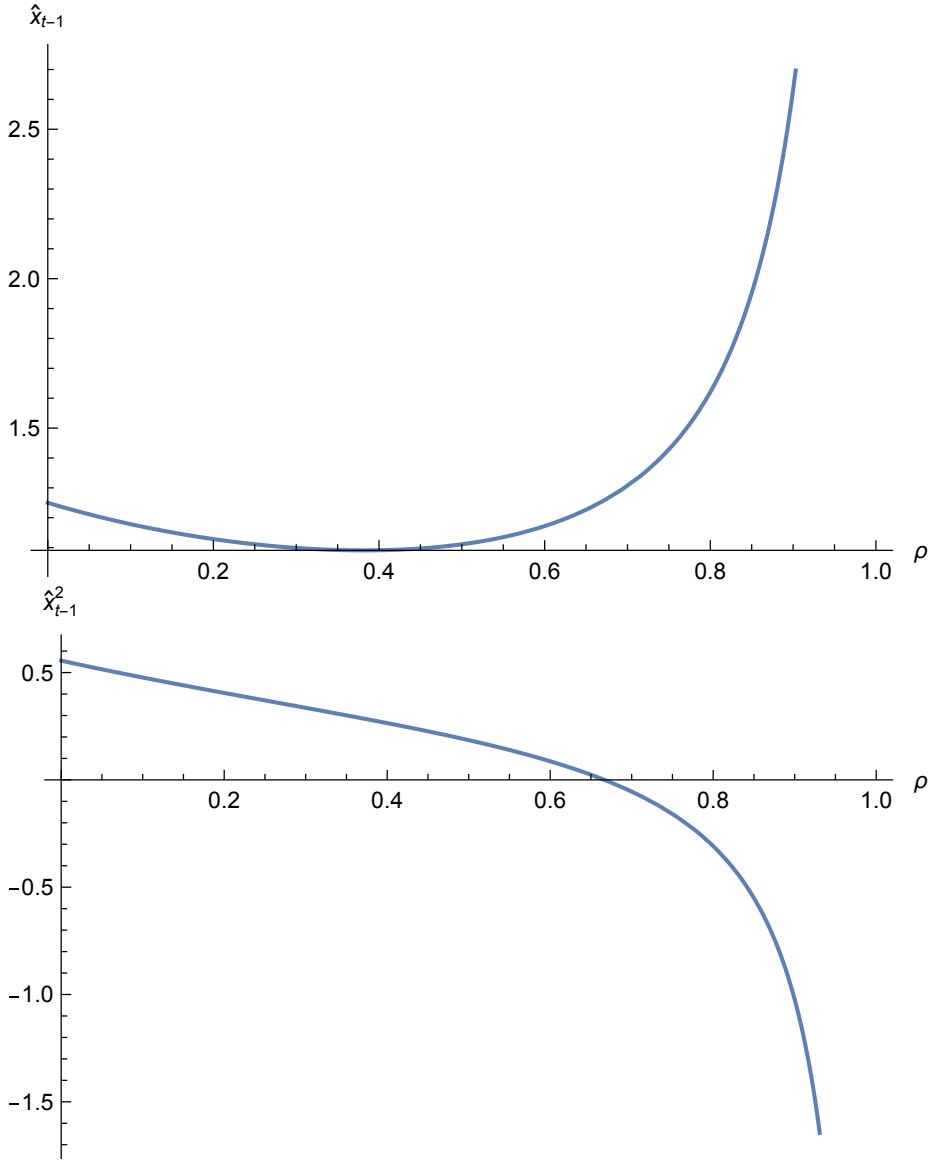


Figure 4: Risky investments $\hat{x}^1_{t-1}$ in the more attractive risky asset (upper panel) and $\hat{x}^2_{t-1}$ in the less attractive risky asset as a function of the correlation $\rho \in (0, 1)$.

that with maximal diversification, the individual investment opportunities can be combined very efficiently.

How does this change for the more realistic case of two positively correlated risky assets? In view of (4.3), the squared Sharpe ratio of the mean-variance optimal portfolio then is

$$\mathrm{SR}^2(\rho) = \mu_{t-1}^\top \Sigma_{t-1}^{-1} \mu_{t-1} = \frac{1}{1-\rho^2}\mathrm{SR}_1^2 - \frac{2\rho}{1-\rho^2}\mathrm{SR}_1\mathrm{SR}_2 + \frac{1}{1-\rho^2}\mathrm{SR}_2^2.$$

Differentiation shows that

$$\frac{d}{d\rho}\mathrm{SR}^2(\rho) = \frac{1}{(1-\rho^2)^2}\left(2\rho\mathrm{SR}_1^2 - (1+\rho^2)\mathrm{SR}_1\mathrm{SR}_2 + 2\rho\mathrm{SR}_2^2\right).$$

We see that correlation initially decreases the Sharpe ratio as the benefits or diversification are eroded, but the this effect is reversed for sufficiently strong correlation for which shorting the less attractive asset to purchase the more attractive one eventually almost becomes an arbitrage. For our example parameters from above, this is illustrated in Figure 5.

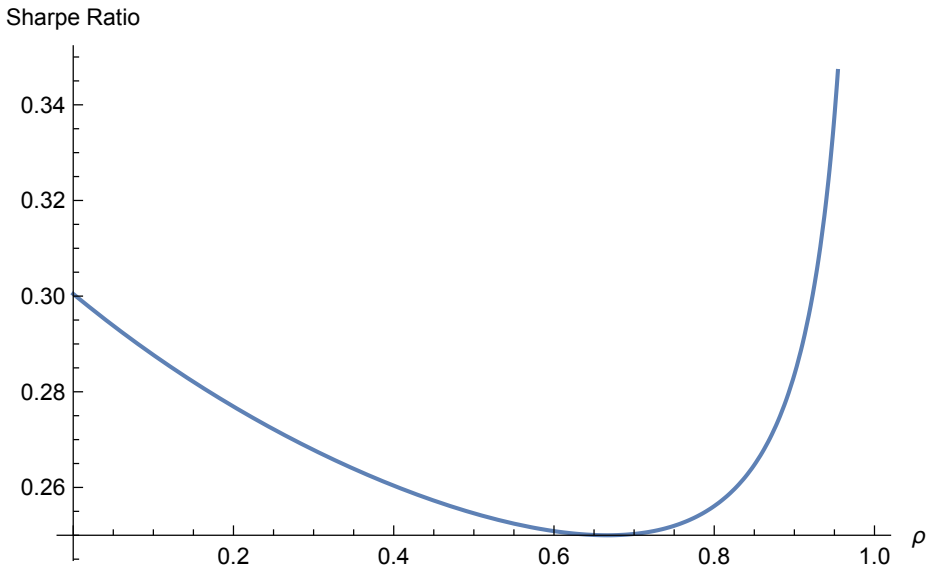


Figure 5: Sharpe ratio $\mathrm{SR}(\rho)$ of the mean-variance optimal portfolios plotted against the correlation $\rho \in (0,1)$ between the risky returns.

Efficient Frontier Another close link between Sharpe ratios and mean-variance optimization is revealed by relating the goal functional (4.1) to a constrained optimization problem. To wit, consider the problem of maximizing the expected excess return $x_{t-1}^\top \mu_{t-1}$ of the portfolio for a *fixed* portfolio variance $x_{t-1}^\top \Sigma_{t-1} x_{t-1} = \sigma^2$. (Such portfolios are called "efficient".) Equivalently, this means that the Sharpe ratio is maximized over portfolios with a given risk budget.

Then, for the optimal portfolio $\hat{x}_{t-1}$ there exists a Lagrange multiplier $\lambda > 0$ such that $(\hat{x}_{t-1}, \lambda)$ is a stationary point of the Lagrangian

$$\mathcal{L}(x, \lambda) = x^\top \mu_{t-1} - \lambda \left(x^\top \Sigma_{t-1} x - \sigma^2 \right).$$

In particular, $\hat{x}_{t-1}$ maximizes the strictly concave function $x \mapsto x^\top \mu_{t-1} - \lambda x^\top \Sigma_{t-1} x$, i.e., is mean-variance optimal for some risk aversion parameter. Conversely, the mean-variance optimal portfolio for any risk-aversion parameter γ is efficient, because it evidently maximizes the expected return over all competing portfolios with the same variance – otherwise, it could not be mean-variance optimal.

For our two risky assets with expected excess returns $\mu_{t-1} = (0.05, 0.05)^\top$ and covariance matrix

$$\Sigma_{t-1} = \begin{pmatrix} 0.2^2 & \rho \times 0.2 \times 0.3 \\ \rho \times 0.2 \times 0.3 & 0.3^2 \end{pmatrix} \quad \text{with } \rho = 0.2,$$

this is illustrated in Figure 6, which shows how the mean-optimal portfolio for different risk aversions trace out the "efficient frontier" among all possible portfolio allocations.

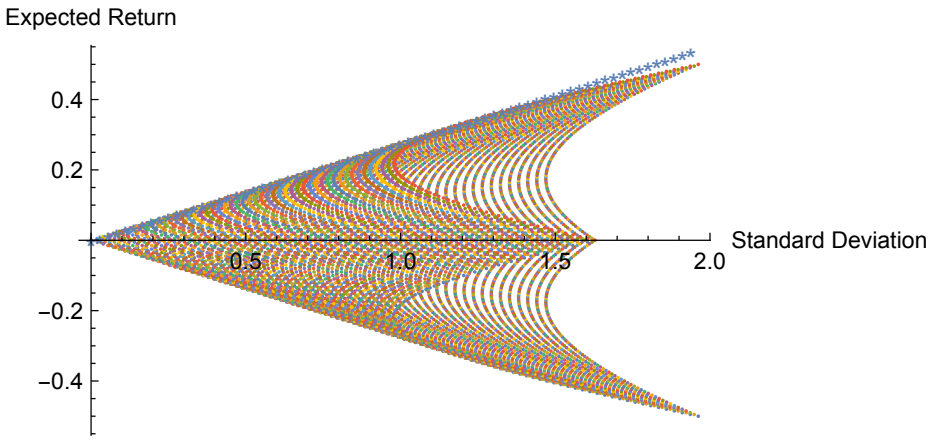


Figure 6: Expected excess returns plotted against the corresponding standard deviations for arbitrary portfolios (brights dots) and mean-variance optimal portfolios for different risk aversions (blue stars).

4.2 Extensions

The mean-variance approach discussed so far is the simplest and most tractable formalism for portfolio construction. However, it suffers from various limitation that can be addressed by more sophisticated models.

Robust Optimization and Shrinkage As already alluded to above, one problem is that – even if the model used was a perfect description of reality – covariances and, in particular, expected returns can only be estimated with substantial measurement errors. One way to

address this *uncertainty* about asset returns is to take a "robust approach". This means that one does not restrict to a single model (that is, the probability measure $\mathbb{P}$ under which means and variances are computed), but considers a whole class of alternatives (i.e., a set $\mathcal{P}$ of probability measures). A very cautious approach to take this into account then is to maximize the performance in the worst-case model:

$$\sup_{x_{t-1}} \inf_{\mathbb{P}\in\mathcal{P}} \left\{ \mathrm{E}_{t-1}^{\mathbb{P}} \left[W_t^{x_{t-1}} \right] - \frac{\gamma}{2} \mathrm{Var}_{t-1}^{\mathbb{P}} \left[X_t^{x_{t-1}} \right] \right\}.$$

However, this criterion typically is too conservative to be practically useful – without some exposure to risk and uncertainty one typically has to avoid risky investments completely. A middle ground is to penalize models that differ too much from a reference model $\bar{\mathbb{P}}$, say the point estimates obtained from a statistical analysis of historical time series:

$$\sup_{x_{t-1}} \inf_{\mathbb{P}\in\mathcal{P}} \left\{ \mathrm{E}_{t-1}^{\mathbb{P}} \left[W_t^{x_{t-1}} \right] - \frac{\gamma}{2} \mathrm{Var}_{t-1}^{\mathbb{P}} \left[X_t^{x_{t-1}} \right] + \frac{1}{\eta} d(\bar{\mathbb{P}}, \mathbb{P}) \right\}.$$

Here, $\eta > 0$ is a scaling parameter that measures how important model uncertainty is compared to risk.[14] The functional $d(\cdot,\cdot)$ is a distance on the space of probability measures; the most tractable choice is the *relative entropy*

$$d_{\mathrm{ent}}(\bar{\mathbb{P}}, \mathbb{P}) = \mathrm{E}_{t-1}^{\bar{\mathbb{P}}} \left[\log \left(\frac{d\bar{\mathbb{P}}}{d\mathbb{P}} \right) \right] = \int_{-\infty}^{\infty} \log \left(\frac{f_{\bar{\mathbb{P}}}(x)}{f_{\mathbb{P}}(x)} \right) f_{\bar{\mathbb{P}}}(x) dx.$$

(Here, the second equality holds if both probability measures have densities with respect to the Lebesgue measure.) How does this work in practice? Suppose the class of models we consider for the excess returns of a risky asset are the normal distributions $\mathcal{N}(\mu, \sigma^2)$ with known variance σ^2 but unknown mean μ, and the reference model is of the form $\mathcal{N}(\bar{\mu}, \sigma^2)$, where $\bar{\mu}$ is typically the mean of a time series of historical returns. With the density

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

of the normal distribution, we then obtain

$$\begin{aligned} d_{\mathrm{ent}}(\bar{\mathbb{P}}, \mathbb{P}) &= \int_{-\infty}^{\infty} \log \left(\frac{\phi_{\bar{\mu},\sigma^2}(x)}{\phi_{\mu,\sigma^2}(x)} \right) \phi_{\bar{\mu},\sigma^2}(x) dx \\ &= \int_{-\infty}^{\infty} \left(\frac{\mu^2 - \bar{\mu}^2 + 2x(\bar{\mu} - \mu)}{2\sigma^2} \right) \phi_{\bar{\mu},\sigma^2}(x) dx \\ &= \frac{\mu^2 - \bar{\mu}^2 + 2\bar{\mu}(\bar{\mu} - \mu)}{2\sigma^2} = \frac{(\mu - \bar{\mu})^2}{2\sigma^2}. \end{aligned}$$

Whence, for normally distributed returns, the robust mean-variance optimization problem is still linear quadratic:

$$(1 + R^f) W_{t-1} + \sup_{x_{t-1}} \inf_{\mu\in\mathbb{R}} \left\{ \mu x_{t-1} - \frac{\gamma}{2} x_{t-1}^2 \sigma^2 + \frac{1}{\eta} \frac{(\mu - \bar{\mu})^2}{2\sigma^2} \right\}. \tag{4.4}$$

[14] Indeed, for $\eta \uparrow \infty$, we recover the worst case model. Conversely, for $\eta \downarrow 0$ we are back in the standard mean-variance framework for the reference model $\bar{\mathbb{P}}$.

