

The Elements of Quantitative Investing

Giuseppe A. Paleologo

June 21, 2024

Contents

Contents	2
Preface	9
I Before the Trade	17
1 The Map and the Territory	19
1.1 The Securities	20
1.2 Modes of Exchange	22
1.3 Who Are the Market Participants?	23
1.3.1 The Sell Side	24
1.3.2 The Buy Side	26
1.4 Where Do Excess Return Come From?	29
1.5 The Elements of Quantitative Investing	32
2 Returns: Properties and Models	37
2.1 Returns	38
2.1.1 Definitions	38
2.1.2 Excess Returns	39
2.1.3 Log Returns	39
2.1.4 Estimating Prices and Returns	41
2.1.5 Stylized Facts	42
2.2 Conditional Heteroscedastic Models (CHM)	46
2.2.1 GARCH as random recursive equations*	47
2.2.2 GARCH(1,1) and Return Stylized Facts	48
2.2.3 *GARCH(1,1) Estimation	51
2.2.4 Realized Volatility	52
2.2.5 Combining CHM and Realized Volatility	55
2.3 State-Space Estimation of Variance	56
2.3.1 Muth's Original Model: EWMA	56

2.3.2	The Harvey-Shephard Model★	59
2.4	Further Reading	61
2.5	★Appendix	62
2.5.1	The Kalman Filter	62
2.5.2	Kalman Filter Examples	64
2.6	Exercises	67
3	Linear Models of Returns: The Basics	71
3.1	Factor Models	72
3.2	Interpretations of Factor Models	74
3.2.1	Graphical Model	75
3.2.2	Superposition of Effects	76
3.2.3	Single-Asset Product	76
3.3	Alpha Spanned and Alpha Orthogonal	77
3.4	Transformations	80
3.4.1	Rotations	80
3.4.2	Projections	81
3.4.3	Push-Outs	82
3.5	Applications	83
3.5.1	Performance Attribution	83
3.5.2	Risk Management: Forecast and Decomposition	84
3.5.3	Portfolio Management	86
3.5.4	Alpha Research	86
3.6	Factor Models Types	88
3.7	Further Reading	88
3.8	★Appendix	89
3.8.1	Linear Regression	89
3.8.2	Linear Regression Decomposition	93
3.8.3	The Frisch-Waugh-Lovell Theorem	93
3.8.4	The Singular Value Decomposition	96
3.9	Exercises	99
4	Portfolio Management: The Basics	103
4.1	Why Mean-Variance Optimization?	104
4.2	Mean-Variance Optimal Portfolios	106
4.3	Trading in Factor Space	111
4.4	Trading in Idio Space	111
4.4.1	Drivers of Information Ratio: Information Coefficient and Diversification	112

4.5	Investment Performance Metrics	115
4.5.1	Expected Return	116
4.5.2	Volatility	116
4.5.3	Sharpe Ratio	117
4.5.4	Capacity	118
4.6	*Appendix	121
4.6.1	Convex Optimization	121
4.6.2	Duality	122
4.6.3	Local Analysis	124
4.6.4	Solutions To Specific Optimization Problems	128
5	MVO and Its Discontents	129
5.1	Shortcomings of Naïve MVO	129
5.2	Constraints and Modified Objectives	133
5.2.1	Types of Constraints	133
5.2.2	Do Constraints Improve or Worsen Performance?	137
5.2.3	Constraints as Penalties	138
5.3	How Does Estimation Error Affect Sharpe Ratio?	142
5.3.1	The Impact of Alpha Error	144
5.3.2	The Impact of Risk Error	145
5.4	Trading Sharpe For Capacity	146
5.5	*Appendix: Theorems on Sharpe Efficiency Loss	147
6	Evaluating Alpha	151
6.1	Backtesting Best Practices	152
6.2	The Backtesting Protocol	158
6.2.1	Cross-Validation and Walk Forward	158
6.3	The Rademacher Anti-Serum	163
6.3.1	Setup	163
6.3.2	Main result and Interpretation	165
6.4	*Appendix: Proofs	169
7	Evaluating Risk	173
7.1	Evaluating Alpha	173
7.2	Evaluating The Covariance Matrix	175
7.2.1	Robust Loss Functions for Volatility Estimation	175
7.2.2	Application to Multivariate Returns	176
7.3	Evaluating the Precision Matrix	179
7.3.1	Minimum-Variance Portfolios	179

7.3.2	Mahalanobis Distance	180
7.4	Ancillary Tests	181
7.4.1	Beta vs realized beta	181
7.4.2	Model Turnover	181
7.5	Further Reading	181
8	Fundamental Factor Models	183
8.1	The Inputs and the Process	183
8.1.1	The Inputs	184
8.1.2	The Process	186
8.2	Cross-Sectional Regression	188
8.2.1	Rank-Deficient Loadings Matrices	190
8.2.2	Conditions for Constrained Identification*	191
8.3	Estimating The Factor Covariance Matrix	192
8.3.1	Factor Covariance Shrinkage	193
8.3.2	Dynamic Conditional Correlation	194
8.3.3	Short-Term Factor Updating	195
8.3.4	Correcting for Autocorrelation in Factor Returns	197
8.4	Estimating the Idiosyncratic Covariance Matrix	197
8.4.1	Exponential Weighting	197
8.4.2	Visual Inspection	198
8.4.3	Short-Term Idio Update	198
8.4.4	Off-Diagonal Clustering	199
8.4.5	Shrinking of Variances	200
8.5	Winsorization of Returns	201
8.6	Selecting Factors: the Large Number of Predictor Case	204
8.7	Multi-Country Models	205
8.7.1	Model Linkage	205
8.7.2	*Currency Rebasing	205
8.8	A Tour of Factors	208
8.9	Further Reading	208
9	Statistical Factor Models	211
9.1	Statistical Models: The Basics	212
9.1.1	Best Low-Rank Approximation and PCA	212
9.1.2	Maximum Likelihood Estimation and PCA	215
9.1.3	Cross-Sectional and Time-Series Regressions via SVD	218
9.2	Beyond the Basics	218
9.2.1	The Spiked Covariance Model	219

9.2.2	Spectral Limit Behavior of the Spiked Covariance Model	221
9.2.3	Optimal Shrinkage of Eigenvalues	223
9.2.4	Eigenvalues: Experiments Vs. Theory	225
9.2.5	Choosing the Number of Factors	226
9.3	Real-Life Stylized Behavior of PCA	228
9.3.1	Concentration of Eigenvalues	228
9.3.2	Turnover of Eigenvectors	230
9.4	Interpreting Principal Components	235
9.4.1	The Clustering View	235
9.4.2	The Regression View	236
9.5	Statistical Model Estimation in Practice	237
9.5.1	Weighted and Two-Stage PCA	238
9.5.2	Implementing Statistical Models in Production	240
9.6	Further Reading	244
9.7	Exercises	245

II During The Trade 253