The first-order condition for the "worst-case model" is

$$0 = x_{t-1} + \frac{1}{\eta}\frac{\mu - \bar{\mu}}{\sigma^2} \quad \rightsquigarrow \quad \hat{\mu} = \bar{\mu} - \eta\sigma^2 x_{t-1}.$$

After plugging this back into (4.4), the robust optimization problem becomes

$$(1 + R^f)W_{t-1} + \sup_{x_{t-1}} \left\{ \bar{\mu} x_{t-1} - \frac{\gamma + \eta}{2} x_{t-1}^2 \sigma^2 \right\}.$$

As a consequence, we are back in the standard mean-variance framework, but with the risk-aversion parameter γ increased by the "uncertainty aversion" η. The corresponding optimal portfolio is

$$\hat{x}_{t-1} = \frac{\bar{\mu}}{(\gamma + \eta)\sigma^2}.$$

Thus, compared to the standard model that disregards model uncertainty, the optimal risky investment (or equivalently, the point estimate for the expected return) is "shrunk" towards a full safe investment (corresponding a zero excess return estimate). The effects of risk and uncertainty can be different for non-normal distributions, but the intuition gleaned from this simple model is still useful: when faced with substantial estimation errors for the expected returns, the cautious way to proceed is to shrink them towards zero.

The variance of a single risky asset can be estimated reasonably well; this is why we have treated it as known in the discussion above. However, in the practically relevant case where many risky assets are considered, estimating the covariance matrix of the corresponding returns also becomes problematic. One problem is the number of parameters involved: for $N = 100$ risky assets, we already have to estimate $N(N-1)/2 = 4950$ covariances, but five years of monthly data only correspond to 6000 data points. A second problem is that the optimal portfolio

$$\hat{x}_{t-1} = (\gamma \Sigma_{t-1})^{-1} \mu_{t-1}$$

is very sensitive to the smallest eigenvalues of the covariance matrix Σ_{t-1}, since these will lead to large investments in the optimal portfolio. To address this, empirical covariance matrices are often shrunk towards a diagonal matrix in practice. Moreover, high-dimensional mean-variance optimization is almost never used without imposing some position limits that prevent excessive leverage, for example. While this rules out explicit solutions, one can instead turn to efficient numerical methods for convex optimization problems.

Expected Utility Maximization Another problematic feature of the mean-variance goal functional

$$\sup_{x_{t-1}} \left\{ \mathrm{E}_{t-1}[W_t^{x_{t-1}}] - \frac{\gamma}{2} \mathrm{Var}_{t-1}[W_t^{x_{t-1}}] \right\} = (1 + R^f)W_{t-1} + \sup_{x_{t-1}} \left\{ x_{t-1}^\top \mu_{t-1} - \frac{\gamma}{2} x_{t-1}^\top \Sigma_{t-1} x_{t-1} \right\}$$

is that it is not monotone with respect to wealth $W_t^{x_{t-1}}$. Whence, a mean-variance investor may turn down a free lottery ticket if the variance of its payoff is sufficiently large compared to its expectation. To avoid this feature and the paradoxical behaviours it can generate,

one can instead consider an *increasing* concave "utility function" $U(\cdot)$ of wealth, and in turn maximize the expected utility:

$$\sup_{x_{t-1}} \mathrm{E}_{t-1}[U(W_t^{x_{t-1}})].$$

The disadvantage of this approach is (i) that it is more difficult to interpret and (ii) that the computations are typically more involved because the corresponding first-order conditions are generally nonlinear. However, for the *exponential utility function* $U(x) = -\exp(-\gamma x)$ and (conditionally) normally distributed excess returns $R_t^e \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$ we have

$$\begin{aligned}\mathrm{E}_{t-1}[U(W_t^{x_{t-1}}] &= \mathrm{E}_{t-1}\left[-\exp(-\gamma((1+R^f)W_{t-1} + x_{t-1}R_t^e))\right] \\ &= -\exp(-\gamma(1+R^f)W_{t-1})\exp\left(-\gamma x_{t-1}^\top \mu_{t-1} + \frac{\gamma^2}{2} x_{t-1}^\top \Sigma_{t-1} x_{t-1}\right).\end{aligned}$$

(Here, we have used the moment-generating function of the multivariate normal distribution in the second step.) Whence, maximizing expected exponential utility is equivalent to mean-variance optimization for normally distributed returns. More generally, mean-variance and expected utility maximization typically yield similar results for returns that are not too far away from a Gaussian distribution, in that they not too skewed and have tails that are not too heavy.

Asset Pricing with Expected Utility As an application of expected utility maximization, let us come briefly reconsider equilibrium asset pricing by a representative agent in this context. For a given excess return R_t^e, the chain rule shows that the first-order condition for the optimal holdings is

$$0 = \mathrm{E}_{t-1}[U'(W_t^{x_{t-1}})R_t^e] \tag{4.5}$$

because $W_t^{x_{t-1}} = (1+r^f)W_{t-1} + x_{t-1}^\top R_t^e$. Solving for the optimal $\hat{x}_{t-1}$ is generally not possible in closed form, because this equation is typically not linear.

However, in equilibrium, the *representative* agent holds the entire supply s of the risky assets and no risk-free assets (since these are in zero net supply). Whence, in a single-period model with liquidating dividends $P_t = D_t$ at time t, her wealth at time t is known: $s^\top D_t$. The corresponding excess return is

$$R_t^{n,e} = \frac{D_t^n - P_{t-1}^n}{P_{t-1}^n} - R^f = \frac{D_t^n}{P_{t-1}^n} - (1+R^f),$$

so if an equilibrium exists it must satisfy

$$(1+R^f)\mathrm{E}_{t-1}[U'(s^\top D_t)] = \frac{\mathrm{E}_{t-1}[U'(s^\top D_t)D_t]}{P_{t-1}}.$$

The equilibrium price at time $t-1$ therefore must be

$$P_{t-1} = \mathrm{E}_{t-1}[\xi_{t-1,t} D_t], \tag{4.6}$$

where we have defined the *stochastic discount factor*

$$\xi_{t-1,t} = \frac{1}{1+R^f}\frac{U'(s^\top D_t)}{\mathrm{E}_{t-1}[U'(s^\top D_t)]}.$$

For increasing utility functions, $U' > 0$ so that the stochastic discount factor is positive. Whence, assets with positive payoffs always have positive prices, unlike for mean-variance preferences.

If there is no uncertainty about the aggregate dividends $s^\top D_t$, then the stochastic discount factor reduces to discounting with the risk-free rate. In general, assets with the same expected payoff are value more highly if they have positive covariance with the stochastic discount factor. For concave utilities, U' is decreasing. Whence, assets are valued more highly if the have high payoffs when the aggregate dividend payments in the market are low.

Another way to interpret (4.6) in analogy to option pricing theory is to pass to *discounted* prices expressed in terms of the safe asset. Then, (4.6) states that the discounted equilibrium price at time $t-1$ is the expectation of the discounted dividend at time t, computed under the pricing measure $\mathbb{Q}$ with density $U'(s^\top D_t)/E_{t-1}[U'(s^\top D_t)]$ relative to the physical probability $\mathbb{P}$. Thus, in analogy to the fundamental theorem of asset pricing, all discounted asset prices are the conditional expectations of their terminal values here, computed under a "risk-neutral" measure. Here, in a simple model, we have derived this pricing measure from "first principles", that is, from aggregate uncertainty in the economy and risk preferences of the representative agent. But in much applied work a stochastic discount factor is simply specified exogenously, just like the risk-neutral models used in derivative pricing.

Multi-Period Optimization So far, we have focused one one-period models where the investment decision from time $t-1$ to t only takes into account information available at time $t-1$. Put differently, the initial decision is not updated dynamically as new information comes in before the end of the investment horizon is reached.

A more realistic model is to start from an initial wealth W_0 at time 0, and then make investment decisions x_{t-1} at times $0, 1, \ldots, T-1$ based on the information that has become available until then. The final wealth will then be

$$W_T^x = (1+R^f)W_{t-1}^x + x_{T-1}^\top R_T^e = \ldots = (1+R^f)^T W_0 + \sum_{t=1}^{T} x_{t-1}^\top R_t^e.$$

For an increasing utility function $U(\cdot)$, we can then maximize the expected utility at the initial time:

$$\sup_{x=(x_0,\ldots,x_{T-1})} \mathrm{E}_0\left[U(W_T^x)\right].$$

Here, the supremum is taken over all investment strategies that are adapted to the information flow of the model. Such dynamic *stochastic control problems* are generally difficult to solve, because they require a tradeoff between the immediate effects of the current decision and its long-term consequences.

A simple example where this is not necessary is if the utility function is exponential

($U(x) = -\exp(-\gamma x)$ and the excess returns R_t^e are independent over time. Then:

$$\begin{aligned}\mathrm{E}_0[U(W_T^x)] &= \exp(-\gamma(1+R^f)^T W_0)\mathrm{E}_0\left[-\exp\left(-\gamma\sum_{t=1}^{T} x_{t-1}^\top R_t^e\right)\right] \\ &= \exp(-\gamma(1+R^f)^T W_0)\prod_{t=1}^{T}\mathrm{E}_0\left[-\exp\left(-\gamma x_{t-1}^\top R_t^e\right)\right].\end{aligned}$$

Whence, by the tower property of conditional expectations, the investment decision x_{t-1} is optimized by maximizing the same one-period functional $\mathrm{E}_{t-1}[U(x_{t-1}^\top R_t^e)]$ we have considered above. This separation of investment choices breaks down when asset returns are not independent, e.g., due to stochastic volatility or predictive signals about expected returns. General tools for approaching such optimization problems are developed in Eyal's courses on *Stochastic Control* and *High-Frequency Trading*.

Multi-Period Asset Pricing Maybe surprisingly, representative-agent equilibrium models are still tractable in a multiperiod context, however, without the use of such sophisticated tools. For simplicity, we focus on the simplest case where only one liquidating dividend is paid at the terminal time T. (More flexible equilibrium models with intermediate dividends require intermediate consumption, to model "where the dividends go".) Then, the first-order condition for optimality of any of the investment decisions x_{t-1} is

$$\begin{aligned}0 = \mathrm{E}_0[U'(W_T^x)R_t^e] &= \mathrm{E}_0\left[\mathrm{E}_t\left[U'(W_T^x)\left(\frac{P_t - P_{t-1}}{P_{t-1}} - R^f\right)\right]\right] \\ &= \mathrm{E}_0\left[\frac{\mathrm{E}_{t-1}\left[U'(W_T^x)P_t\right]}{P_{t-1}} - (1+R^f)\mathrm{E}_{t-1}\left[U'(W_T^x)\right]\right].\end{aligned}$$

(Here, we have once again used the tower property of conditional expectations in the second step and taken out what is known at time $t-1$ in the third.) This first-order condition is difficult to simplify further, because the terminal wealth W_T^x generally depends in the unknown control $x = (x_0, \ldots, x_{T-1})$ in a complex manner.

However, in a representative-agent equilibrium, the representative agent holds no safe assets and all risky shares at all times. As a consequence, at the terminal time the wealth of the representative agent is simply the exogenously given aggregate dividend $s^\top D_T$. Whence, to satisfy the first-order conditions derived above at the last trading time $T-1$, we need

$$P_{T-1} = \mathrm{E}_{T-1}[\xi_{T-1,T} D_T].$$

Here, just as in the one-period model considered above, the stochastic discount factor from time $T-1$ to T is

$$\xi_{T-1,T} = \frac{1}{1+R^f}\frac{U'(s^\top D_T)}{\mathrm{E}_{T-1}[U'(s^\top D_T)]}. \tag{4.7}$$

At the earlier times, we need

$$P_{t-1} = \mathrm{E}_{t-1}[\xi_{t-1,t} P_t],$$

where, by another application of the tower property, the corresponding stochastic discount factor from time $t-1$ to t is

$$\xi_{t-1,t} = \frac{1}{1+R^f} \frac{\mathrm{E}_t[U'(s^\top D_T)]}{\mathrm{E}_{t-1}[U'(s^\top D_T)]}.$$

As a consequence,

$$\xi_t P_t = \mathrm{E}_t[\xi_T D_T] \quad \text{where } \xi_t = \prod_{s=1}^{t} \xi_{s-1,s}$$

is the discount factor from time 0 to t. Equivalently, asset prices discounted with the risk-free rate are the conditional expectations of the terminal dividends under the equilibrium pricing measure with density process given in terms of the expectations of the terminal dividend. In particular, all discounted asset prices are martingales under this risk-neutral measure.

In more general models with intermediate dividends and consumption, similar relationships hold. For example, in an infinite-horizon setting (with appropriate discounting rather than a terminal time), equilibrium prices then are the sum of future expected dividends, discounted with an appropriate stochastic discount factor:

$$\xi_t P_t = \sum_{s=t+1}^{\infty} \mathrm{E}_t[\xi_s D_s]. \tag{4.8}$$

This relation is called the "dividend discount model" (or the "discounted cash flow model" or the "present value model"). The simplest version of this model obtains for a constant discount rate, $\xi_t = (1+k)^{-t}$. (For example, in (4.7) this happens if there is no "aggregate uncertainty" in that the total dividends $s^\top D_T$ are not random.) For *stochastic* discount factors, the relationship (4.8) once again shows that among dividends with the same expected values, the ones are valued more highly that have positive covariance with the discount factor. This means that dividends are valued more if they are high when the market values them highly (because the aggregate dividend in (4.7) is low).

The dividend discount model plays a central role in equity valuation. We will discuss this in more detail in Section 6.1.

4.3 Risk Management

Measuring Risk Risk can and should be measured in different ways. One very common measure of risk is *volatility*, that is, the standard deviation of an uncertain payoff or return. This is natural for normal distributions, where the standard deviation in fact describes the entire distribution of events around the expected value. More generally, volatility is a useful risk measure for distributions that are relatively symmetric (i.e., not the lottery tickets alluded to above, for example) and don't have extreme exposure to crash risk. For normal distributions, two standard deviations from the mean are uncommon and five standard-deviation events almost never happen. But for the returns of hedge-fund strategies, two standard deviation events are common and five standard deviation events certainly do happen. Hence, it is not prudent to rely on volatility alone as a risk measure in this context.

Another very popular risk measure is the *value-at-risk (VaR)*. This measures the maximum loss that can occur with a certain confidence. For example, the 99%-VaR is the most you can loose with 99% confidence:

$$\mathbb{P}\left[\text{Loss } \leq \text{Var}_{99\%}\right] \geq 0.99.$$

To estimate the 99%-VaR from historical data, you simply sort your past returns, and find a return for which 1% of returns are worse and 99% are better. (Of course, one has to be careful if the positions of the trading strategy and the market environment have changed significantly in the meantime.)

An issue with the VaR is that it says nothing about the magnitude of the losses if this threshold is exceeded. This is addressed by the so-called *expected shortfall*, which is the expected loss given that the VaR is exceeded:

$$\text{ES} = \text{E}[\text{Loss}|\text{Loss} > \text{Var}].$$

Of course, this tail risk measure is quite difficult to estimate reliably, since one is taking the average of only very few historical data points (1% of the entire data if the 99%-VaR is used).

Another important class of risk measures is based on various *stress tests*. This means that portfolio returns are simulated in a number of extreme market scenarios. This can include past events such as the market crashes after the collapse of LTCM or Lehman Brothers, as well as hypothetical future events such as the failure of a sovereign state, a spike in volatility, or a sharp increase in margin requirements. Unlike volatility and VaR (which measure the risk during relatively "normal" markets), stress tests explore extreme events for which not enough historical data is available to allow an accurate statistical estimation. For hedge funds, the main goal is to make sure that none of their positions are so large that they put the entire fund at risk of blowing up in a stress scenario.

Prospective and Reactive Risk Management Once a methodology for measuring risks is established, the next step is to manage these exposures. Here, *prospective* risk management refers to controlling risk before a bad event occurs. This can be done by diversifying investments, imposing risk limits, and by tail hedging via options, for example.

In contrast, *reactive* risk control reacts to bad events and seeks to limit the losses as they evolve. One typical mechanism for this is *drawdown control*, which is particularly important for leveraged hedge funds who cannot simply "ride out" their way through a crisis. Therefore, one may want to minimize the risk that the drawdown becomes worse then some prespecified *maximum acceptable drawdown (MADD)*, say 25%, from the historical maximum of the fund value. One sensible way to implement this is to impose

$$\text{VaR}_t \leq \text{MADD} - \text{DD}_t.$$

The right-hand side of this inequality is the distance between the maximum acceptable drawdown and the current drawdown, that is, the maximum acceptable loss given what has already been lost. The value-at-risk on the left hand side is the largest amount that can be lost with a certain confidence, given current positions and current market risk. Hence, this constraint requires that the current risk must be kept small enough so that losses do

not push the drawdown beyond MADD with a certain confidence level. If this inequality is violated, the hedge fund should reduce risk by unwinding positions, until the VaR is reduced sufficiently to satisfy the inequality again.

A prespecified drawdown control (or other reactive risk management strategy) has the advantage of creating a clear plan for how to handle adversity. Indeed, reducing risk after losing on a position is painful, since the trader feels that a loss is being locked in if the trade is unwound. This creates an incentive for trying to ride out the situation in order to recover the losses – which may in turn lead to growing losses and finally disaster.

5 Implementing Trading Strategies

Implementing trading strategies with substantial turnover can be difficult due to *transaction costs*. These manifest themselves in various forms. First, most (in particular, small) investors pay commissions and other direct costs on each trade. Second, there is a "bid-ask spread" between the lowest price currently asked if you want to purchase a share, compared to the highest price that is bid if you want to sell it. Third (and most importantly for large investors), large trades executed quickly adversely affect the quoted market prices. Indeed, if you have to liquidate a large position quickly, then this pushes down the market price, leading to lower proceeds than a "paper trade" carried out at the initial market quote.

5.1 Optimal Trading with Transaction Costs

In a world without trading costs, it is optimal to react to any new piece of information, even if this requires to trade in and out of large positions very frequently. When transaction costs are taken into account, it becomes apparent that this is not efficient. Instead, one needs to strike a balance between the gains and costs of trading. How to do this optimally depends on the nature of the most relevant trading costs.

Decreasing Costs (as a function of trade size) In an Over-the-Counter (OTC) market, you often have to call a dealer on the phone to trade. It takes the dealer a similar amount of time to process a large trade as a small one. Therefore, transaction costs in such markets tend to be larger for small orders as a percentage of the price or per share traded. How is this reflected in optimal trading strategies? Since small trades are very unfavorable, it is optimal not to trade at all until the actual position deviates too far from a continuously adjusted "target position". When this happens, you call the dealer and trade back to your preferred allocation.

Constant Costs: Bid-Ask Spreads These costs are also called *proportional* costs since they are proportional to trade size (but constant per unit traded). They are most relevant for trading strategies that trade positions that are small enough to be filled at the best bid-ask prices in electronic limit-order books. This is particularly relevant for "large tick stocks", where the exchange imposes a large minimal difference between bid and ask prices. This in turn makes liquidity provision profitable enough for a lot of liquidity to be provided at the bid-ask prices. Since constant costs penalize small trades less severely than decreasing costs, the corresponding optimal strategies do not prescribe to trade all the way back to a target

strategy once a certain threshold is breached. Instead, one performs just enough trading to keep the deviation from the target small enough.

Increasing Costs (as a function of trade size): Market Impact If the trades associated to a strategy are large enough to exhaust the liquidity provided at the best bid and ask prices, then it causes market impact. The adverse effect on the average execution price is more severe the larger the position that is traded. Whence, if these costs are the most relevant ones, then it it optimal to split up larger trades into many small ones and gradually trade towards the target allocation, e.g., by reducing the deviation by a certain percentage each day.

The different forms of these three types optimal trading strategies are illustrated in Figure 7, taken from Pedersen (2015). A proper analysis of the impact of transaction costs on optimal trading strategies leads to difficult stochastic control problems and is beyond our scope here; see, e.g., Muhle-Karbe et al. (2017) for an introduction to the related literature. To illustrate some of the concepts with elementary tools, we now discuss a particularly simple optimization problem with market impact, namely the execution of a single exogenously given trade.

5.2 Optimal Execution

Model We consider an agent with initial position of $\Phi > 0$ of a risky asset that needs to be liquidated in T trading rounds. Write φ_t for the position after t trading rounds and denote by $\Delta\varphi_t = \varphi_t - \varphi_{t-1}$ the number of shares traded in round t. In view of the hard liquidation constraint, the agent chooses from the (adapted) strategies $(\varphi_t)_{t=0,\dots,T}$ that satisfy $\varphi_0 = \Phi$ and $\varphi_T = 0$.

The unaffected price process $(\bar{P}_t)_{t=1,\dots,T}$ of the risky asset is a martingale.[15] As suggested by Almgren and Chriss (2001), trades have both a temporary and a permanent impact, in that the execution price per share in trading round t is

$$P_t = \bar{P}_t + \beta(\varphi_t - \Phi) + \lambda\Delta\varphi_t.$$

Here, $\beta > 0$ describes the *permanent impact* of all past trades $\sum_{s=1}^{t} \Delta\varphi_s = \varphi_t - \Phi$, while $\lambda > 0$ measures the *temporary impact* of the current trade $\Delta\varphi_t$. Accordingly, the liquidation proceeds of a strategy $(\Delta\varphi_t)_{t=1,\dots,T}$ are

$$-\sum_{t=1}^{T} P_t\Delta\varphi_t = -\sum_{t=1}^{T} \lambda\Delta\varphi_t^2 - \sum_{t=1}^{T} \beta(\varphi_t - \Phi)(\varphi_t - \varphi_{t-1}) - \sum_{t=1}^{T} \bar{P}_t(\varphi_t - \varphi_{t-1}). \tag{5.1}$$

[15] This means that it has no predictable trend, which is is a reasonable assumption over the short time horizons typical for the implementation of a single order. This assumption will simplify the dynamic optimization to a deterministic problem below.

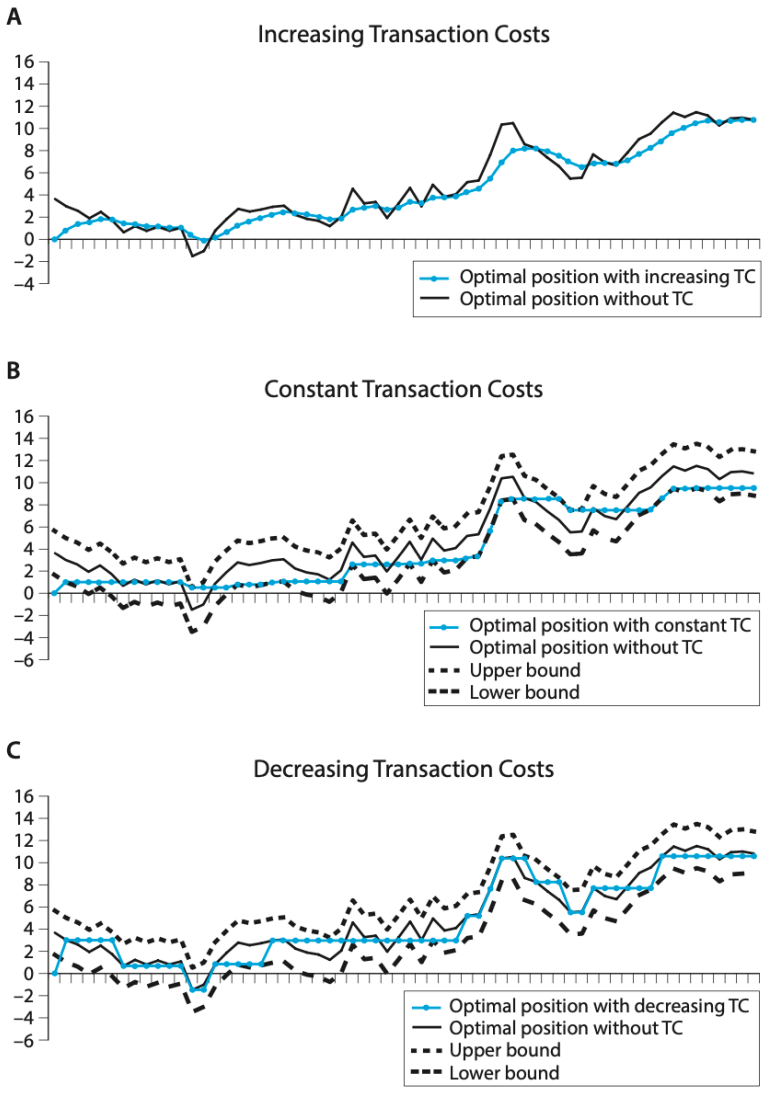


Figure 7: Optimal trading strategies for different forms of transaction costs.

The second term in (5.1) can be rewritten as

$$\begin{aligned}\sum_{t=1}^{T}\beta(\varphi_t-\Phi)(\varphi_{t-1}-\varphi_t) &= \frac{\beta}{2}\sum_{t=1}^{T}\left(\varphi_{t-1}^2-\varphi_t^2-(\varphi_t-\varphi_{t-1})^2-2\Phi(\varphi_{t-1}-\varphi_t)\right)\\ &= \frac{\beta}{2}\left(\varphi_0^2-\varphi_T^2-\sum_{t=1}^{T}\Delta\varphi_t^2-2\Phi\varphi_0+2\Phi\varphi_T\right)\\ &= \frac{\beta}{2}\left(-\Phi^2-\sum_{t=1}^{T}\Delta\varphi_t^2\right).\end{aligned}$$

Similarly, the third term in (5.1) can be rewritten as

$$\begin{aligned}-\sum_{t=1}^{T}\bar{P}_t(\varphi_t-\varphi_{t-1}) &= -\varphi_T\bar{P}_T+\varphi_0\bar{P}_0+\sum_{t=1}^{T}\varphi_{t-1}(\bar{P}_t-\bar{P}_{t-1})\\ &= \Phi\bar{P}_0+\sum_{t=1}^{T}\varphi_{t-1}\Delta\bar{P}_t.\end{aligned}$$

Together with the martingale property of the unaffected price $\bar{P}$ (which implies that the conditional expectation of changes in the unaffected price are zero), it therefore follows that the *expected* liquidation proceeds are

$$\mathrm{E}_0\left[-\sum_{t=1}^{T}P_t\Delta\varphi_t\right]=\Phi\bar{P}_0-\frac{\beta}{2}\Phi^2-\tilde{\lambda}\sum_{t=1}^{T}\mathrm{E}_0\left[\Delta\varphi_t^2\right],\quad\text{where }\tilde{\lambda}=\lambda+\frac{\beta}{2}.$$

Maximizing the liquidation proceeds (or, equivalently, minimizing the shortfall relative to the frictionless benchmark ΦP_0) already is a nontrivial and well-posed problem. Nevertheless, it makes sense to add an additional cost for holding positions, in order to penalize liquidation programs that trade too slowly and thereby expose the investor to the risk of adverse price changes. The simplest way to model this is to impose a cost γ on the average squared holdings:

$$\sum_{t=1}^{T}\gamma\mathrm{E}_0[\varphi_t^2].$$

Neglecting the terms $\Phi P_0-\frac{\beta}{2}\Phi^2$ that do not depend on the choice of the execution strategy, this leads to the following quadratic goal functional:

$$\sum_{t=1}^{T}\mathrm{E}_0\left[\tilde{\lambda}\Delta\varphi_t^2+\gamma\varphi_t^2\right]\to\min!\tag{5.2}$$

Now, observe that – by the assumption that the price process is a martingale – the random price changes have disappeared from the goal functional. Accordingly, it is not surprising that the optimal execution strategy is deterministic. To see this, consider any (potentially random) strategy and replace it by its expectation. Then, the initial and terminal constraints are still satisfied but it follows from Jensen's inequality that the risk and trading costs are

less than or equal to their counterparts for the original random strategy. Whence, we can restrict ourselves to deterministic execution strategies, and (5.2) in turn reduces to the optimization problem

$$\sum_{t=1}^{T} \Big[\tilde{\lambda} \Delta \varphi_t^2 + \gamma \varphi_t^2 \Big] \to \min! \tag{5.3}$$

in $\mathbb{R}^T$ (or, more precisely $\mathbb{R}^{T-1}$, since the terminal position is fixed by the hard liquidation constraint). To determine the optimizer, it now suffices to find the (unique) root of the first-order conditions, obtained by setting the derivatives with respect to the control variables φ_t equal to zero:

$$0 = 2\tilde{\lambda}(\varphi_t - \varphi_{t-1}) - 2\tilde{\lambda}(\varphi_{t+1} - \varphi_t) + 2\gamma\varphi_t, \quad t = 1, \dots, T-1.$$

Therefore, the optimal holdings $(\varphi_t)_{t=0,\dots,T}$ satisfies the following linear difference equation:

$$\varphi_{t+1} - \left(2 + \frac{\gamma}{\tilde{\lambda}}\right) \varphi_t + \varphi_{t-1} = 0, \quad t = 1, \dots T-1, \quad \text{with } \varphi_0 = \Phi \text{ and } \varphi_T = 0. \tag{5.4}$$

To solve this equation, first disregard the initial and terminal conditions. Plugging the trial solution $\varphi_t = e^{\kappa t}$ into the difference equation, we find that κ needs to solve

$$0 = e^{\kappa} - \left(2 + \frac{\gamma}{\tilde{\lambda}}\right) + e^{-\kappa} = 2\Big(\cosh(\kappa) - 1\Big) - \frac{\gamma}{\tilde{\lambda}}.$$

Here, we have used the definition of the hyperbolic cosine in the second step, $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$. As this function is symmetric and maps the real line to $[1, \infty)$, there are exactly two solutions $\pm\kappa$ (with $\kappa > 0$) of the above equation for any values $\gamma, \tilde{\lambda} > 0$. (There is no explicit expression for $\pm\kappa$, though.) Any linear combination $\alpha e^{\kappa n} + \beta e^{-\kappa n}$ is in turn also a solution of the difference equation (5.4). The correct weights are then identified by matching the initial and terminal conditions, which leads to

$$\alpha + \beta = \Phi, \quad \alpha e^{\kappa T} + \beta e^{-\kappa T} = 0.$$

As a consequence, $\alpha = \Phi - \beta$, so that β has to solve

$$0 = (\Phi - \beta)e^{\kappa T} + \beta e^{-\kappa T} \quad \rightsquigarrow \quad \beta = \frac{e^{\kappa T}\Phi}{e^{\kappa T} - e^{-\kappa T}} \quad \rightsquigarrow \quad \alpha = \Phi - \beta = -\frac{e^{-\kappa T}\Phi}{e^{\kappa T} - e^{-\kappa T}}.$$

In summary, the solution of our difference equation with correct boundary conditions is

$$\varphi_t = -\frac{e^{-\kappa(T-t)}\Phi}{e^{\kappa T} - e^{-\kappa T}} + \frac{e^{\kappa(T-t)}\Phi}{e^{\kappa T} - e^{-\kappa T}} = \frac{\sinh(\kappa(T-t))}{\sinh(\kappa T)}\Phi.$$

(Here, we have used the definition of the hyperbolic sine in the second step, $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$.) As a result, the optimal holdings decrease monotonically from Φ to 0, with an "urgency" parameter κ that is determined by the ratio of the holding cost γ and the "effective" trading cost $\tilde{\lambda}$. For very large γ, κ tends to ∞ and the position is almost completely liquidated already in the first trading round. In contrast, for $\gamma \to 0$, we have $\kappa \to 0$. l'Hospital's rule in turn shows that the optimal position converges to $\varphi_t = \frac{T-t}{T}\Phi$. This means that agents without inventory costs optimally liquidate at a constant rate ("TWAP") irrespective of their price impact costs. Intermediate values of the inventory cost interpolate between these two extreme regimes, as illustrated in Figure 8.

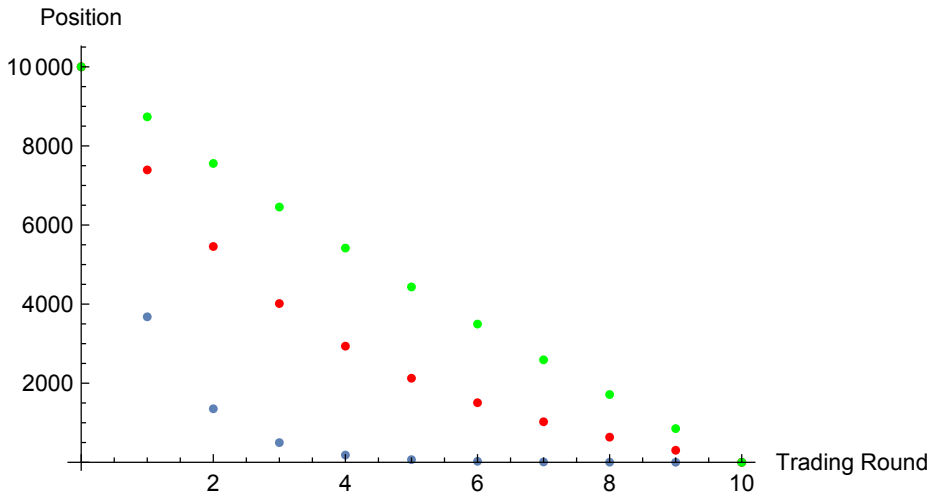


Figure 8: Almgren-Chriss execution paths for $T = 10$ trading rounds, initial position $\Phi = 10,000$, and urgency parameters $\kappa = 1$ (blue), $\kappa = 0.3$ (red), and $\kappa = 0.1$ (green).

5.3 Measuring Price Impact

We now briefly discuss how to measure the price impact of large trades from historical data. Recall that

$$\text{Price change} = \text{noise} + \text{alpha} + \text{price impact}$$

For short-time horizons, the alpha can typically be neglected (like through the martingale assumption in the Almgren/Chriss model).[16] The noise term usually is by far the largest for short horizons, but vanishes *on average*. Therefore, it can be averaged out over many trades of similar size. To estimate market impact as a function of trade size, one therefore average the price changes following trades with similar size.

How to do this in practice depends on the type of data that is available. Many academic papers are based on public datasets, where trades are anonymous and often not even identified as buyer-initiated or seller-initiated. This makes it difficult the piece to together the "metaorders" the investors have broken up into small pieces to reduce price impact.

In contrast, "semi-public data" collected by brokers, for example, groups trades into different metaorders, but also contains no information about the investors that executed these trades (e.g., on whether they used sophisticated algorithms or not) and their trading motives. Such data is used in Almgren et al. (2005), for example.

Finally, the finest analyses are possible with proprietary data collected by sophisticated market participants themselves. In this case, all information is available. Such datasets are used in research by quantitative asset managers like AQR and CFM, for example.

Fortunately (for researchers and less sophisticated market participants), this research

[16] Note, however, that this can be a problem for trading strategies based on fast intraday trading signals. In this case, it is a very challenging problem to distinguish whether prices went up because your signal correctly predicted this, or whether this happened because of price impact.

indicates that estimate based on public data already yield useful ballpark numbers. Some of the classic measures used in this context are "Amihud's measure" (Amihud, 2002). This considers a moving average of absolute daily percentage returns divided by daily trading volume (in dollars). The intuition is that a market is illiquid if prices move a lot with little volume. Note that this measure is not derived from any model, and makes no statement about a particular form of form of price impact as a function of the traded order size.

The most well-known theoretical model that addresses this is "Kyle's model" (Kyle, 1985) from the Market Microstructure literature, discussed in detail in Emma's *Market Microstucture* elective. This model suggests that price impact should be linear in trade size, and proportional to volatility of prices divided by trading volume. The intuition for this is that there is more scope for adverse selection by "insiders" in markets with lots of price changes relative to uninformed "noise trading".

How do these theoretical results compare to empirical estimates based on proprietary trade data? Reassuringly, estimate for linear price impact lead to similar orders of magnitude (Muhle-Karbe et al., 2020). However, empirical evidence strongly suggests that price impact is not really linear in trade size. Instead various studies consistently find that it scales approximately with the *square-root* of the order size Q:

$$\text{Impact}(Q) = Y \times \text{Volatility} \times \sqrt{\frac{Q}{\text{Volume}}} \tag{5.5}$$

where Y is a scaling constant of order 1. Remarkably, this holds for many different assets and markets. This is illustrated in Figure 9 from Loeb (1983) for the bid-ask spreads quoted by market-making specialists on NASDAQ in the 1980s.

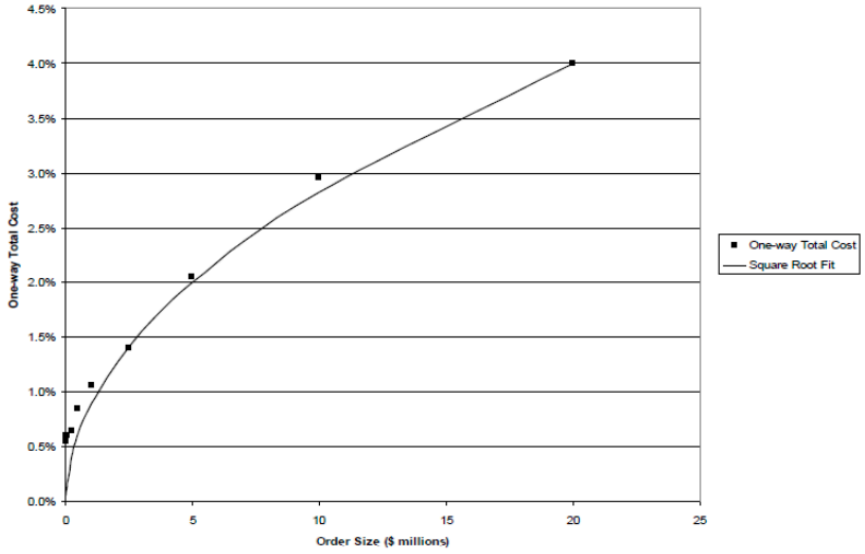
Figure 10 taken from an AQR research report finds a completely analogous result for international stocks traded via electronic limit order books in 2018.

Finally, Figure 11 present the corresponding fit by CFM researchers for trades in Bitcoin, which again display virtually the same behavior.

How to apply the square-root formula (5.5)? Trading volume is published regularly and therefore easy to observe, and volatility is relatively easy to estimate from time series of prices. In contrast, estimating the multiplier Y is more tricky and requires proprietary trade data. Even with such data, it is challenging to pin down its time evolution, as trades at different time points need to be averaged to get rid of noise. However, estimates for Y appear to be remarkably stable over time and close to 1. Therefore, this scaling constant can be ignored at first order.

Many further challenging questions for research and practice remain open in this context. For example, it is crucial to understand not just how each trade impacts the current market price, but also whether and how this impact decays after the metaorder is executed. This depends on information content of the order: if the order reveals new information, prices should adjust permanently. If not, then liquidity should recover and prices return more or less to their original levels eventually.

Another crucial issue subject to much current research is "cross price impact", that is, the effect that trading one security has on the prices of other correlated assets. Multivariate



US Stocks, Loeb

Figure 9: Estimated price impact cost as a function of order size and square-root fit.

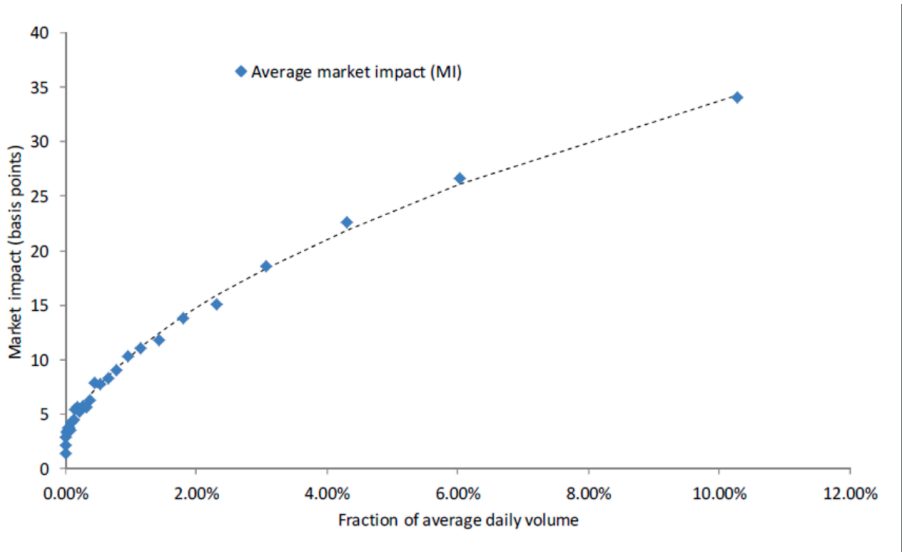


Figure 10: Estimated price impact cost as a function of order size and square-root fit.

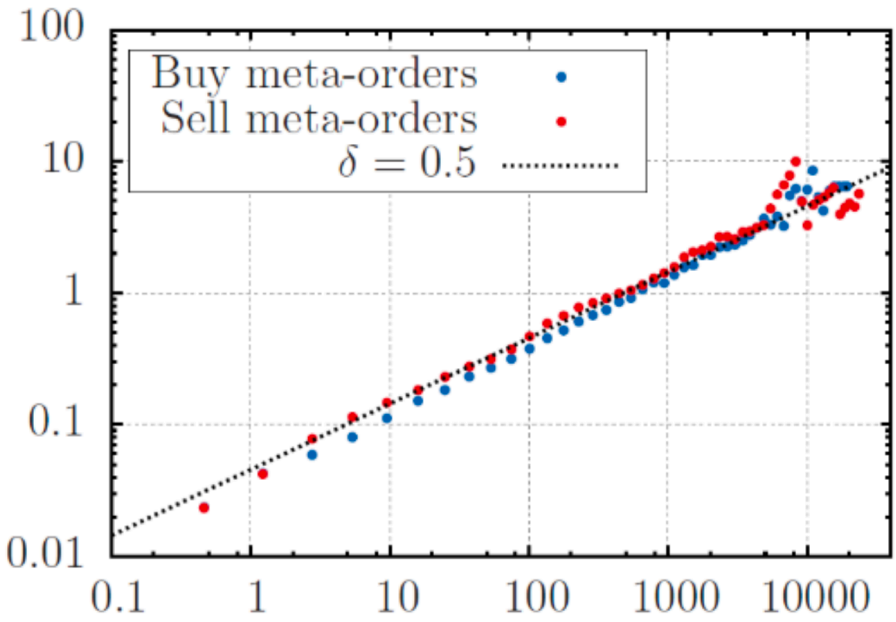


Figure 11: Estimated price impact cost as a function of order size and square-root fit (in log-log scale).

versions of the Kyle model suggest linear price impact proportional to a product of covariance matrices of prices and trading volume. But no multivariate analogue of the square-root law has been identified so far.

6 Equity Strategies

In this section, we dicuss a number of the most commonly used trading strategies for equities (i.e., stocks). In each case, the goal is to identify which stocks have high and low expected returns. The stocks with high-expected returns are then bought, potentially by shorting some of the stocks with low expected returns. Active investors with no-shorting constraints (such as pension funds) instead overweight the stocks with high expected returns relative to their relative market capitalizations, and underweight the stocks whose forecasted expected returns are low.

We will fist have a look at *discretionary equity investments*. This means that traders buy and sell stocks based on their "discretionary" views, i.e., their overall assessment of the stocks they have analyzed. This analysis can include equity valuation models, discussions with management, and many other sources of information.

Discretionary equity investors typically buy more stocks than they sell short. The opposite is true for *dedicated short bias* hedge funds, which focus on shorting stocks that they expect to decline markedly. They typically look for signs of fraud, overstated earnings, or poor business plans.

Whereas the trading decisions in these two types of strategies are ultimately based on human judgements, *quantitative trading strategies* invest systematically based on a formal model. Typically, the goal is to develop a small edge on many small diversified trades that cannot be easily analyzed using non-quantitative methods. To this end, quantitative traders use tools from statistics, engineering, computer science, and finance together with lots of data to identify relationships that have not been incorporated in prices immediately. They then develop algorithms that generate trading signals based on these relations, take into account trading costs in an appropriate manner, and route trades to different exchanges in an efficient way. In other words, quantitative trading is done by feeding data into computers that run various programs with human oversight.

The advantage of discretionary strategies compared to quantitative ones is that they can much more easily take into account "soft information" such as personal conversations. (Even though this edge is being diminished with the advent of textual analysis, for example.) However, the extensive manual labor these strategies require limits the number of securities that can be analyzed. Moreover, human discretion opens the door to a number of behavioral biases.

6.1 Equity Valuation and Investing

All equity strategies rely on evaluating the future prospects of different stocks. In the words of Warren Buffet:

> *Intrinsic value is an all-important concept that offers the only logical approach to evaluating the relative attractiveness of investments and businesses. Intrinsic value can be defined simply: it is the discounted value of the cash that can be taken out of a business during its remaining life.*

Intrinsic Value and the Dividend Discount Model The "value" of a stock is often called *intrinsic value* (or *fundamental value*) to distinguish it from the market price of the stock. Believers in the efficient market hypothesis consider price and value to be identical. In contrast, "value investors" search for stocks whose market price they deem cheap relative to the intrinsic value.

Let us write V_t for the intrinsic value of a stock at time t. The latter is ultimately derived from the cash flows that are returned to share holders. As is customary, we refer to these cash flows as "dividends", but note that this can also include other cash transfers to shareholders (such as share repurchases). Of course, we cannot just add up future dividends because we need to take into account both the time value of money and compensation for uncertainty. The intrinsic value V_{t-1} at time $t-1$ depends on the next dividend D_t, the intrinsic value V_t after the latter is paid, and the *discount rate* k_t (sometimes also called *required rate of return*):

$$V_{t-1} = \mathrm{E}_{t-1}\left[\frac{D_t + V_t}{1+k_t}\right].$$

Iterating this procedure, we obtain

$$V_{t-1} = \mathrm{E}_{t-1}\left[\frac{D_t}{1+k_t} + \frac{D_{t+1}}{(1+k_t)(1+k_{t+1})} + \ldots\right] = \mathrm{E}_{t-1}\left[\sum_{s=t}^{\infty} \frac{D_s}{\prod_{u=0}^{s-1}(1+k_{t+u})}\right]. \tag{6.1}$$

This is exactly the same relationship we have seen in Section 4.2 for equilibrium *prices* in asset pricing models, where the discount factors $\xi_{t-1,t} = (1 + k_t)^{-1}$ were derived from the aggregate uncertainty in the economy and the risk preferences of the representative agent. However, even in a model like this, the valuation of each individual agent may still differ from the market valuation. Let us illustrate this important point with a simple toy model where individual agents have different beliefs about the distribution of future dividends.

Heterogenous Beliefs and Valuations For simplicity, consider a single-period economy with a single risky asset. The asset is traded by two types of agents, which have the same exponential utility function $U(x) = -\exp(-\gamma x)$ but different beliefs about the liquidating dividend D_t the asset will pay at the terminal time t. To wit, conditional on the information available at time $t-1$, the more optimistic agent 1 believes that $D_t \sim \mathcal{N}(\mu_1, \sigma^2)$ whereas the more pessimistic agent 2 believes that $D_t \sim \mathcal{N}(\mu_2, \sigma^2)$ where $\mu_2 < \mu_1$.

Then, the optimization problem of agent n is to maximize

$$\begin{aligned}
\mathrm{E}^n_{t-1}[U(W_t^{n,x_{t-1}})] &= \mathrm{E}^n_{t-1}\left[U\left(W^n_{t-1}(1+R^f) + x_{t-1}R^e_t\right)\right] \\
&= -\exp\left(-\gamma W^n_{t-1}(1+R^f)\right)\mathrm{E}^n_{t-1}\left[\exp\left(-\gamma x_{t-1}\left(\frac{D_t - P_{t-1}}{P_{t-1}} - R^f\right)\right)\right] \\
&= \exp\left(-\gamma(1+R^f)(W^n_{t-1} - x_{t-1}\right)\mathrm{E}^n_{t-1}\left[\exp\left(-\frac{\gamma x_{t-1}}{P_{t-1}}D_t\right)\right] \\
&= -\exp\left(-\gamma(1+R^f)(W^n_{t-1} - x_{t-1})\right)\exp\left(-\frac{\gamma x_{t-1}}{P_{t-1}}\mu_n + \frac{\gamma^2 x^2_{t-1}}{2P^2_{t-1}}\sigma^2\right).
\end{aligned}$$

By maximizing the quadratic function of the investment x_{t-1} in the exponential, we find that the optimum is

$$\hat{x}^n_{t-1} = \frac{\mu_n P_{t-1} - (1+R^f)P^2_{t-1}}{\gamma\sigma^2}.$$

Recall that $\hat{x}_{t-1}$ is dollar amount invested into the risky asset at time t, so the corresponding number of risky shares is

$$\hat{\varphi}^n_{t-1} = \frac{\hat{x}^n_{t-1}}{P_{t-1}} = \frac{\mu_n - (1+R^f)P_{t-1}}{\gamma\sigma^2}.$$

As a consequence, matching the total number of risky shares $\hat{\varphi}^1_{t-1} + \hat{\varphi}^2_{t-1}$ the agents want to hold to the given supply s requires

$$s = \frac{\mu_1 - (1+R^f)P_{t-1}}{\gamma\sigma^2} + \frac{\mu_2 - (1+R^f)P_{t-1}}{\gamma\sigma^2} = \frac{\mu_1 + \mu_2 - 2(1+R^f)P_{t-1}}{\gamma\sigma^2}.$$

Whence, the market-clearing equilibrium price at time $t-1$ is

$$P_{t-1} = \frac{1}{1+R^f}\left(\frac{\mu_1 + \mu_2}{2} - \frac{\gamma}{2}s\sigma^2\right).$$

Exactly the same result obtains for a single representative agent with aggregate risk aversion $\gamma/2$, who has the same beliefs $(\mu_1 + \mu_2)/2$ about expected dividends as the *average* of the

individual agents. For the more optimistic agent, the asset will then appear undervalued at the market price, whereas it will seem to be overvalued for the more pessimistic agent. Whence, agent 1 will hold more risky shares than agent 2 and, if the beliefs are different enough relative to the agents' risk aversion, then agent 2 will even short the risky asset.

Gordon's Growth Model Putting the dividend discount model (6.1) in action requires three inputs: (i) the distribution of future dividends, (ii) the distribution of future discount factors, and (iii) the dependence between these random variables. Unfortunately, none of these quantities is easy to estimate in practice. To simplify the analysis, it is therefore often assumed that the discount rate is constant. In this case, (6.1) simplifies to

$$V_{t-1} = \sum_{s=1}^{\infty} \frac{\mathrm{E}_{t-1}[D_{t-1+s}]}{(1+k)^s}. \tag{6.2}$$

Then, it only remains to estimate the expected evolution of future dividends. The simplest model for this is a constant (expected) growth rate, $\mathrm{E}_t[D_{t-1+s}] = (1+g)^s D_{t-1}$. Then (given that the discount rate is big enough, $k > g$), (6.3) becomes

$$V_{t-1} = \sum_{s=1}^{\infty} \frac{(1+g)^s}{(1+k)^s} D_{t-1} = \frac{(1+g)}{k-g} D_{t-1}. \tag{6.3}$$

This means that the intrinsic value is higher if current dividends are higher, if the dividend growth rate is higher, or if the required return is lower. This makes sense intuitively. However, even with more flexible econometric models for the evolution of dividends (so-called "vector autoregressive models"), the dividend discount model with a constant discount rate can only explain a modest part of the fluctuations exhibited by real stock prices. This is illustrated in Figure 12, taken from Campbell et al. (1997). This suggests that fluctuations of the stochastic discount factor also play an important role in determining asset prices and their volatilities.

Other Approaches to Equity Valuation Estimating dividends is sometimes difficult, in particular, for new companies that have never paid any dividends so far. Stock analysts then often focus on companies' book values and other accounting variables.

Another standard approach to equity valuation is *relative valuation*, based on other comparable stocks. For example, one can value a stock at $E \times P/E$, where E is the firm's earnings and P/E is the price/earnings ratio of other comparable stocks (e.g., a historical average in the same industry sector). Naturally, one can use the same approach for any other characteristic, as long as the latter is representative of the firm and its future prospects. Of course, relative valuation says nothing about whether the entire market is over or undervalued, but it can be informative about which stocks are expensive or cheap relative to others.

6.2 Discretionary Equity Investing – Value and Quality

Value Investing The idea of "value" investing sounds simple enough – buy securities that appear cheap while possible shorting securities that appear expensive (relative to their

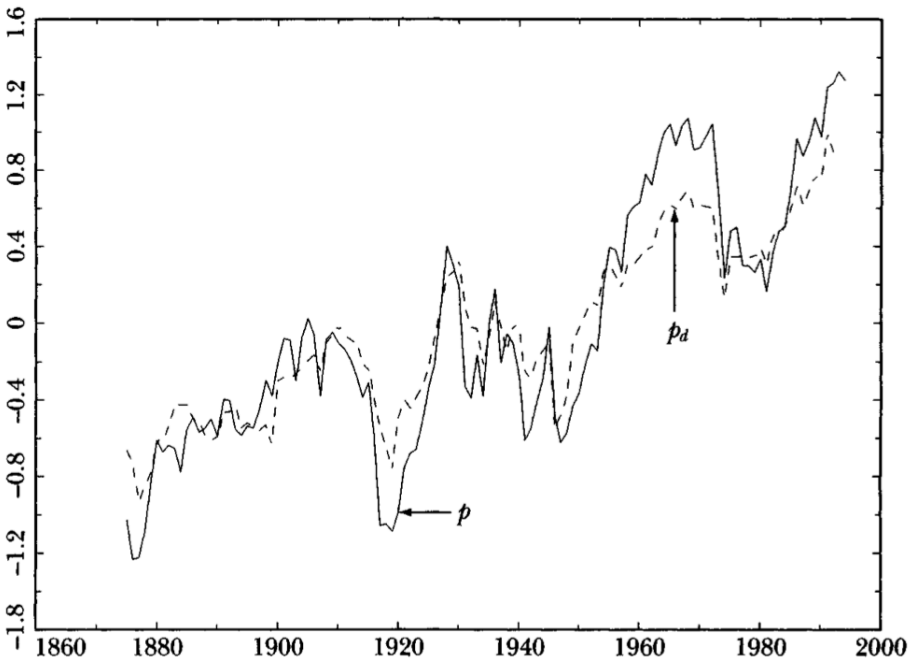


Figure 12: Log real stock price (p) and estimated dividend component (p_d) for annual US data from 1876 to 1994.

intrinsic values). Implementing this in practice is harder than it sounds, though. Indeed, assets are typically expensive for a reason, in that there is something investors love about them (and conversely, something that makes them uncomfortable about cheap stocks). Value investing therefore means going against the conventional wisdom, which is never easy (both psychologically and to justify to management).

There are many different ways to implement value investing. For example, some investors (like Warren Buffet) invest for the long term, seeking to buy stocks for less than the value than the future dividends they will collect over time. Other value investors look to buy cheap stocks and sell them over the medium term, with the hope that the initial pricing error is corrected eventually.

One typical simple implementation of value investing is to buy stocks with a high book value compared to their market value. Even this simple strategy was profitable historically, but has done very poorly in recent years. This is illustrated in Figure 13, which plot the cumulative sum (i.e., without compounding) of the HML factor from Ken French's website(`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html`). The left panel displays the very encouraging performance from 1990 until 2010. The right panel extends the time series to 1990-2020, where the value strategy has clearly performed much more poorly.

Value Trap When you buy a stock with a very low price, e.g., a very low book-to-market ratio, you must always ask yourself the crucial question: does the stock look cheap because

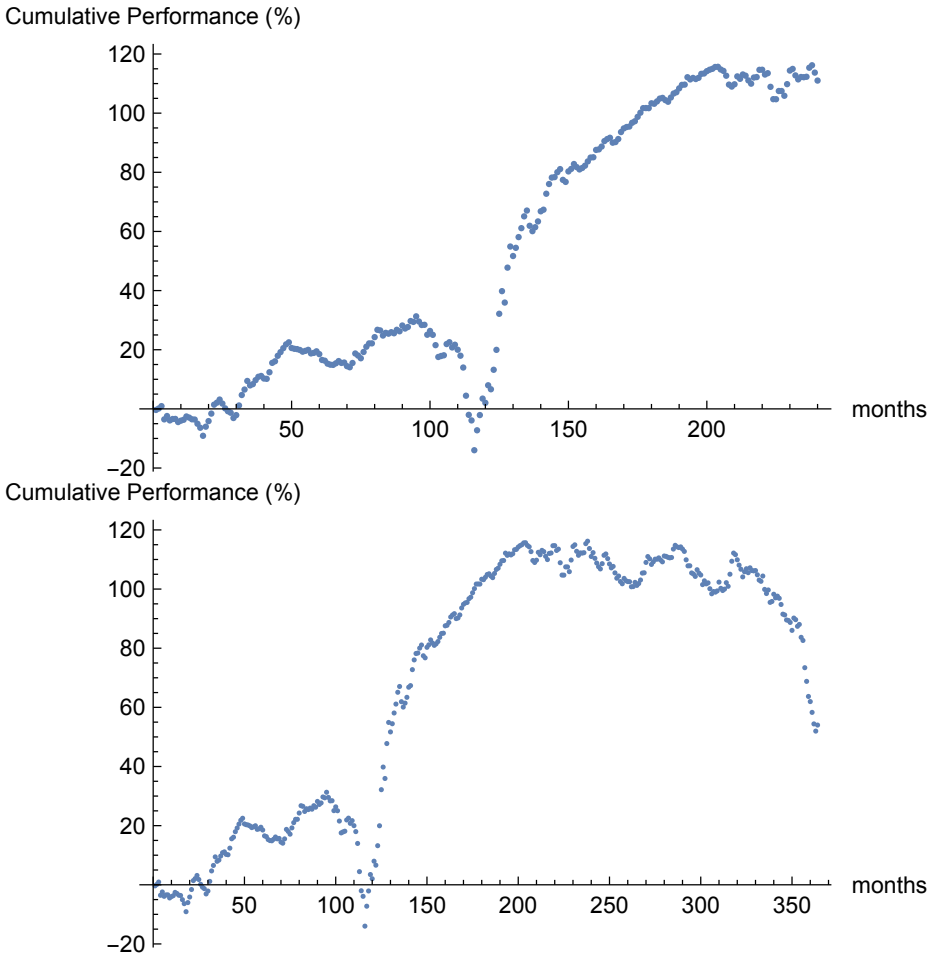


Figure 13: Cumulative performance of the value factor 1990-2010 (upper panel) and 1990-2020 (lower panel).

it *is* cheap, or because it deserves to be cheap? Since stock prices are the outcome of trading between many sophisticated investors, prices reflect a lot of information. Thus, if a stock looks cheap, there is often a good reason for that in that its growth is likely to be low. For example, the price of a bank stock may be low because the market recognizes that it will have to write down many of the loans it has issued. The risk that value investors thereby end up owning fundamentally flawed companies is called the *value trap*.

More generally, suppose that you have bought a stock with an unusually low price-to-book ratio (relative to the historical values of similar firms), because you think this measure of "cheapness" will normalize over time. If this belief is correct, does that mean you are guaranteed a profit? Not necessarily – this depends on what is going to adjust, the price or the book value. If mean-reversion is driven by a rising price, then the trade will indeed make money. However, you may loose money if the mean reversion is driven by a falling book value. This means that the stock experiences low earnings that make it live up to the market's expectations.

Quality Investing A way to at least partially mitigate the value trap is to also focus on some "quality characteristics" of the corresponding stocks. Following Gordon's growth model, typical examples include profitability and payout ratios (which determine dividends) as well as safety (which influences required returns) and growth (in profits!). But lots of other measures are also used in this context, e.g., assessments of how well a company is managed.

Believers in full market efficiency would agree that high-quality firms can be identified, but claim that efficiency implies that the very high prices of such firms already fully reflect these favorable characteristics. In contrast, quality investors believe that it pays to find high-quality firms, because prices do not always fully reflect this information so that their future average returns are high on average.

Value and Quality A prototypical example of combining value and quality investing is Berkshire Hathaway, the investment company run by Warren Buffet. Much of its impressive investment performance over more than 30 years can be explained by a focus on cheap high-quality stocks. In the words of Warren Buffet:

> *Whether we are talking about socks or stocks, I like buying quality merchandise when it is marked down.*

Momentum A different class of popular trading strategies looks for "trends" in the time series of asset returns. For example, time series *momentum* strategies go long when a market has experienced positive excess returns over a certain lookback horizon (e.g., 1, 3, or 12 months), and short otherwise. These strategies produce positive expected returns if market prices exhibit trends, but why should that be the case?

One intuitively appealing mechanism for this that has been proposed is that prices initially underreact to new information, so that trends arise as prices move slowly to more fully reflect the changed fundamental values. These trends have the potential to continue even further, causing an eventual overreaction. Naturally, all trends must come to an end as

deviations from fundamental values cannot continue indefinitely in an efficiently inefficient market.

6.3 Dedicated Short Bias

Whereas most discretionary equity investors focus on buying stocks, a much smaller group of hedge funds focuses on short selling. This relies on many of the same valuation techniques, but the goal now is to identify companies with potential problems. This can include outright fraud, but also overstated earnings, aggressive accounting techniques, or fundamentally flawed business plans.[17]

Shortselling is more challenging than buying stocks for a number of reasons. Indeed, short selling is up against the general headwind of stock prices increasing on average. For example, shorting a stock that goes up by less than the market is in principle a successful trade, but it may not feel that way. More specifically, this trade has a positive alpha and will make money if appropriately hedged, but it looses money if viewed in isolation.

Shorting stocks is also difficult for a number of technical reasons. First of all, short selling is banned in some countries or for some stocks in certain periods of time (e.g., during the financial crisis). Even if it is allowed, short selling requires to find an owner of the share who is willing to borrow it. When the prices of the shorted stocks rise, short sellers may eventually face a margin call (both because their trades lose money when marked to market and because the notational of the trade increases) and might be forced to close their positions. This may lead to a "short squeeze" – as the short sellers buy back the shares, this drives up the prices even further leading to further buying, more margin calls, etc.

Shortselling Restrictions and Bubbles As discussed in a simple model with heterogeneous beliefs in Section 6.1, market prices aggregate the views of more optimistic and pessimistic agents. However, the views of the pessimists are often incorporated through short selling. What happens if this is not allowed?

Recall that in our model with heterogeneous beliefs, the equilibrium price was

$$P_{t-1} = \frac{1}{1+R^f}\left(\frac{\mu_1+\mu_2}{2} - \frac{\gamma}{2}s\sigma^2\right), \tag{6.4}$$

and the pessimistic agent shorts the risky asset if

$$\hat{\varphi}_t^2 = \frac{s}{2} + \frac{\mu_2-\mu_1}{2\gamma\sigma^2} < 0, \tag{6.5}$$

that is, if the relative pessimism is strong enough to overcome the risk premium. If short sales are not allowed in this model, then each agent's optimization is done over positive investments only. For a given initial price level P_{t-1}, this leads to the optimal share holdings

$$\hat{\varphi}_{t-1}^1 = \max\left\{\frac{\mu_1-(1+R^f)P_{t-1}}{\gamma\sigma^2}, 0\right\}, \quad \hat{\varphi}_{t-1}^2 = \max\left\{\frac{\mu_2-(1+R^f)P_{t-1}}{\gamma\sigma^2}, 0\right\}.$$

[17] For example, high current profits based on a technology that is about to become obsolete – as happened to the producers of the fax machine or non-smart cellphones. If and when the same fate awaits traditional car producers is a current open question, for example.

As the no-shorting constraints increases the holdings of each agent relative to the unconstrained model, the equilibrium price is increased in order to still clear the market. This makes it even less attractive for the pessimist to long the risky asset, so that all shares are held by the optimist at the single-agent equilibrium price:[18]

$$P_{t-1}^{\text{NoShort}} = \frac{1}{1+R^f}\left(\mu_1 - s\gamma\sigma^2\right).$$

No-shorting constraints therefore exclude the pessimists from the market here. As a consequence, the market price coincides with the valuation of the optimists, whereas the views of the pessimists are not incorporated. If each agent receives a signal about fundamentals with iid noise, for example, then the constraints clearly make the market less efficient. However, at least the market price is still supported by the fundamental views of *some* market participant here.

This no longer remains true in dynamic models where the agents potentially have a "resale option". If not the same agent is more optimistic in all scenarios, then this can lead equilibrium prices that are higher than the valuation of *any* of the individual agents (Harrison and Kreps, 1978). This can be interpreted as a "bubble"; see, e.g., Scheinkman and Xiong (2003) for more details. Even when shortselling is possible (but costly), the above effects can persist (Nutz and Scheinkman, 2020).

Contrary to popular opinion (for example, shorting has been called "blatant thuggery" in hearings held by the US Congress), shortselling therefore has an important role in making markets *more* efficient. As a case in point, several of the witnesses in the above hearing were later prosecuted for fraud.

6.4 Quantitative Equity Investing

Fundamental Quantitative Investing "Fundamental quants" trade on factors such as value, momentum, quality, size, and low risk. The information used for this is similar to discretionary investors, with the difference that quants try to automate this analysis so that a computer can apply it across thousands of assets around the world in a systematic manner.

Fundamental quant strategies are applied in a number of different contexts, ranging from long-short market neutral funds, to long-only benchmark-driven strategies. The underlying building blocks are the same, namely quantitative estimates for which stocks have high expected returns, which ones have low expected returns, and a risk model.

Quant strategies typically involve hundreds or even thousands of stocks, so that most idiosyncratic risk is diversified away. By balancing long and short dollar investments as well as risk, market-neutral strategies also try to eliminate the overall market risk. In addition, some quants also try to eliminate (some) systematic risks associated to certain regions or industries.

When idiosyncratic, market and industry risks have (hopefully) been eliminated, is there any risk left? Yes, the exposure to the factors the strategy is betting on. If the quant is betting on value, for example, the portfolio risk is that the value factor performs poorly,

[18] This price is indeed bigger than its unconstrained counterpart (6.4) given (6.5), so that it is indeed suboptimal for the pessimist to hold any shares.

i.e., "cheap" stocks become even cheaper and "expensive" ones even more expensive, or if the "cheap" stocks turn out not to be cheap after all relative to deteriorating fundamentals.

If this happens, then leveraged quant investors can face a "liquidity spiral" where many overlapping portfolios need to be liquidated at the same time. Here, the overlap comes from the fact that all quant portfolios try to hold stocks with high expected returns and short those with low expected returns, even though they may do this using very different models. This is what happened during the so-called "quant event" of 2007, where long-short market neutral value and momentum strategies experienced massive losses over a few days, see (Pedersen, 2015, pp. 145–149) for more details.

Statistical Arbitrage "Stat arb" strategies are also quantitative, but typically based less on the analysis of fundamental values and more based on arbitrage relations and statistical relationships.

One classical example are "dual-listed stocks", that is, shares of two companies that have merged but remain listed separately on different exchanges. For example, Unilever (which was formed in a merger in 1930) still consists of two different companies, one based in the Netherlands (with shares traded in Euros) and the other based in the UK (with shares traded in GBP). While the two prices follow each other closely, there is sometimes a significant spread of several percentage points between them. This is displayed in Figure 14 from Pedersen (2015).

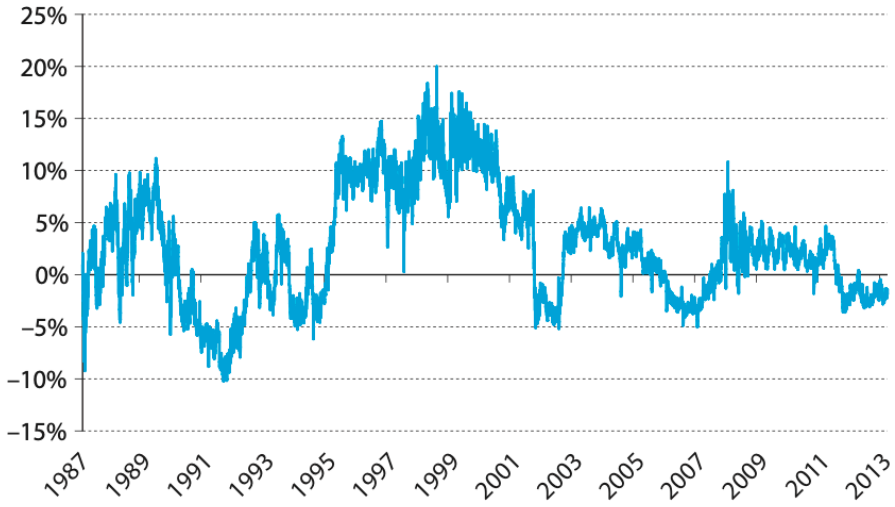


Figure 14: Deviation from parity for Unilever's dual-listed stocks.

Stat arb traders trade on the discrepancies between such "twin stocks". This reduces the arbitrage spreads, but often does not eliminate them completely. Indeed, trading on the spreads requires constant monitoring of the market, understanding the contractual rights of the different shares, implementing the required trades despite transaction costs, hedging currency risk, etc. Whence, while the spread would be pushed to zero in a perfectly efficient

market, efficiently inefficient spreads remain in reality. Often, the less liquid of the twins trades at a discount.

Another well-known stat arb trade focuses on different share classes for the same company. These may have different voting rights, and can also differ substantially in terms of liquidity. Again, the corresponding prices are closely related (and display mean-reverting behavior in the long term), but there can be significant short-term deviations, as illustrated in Figure 15 from Pedersen (2015). In addition, these spreads can also be affected strongly by corporate events such as share repurchases or takeovers.

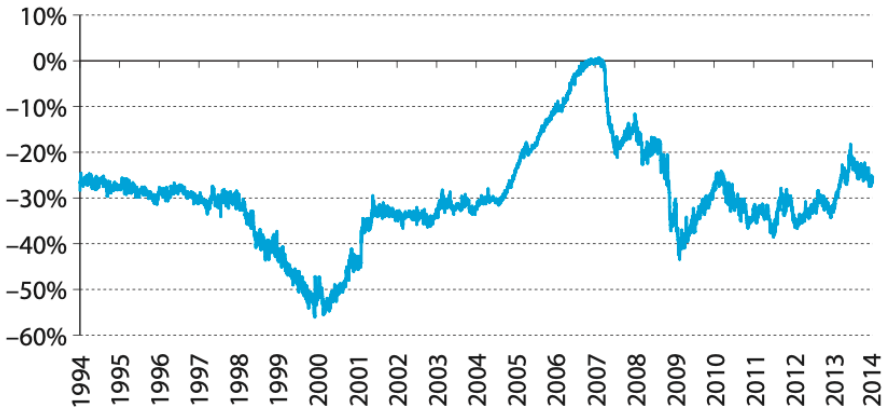


Figure 15: The price discount of BMW preference shares relative to ordinary shares.

A third classical stat arb trade looks at combinations of assets that behave similarly in a statistical sense, without any explicit arbitrage link. A typical example is "pairs trading", where one constructs a long-short portfolio of similar stocks and then bets on its mean-reversion, i.e., price reversals. The same approach is also applied to long-short portfolios of bonds with different maturities, for example.

High-Frequency Trading Another important class of quant strategies is *high-frequency trading (HFT)*. As implied by the name, these strategies operate at much higher frequencies than the other quant strategies. Conversely, they typically trade much smaller positions. Some HFTs provide liquidity by continually updating electronic limit orders. Others take liquidity by exploiting ultra-fast information feeds that provide them with access to information about price changes in other exchanges that are not public knowledge yet ("latency arbitrage"). Another (infamous) strategy is "frontrunning", which seeks to identify and exploit large orders traded by other market participants. Such trading strategies are discussed in more detail in the electives *High-Frequency Trading* and *Market Microstructure*.

7 Arbitrage Pricing and Trading

The textbook definition of an *arbitrage* refers to a trade that can be set up without any initial cash investment, never incurs any losses, and delivers strictly positive gains with some

positive probability. This is typically implemented by selling an (overvalued) security (or portfolio of securities) to buy another (undervalued) security. If the undervalued security is less expensive but always produces the same or better cash flows going forward, then this is indeed an arbitrage.

Such literal arbitrages almost never exist in the real world. Practitioners in turn often use the word "arbitrage" for trades that come close to this, in that buying and selling closely related securities does not guarantee a profit without risk, but nevertheless can be done at attractive relative prices. While such "buy low, sell high" trades can often be expected to be profitable, they typically do require some cash outlay (e.g., for margin requirements), can lead to substantial losses before they "converge", and often incur a non-negligible risk of not converging at all. Such "arbitrage opportunities" therefore in fact often are just compensation for liquidity risk or "deal risk" in connection with corporate events.

7.1 General Arbitrage-Pricing Framework

The (relative) pricing of securities by imposing the absence of arbitrage is ubiquitous in finance going back to the works of Black, Merton, and Scholes on option pricing (Black and Scholes, 1973; Merton, 1973). The basic idea is the following:

1. If there are no arbitrage opportunities and two securities have the same payoff, then they must have the same price at all times.

2. If a (static) portfolio composed of traded assets has the same payoff as another security, then the only price for the security that is compatible with no arbitrage is the value of this *replicating portfolio*.

3. If a (dynamic) *self-financing* portfolio has the same final payoff as a security, then the initial price of the security has to be the same as the initial value of setting up this *dynamic hedging strategy*.

Whence, if we can find a way to replicate a security – by another security or a replicating portfolio – then the security's price is pinned down by the absence of arbitrage. Indeed, if it would trade at any other price, then we could make a riskless profit by buying low and selling high as described above. In the standard arbitrage pricing theory, it is assumed that this never happens because such mispricings are immediately exploited and thereby eliminated.

However, in the real world, investors face transaction costs and funding costs, so that arbitrage trades involve costs and are almost never entirely risk free. The three types of arbitrage trades described above are increasingly influenced by trading costs. While types 1 and 2 involve buy and hold strategies (so that transaction costs only need to be paid when setting up and liquidating the positions), type 3 requires dynamic trading which incurs much higher transaction costs. As a consequence, arbitrage relations based on *dynamic* hedging can more easily break down in an efficiently inefficient market.

7.2 Option Arbitrage

The most well-known example of arbitrage pricing concerns *derivative* securities, whose payoffs are determined by some *underlying* primary assets. The standard examples are call

or put options written on equity indices or individual stocks. If the price of the underlying at time $t = 0, \ldots, T$ is denoted by S_t, then the payoffs of put and call options with maturity T strike price K are

$$P_T = \max\{K - S_T, 0\}, \quad C_T = \max\{S_T - K, 0\}.$$

Options are used for a number of reasons. For example, one can hedge the risk of market crashes by buying "out-of-the-money" put options with smaller strike K than the current market price S_0. These will then expire worthless if the market does not drop below K, but cover any additional losses beyond that threshold. Whence, the price paid for the put option is an insurance premium for crash risk.

Another reason for trading options is that they offer "embedded leverage" To wit, for the same amount of money, you can buy many more call options than stocks, yet both investments offer the same upside in case of large market rises. Of course, risk and return are related and the risk of losing all your money also far greater with options, just like with a leveraged portfolio.

Put-Call Parity The most well-known arbitrage-pricing relation for European call and put options is the *put-call parity*. The idea is to create a synthetic "stock" by buying a European call option and shorting a European put option. At maturity T, this creates the payoff

$$C_T - P_T = S_T - K.$$

In order to match the cash flows of a stock that pays no dividends, we therefore need to add enough cash to the replicating portfolio to have K dollars at hand at maturity T. With a (simply compounded) risk-free rate R^f, this requires an initial cash investment of $K/(1 + R^f)^T$. In the absence of arbitrage, the value of the replicating portfolio must then match the value of the stock at all times, which leads to the following put-call parity:

$$C_t - P_t = S_t - \frac{K}{(1 + R^f)^{T-t}}.$$

This arbitrage-pricing relationship usually holds well in the real world, because it does not require dynamic trading. An exception to this is when some of the securities involved are difficult to short.

Carr-Madan Formula The same kind of static replication argument in fact works for *any* European option with smooth payoff $G(S_T)$, at the cost of requiring positions in calls and puts with *all* strikes. Indeed, for any $\bar{s} > 0$, it holds that (Carr and Madan, 2001):

$$\begin{aligned} G(s) = G(\bar{s}) + G'(\bar{s})(s - \bar{s}) \\ + \int_0^{\bar{s}} G''(K) \max\{K - s, 0\} dK + \int_{\bar{s}}^{\infty} G''(K) \max\{s - K, 0\} dK. \end{aligned}$$

If one discretizes the integrals on a grid with meshwidth ΔK, the option payoff one the left hand side can then be replicated by a static position of $G(\bar{s}) - G'(\bar{s})\bar{s}$ in cash, $G'(\bar{s})$ shares of the underlying, and $G''(K)\Delta K$ shares of a range of calls and puts with strikes K.

In reality, calls and puts with extreme strikes are not liquidly traded, so the usefulness of this representation hinges on beeing able to obtain a good approximation with the available options.

Arbitrage with Dynamic Hedging Most exotic derivative securities depend on the whole path of the underlying and cannot be replicated by static positions in simple options. At best, they can be replicated by *dynamic* trading in the underlying, the risk-free asset, and potentially some further liquidly traded call and put options.

The most well known examples for this are the (discrete-time) binomial model and the (continuous-time) Black-Scholes model. Both of these models are *complete*, in that *any* derivative can be replicated by dynamic trading in the risk-free asset and the underlying alone. Whence, absence of arbitrage pins down a unique price for all such securities. If the quoted option price deviates from this value, then this gives rise an arbitrage opportunity of type 3. This is more difficult to exploit in real life because it involves dynamic trading.

Moreover, in more realistic models (e.g., already the trinomial model in discrete time or stochastic volatility models in continuous time) it is not possible to replicate most options by trading in the underlying and the risk-free asset. Such markets can often be *completed*, in that any derivative can be replicated if a number of call and put options are available for dynamic trading as well. However, Moreover, resulting hedging strategies then depend on the chosen model, and therefore may not be accurate in real life. Moreover, the transaction costs for options are substantially larger than for the underlying stocks, so implementing dynamic hedging strategies based on options in practice is not easy.

7.3 Fixed-Income Arbitrage

Another natural habitat for arbitrage trades are fixed income markets. The reason is that, like option markets, they contain a large number of closely related securities that must obey relationships similar to the put-call parity. Indeed, almost all interest-rate derivatives depend heavily on the risk-free rate, and fixed income arbitrage traders then try to exploit relative mispricings. In doing so, a lot of the common risk can he hedged away by going long and short. However, the intense competition on fixed-income markets and the limited risk involved imply that the available trading opportunities are typically small in an efficiently inefficient market. Achieving high returns then requires significant leverage. Such highly leveraged arbitrage trades can earn moderate profits when they converge ("picking up nickels"), interrupted by occasional dramatic losses when many arbitrageurs have to liquidate their positions at the same time ("the steamroller"). The collapse of LTCM in 1998 is the most well-known example for this.

On-the-run vs. Off-the-run Let us illustrate fixed-income arbitrage by one classic arbitrage trades for government bonds (also pursued by LTCM, for example) that is based on a long-short portfolio of on-the-run and off-the-run treasury bonds.

On-the-run treasuries are newly issued government bonds. Since they have just been issued, they are heavily traded and are thus easy to buy and sell. Furthermore, leveraged investment in on-the-run treasuries are easy to finance because lenders like this liquid collateral. *Off-the-run treasuries* are old bonds, and they tend to have worse market and funding

liquidity. As a result, off-the-run treasuries are cheaper and offer higher yields.[19] The "yield spread" between off-the-run and on-the-run treasuries fluctuates considerably over time, ranging from virtually zero to almost a percentage point i the US from 1990-2010, for example. The spread is low in times of generally high market liquidity, but spikes in times of crisis when investors try to switch to more liquid assets. This is illustrated in Figure 16, taken from Pedersen (2015).

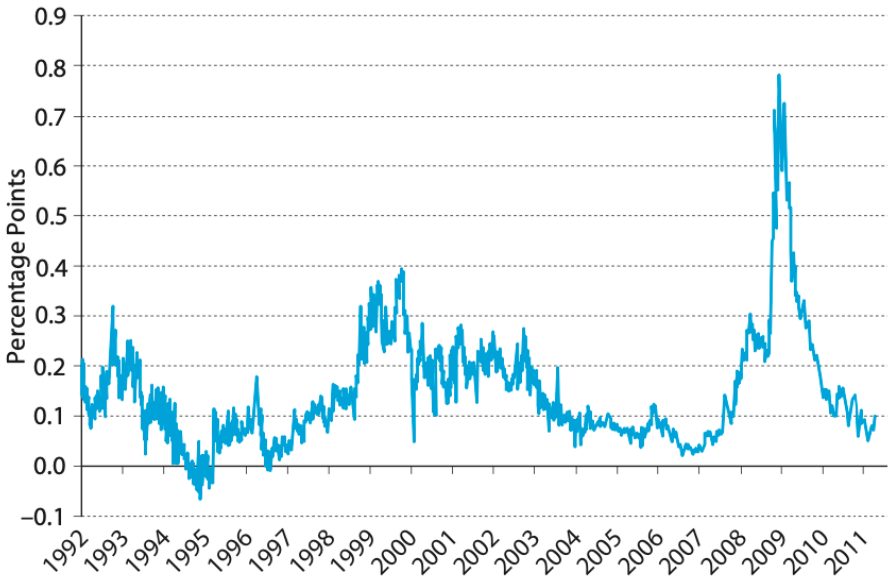


Figure 16: The on-the-run/off-the-run yield spread for 10-year US treasure bonds.

The typical on-the-run/off-the-run trade buys the "cheap" off-the-run treasuries sell short "expensive" on-the-run treasuries. This earns the yield spread, but exposes the trader to the corresponding market liquidity and funding liquidity risk.

7.4 Convertible Bond Arbitrage

Convertible Bonds Another typical fixed-income arbitrage trade based more heavily on option-pricing theory focuses on *convertible bonds*. These are corporate bonds that can be converted into a certain number of stocks of the underlying. Whence, convertible bonds are similar to standard bond combined with American call option. But the payoffs are not quite the same, since the stock is exchanged for the bond itself rather than cash. Moreover, convertible bonds typically also are "callable" (i.e., can be redeemed by the issuer before maturity).

Why do firms use convertible bonds for financing? One reason is that they provide financing that strikes a balance between bonds and equity. To wit, selling convertible bonds dilutes the existing equity of a company less than issuing new stocks. At the same time, convertible bonds replace some of the coupon payments of straight bonds by optionality.

[19]The *yield* of a bond is the internal rate of return if you hold the bond until maturity.

Another reason for the popularity of convertible bonds is that they can usually be sold very quickly (less than one day).

Pricing anf Hedging Convertibles Convertible bonds are priced similarly as American options. In continuous-time models like Black-Scholes, this leads to PDEs; for discrete-time models such as binomial trees, prices can be computed by backward induction.

Convertible bond arbitrage trades typically buy convertible bonds that appear cheap. Then, the resulting risk is (partly) hedged by a (dynamically adjusted) short position in stocks (like for American calls). We don't go into the risk-neutral option pricing details here. Instead, we briefly discuss the intuition, and implications for arbitrage trades. Figure 17 from Pedersen (2015) displays convertible bond prices without default risk for the underlying company. The option price is clearly convex, just like for American calls.

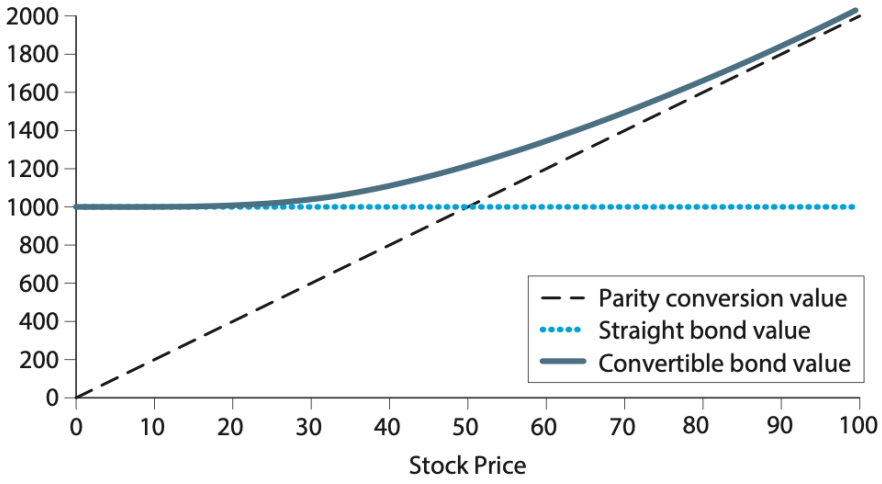


Figure 17: Price of a convertible bond without default risk.

This is no longer true if one takes into account that the company issuing the convertible can default. Figure 18 (also from Pedersen (2015)) displays the price of a convertible bond for a company that defaults if if market value drops below fixed liabilities of 100,000,000 (with equal seniority of all debt assumed for simplicity). The default risk then is similar to being short a put option, and the price of the convertible in turn has convex and concave regions like a long-short portfolio of calls and puts.

To hedge convertible bonds, the risk of price changes for the underlying can be mitigated by delta hedging as usual. In particular, the hedge ratio then is the slope of the option price, just as for other options. This approaches the conversion ratio for very high stock prices. Conversely, it drops for lower stock prices, but can pick up again when default becomes a concern. This is illustrated in Figure 19 from Pedersen (2015).

Like for American options, a key question is ofd course when to convert a convertible. A

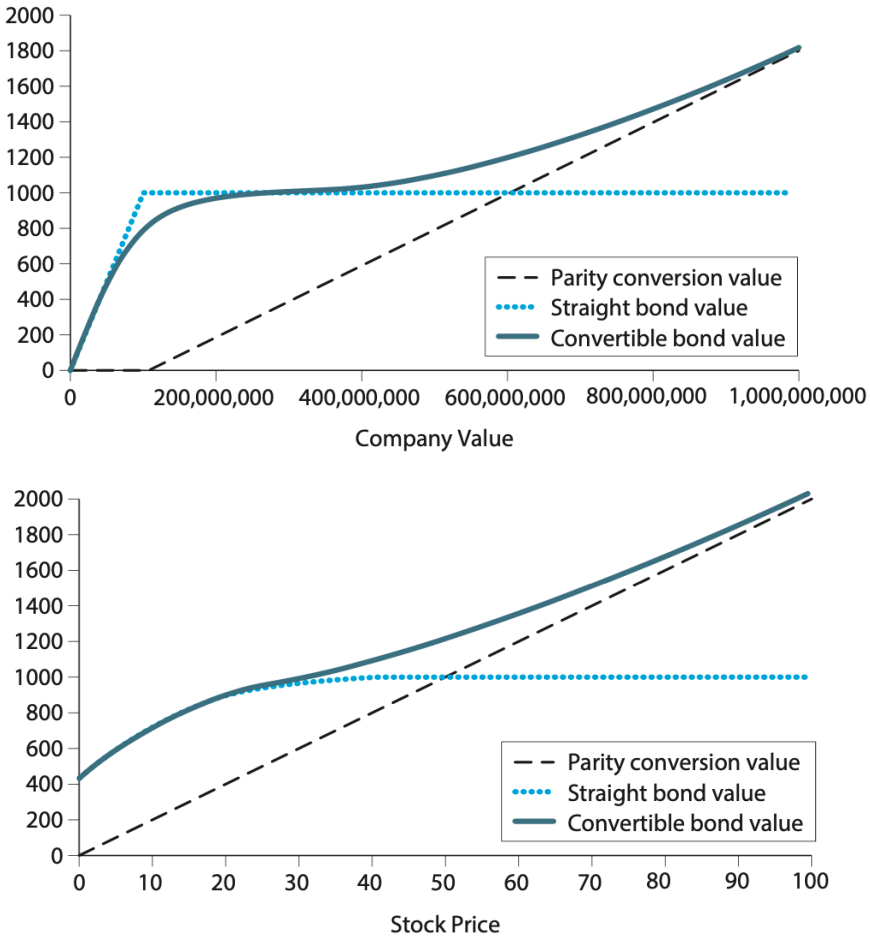


Figure 18: Price of a convertible bond with default risk, plotted against company value (left panel) and against the stock price (right panerl).

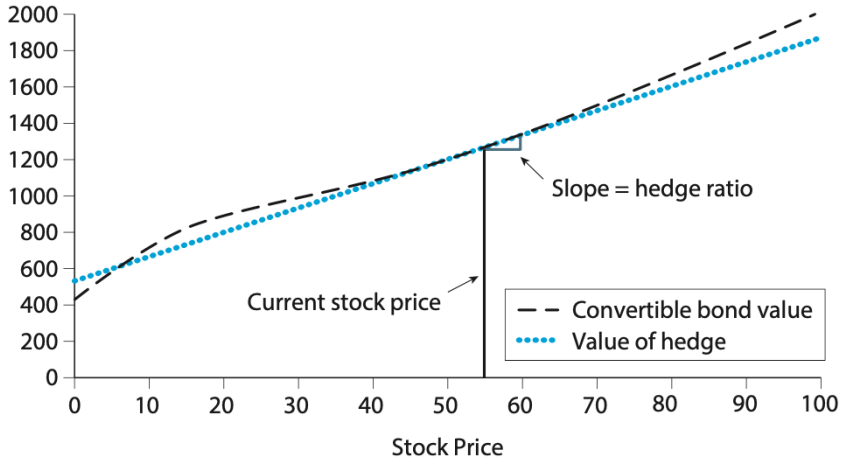


Figure 19: Delta hedge for a convertible bond.

folk theorem says *never*, like for American calls, see Ingersoll (1977) for more details. However, this need not remain true with dividends, because it can pay off to convert just before dividend date. Similarly, early conversion can also be optimal due to financial frictions. For example, one may want to convert early to save funding or transaction costs incurred by the hedging strategy, see Jensen and Pedersen (2016).

Let us now summarize the cash flows of holding a convertible bond. As the holder, you receive coupon payments, but (typically) pay funding costs for leverage as well as dividends and shorting costs for the hedge position.However, the primary driver of profits and losses are price changes of the underlying stock. If the hedge is adjusted discretely (say, once per day), then a surprising feature is that if stock price rises and then drops by the same amount, you can make money! The reason for this is the convexity of the price! Indeed, in the region where the option price is convex, you need to short more after price goes up. As a result, you benefit more when the price goes down again as illustrated in Figure **??** from Pedersen (2015).

Does this mean you can never loose money? Of course not. For example, price fluctuations cause losses occur when the stock price leaves the region where the convertible bond price is convex (where default risk becomes a concern). Moreover, you also loose money if the stock price does not move enough and the value of the conversion option decays over time. In summary, the P&L of holding a convertible therefore typically depends on how price moves you get relative to how many you implicitly paid for. As a consequence, value of the convertible also increases when stock price becomes more volatile.

Convertible-Bind Arbitrage Historically, convertibles trade at a discount relative to the value of their components (bond+option). Why? One reason is that, as illustrated by our discussion above, convertible bonds require expertise to trade. Moreover, they have large transaction costs and face substantial market liquidity risk. Likewise, convertibles are

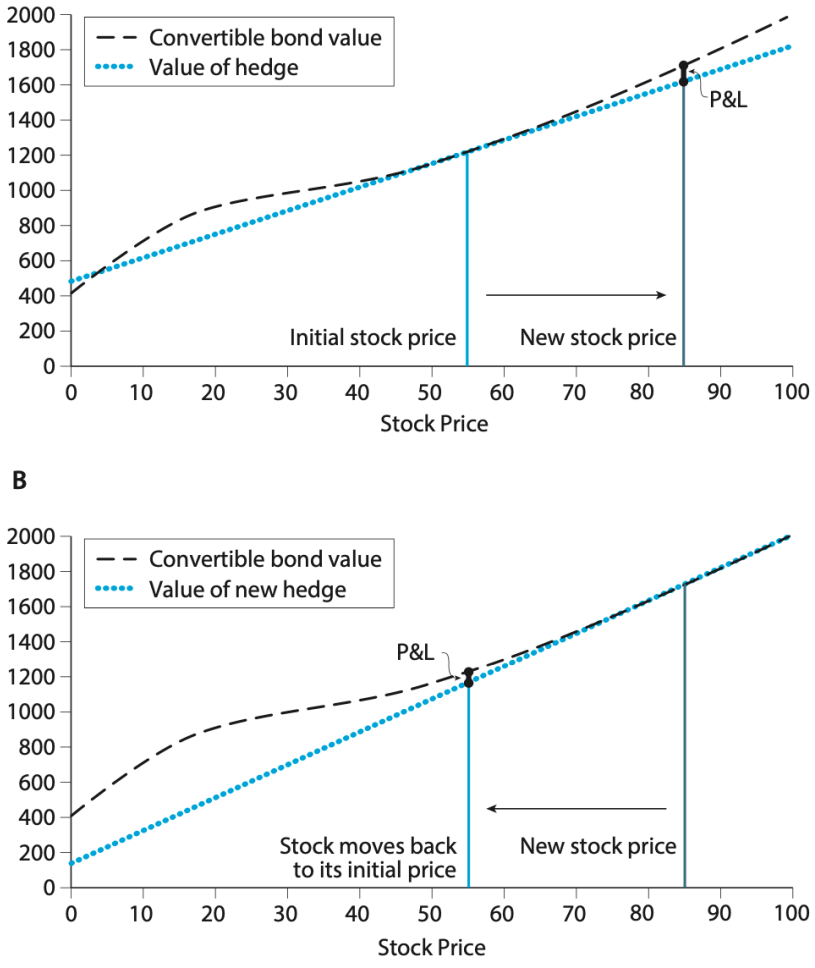


Figure 20: Profits generated by price fluctuations.

difficult and expensive to finance, and face substantial funding liquidity risk. Conversely, quick sales of convertibles provide fast liquidity, so theissuers are willing to accept a discount for that.

In summary, convertible bond arbitrages thus earn premia for a number of different liquidity risks. Unlike market risk, interest rate risk and (to a certain extent) credit risk, these risks are difficult to hedge. Moreover, all of them are interrelated leading to yields of certain convertibles rose *above straight bonds* during the height of the 2008 financial crisis, despite the extra optionality offered by the convertibles, see Figure 21 from Pedersen (2015)! In particular, the trades are not free lunches, but efficiently inefficient compensation for certain risks.

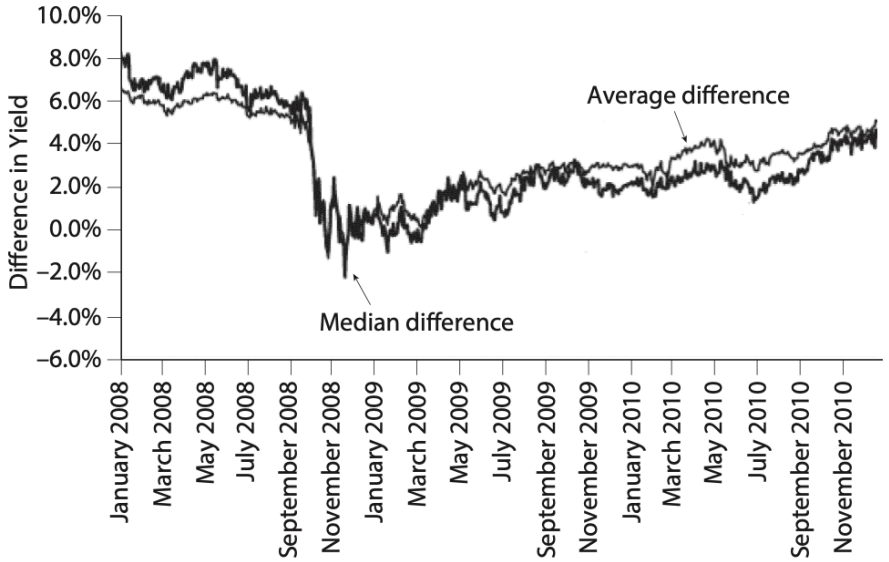


Figure 21: Yield spread between straight bonds and convertible bonds.

References

Acharya, V. V. and Pedersen, L. H. (2005). Asset pricing with liquidity risk. *J. Fin. Econ.*, 77(2):375–410.

Almgren, R. F. and Chriss, N. (2001). Optimal execution of portfolio transactions. *J. Risk*, 3:5–40.

Almgren, R. F., Thum, C., Hauptmann, E., and Li, H. (2005). Direct estimation of equity market impact. *RISK*, July.

Amihud, Y. (2002). Iliquidity and stock returns: cross-section and time-series effects. *J. Fin. Markets*, 5(1):31–56.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *J. Pol. Econ.*, 81(3):637–654.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton University Press, Princeton, NY.

Carhart, M. M. (1997). On persistence in mutual fund performance. *J. Finance*, 52(1):57–82.

Carr, P. and Madan, D. (2001). Optimal positioning in derivative securities. *Quant. Finance*, 1(1):19–37.

Cont, R. and Sirignano, J. (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quant. Finance*, 19(9):1449–1459.

Duffie, D. (2001). *Dynamic asset pricing theory*. Princeton University Press, Princeton, NY.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Fin. Econ.*, 33(1):3–56.

Garleanu, N. and Pedersen, L. H. (2011). Margin-based asset pricing and deviations from the law of one price. *Rev. Fin. Stud.*, 24(6):1980–2022.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *Am. Econ. Rev.*, 70(3):393–408.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Rev. Fin. Stud.*, 33(5):2223–2273.

Harrison, J. M. and Kreps, D. M. (1978). Speculative investor behavior in a stock market with heterogeneous expectations. *Quart. J. Econ.*, 92(2):323–336.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *Rev. Fin. Stud.*, 33(5):2019–2133.

Ingersoll, J. E. (1977). A contingent-claims valuation of convertible securities. *J. Fin. Econ.*, 4(3):289–321.

Jensen, M. V. and Pedersen, L. H. (2016). Early option exercise: Never say never. *J. Fin. Econ.*, 121(2):278–299.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Linnainmaa, J. T. and Roberts, M. R. (2018). The history of the cross-section of stock returns. *Rev. Fin. Stud.*, 31(7):2606–2649.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *J. Finance*, 20(4):587–615.

Loeb, T. F. (1983). Trading cost: The critical link between investment information and results. *Financial Analysts J.*, 39(3):39–44.

Merton, R. C. (1973). Theory of rational option pricing. *Bell J. Econ.*, 4(1):141–183.

Muhle-Karbe, J., Reppen, M., and Soner, H. M. (2017). A primer on portfolio choice with small transaction costs. *Ann. Rev. Fin. Econ.*, 9:301–331.

Muhle-Karbe, J., Shi, X., and Yang, C. (2020). An equilibrium model for the cross-section of liquidity premia. Preprint, available online at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3738500`.

Nutz, M. and Scheinkman, J. A. (2020). Shorting in speculative markets. *J. Finance*, 75(2):995–1036.

Pedersen, L. H. (2015). *Efficiently inefficient: how smart money invests and market prices are determined*. Princeton University Press, Princeton, NY.

Scheinkman, J. A. and Xiong, W. (2003). Overconfidence and speculative bubbles. *J. Pol. Econ.*, 111(6):1183–1220.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance*, 19(3):425–442.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Voodoo correlations in social neuroscience. *Perspect. Psychol. Sci.*, 4(3):274–290.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.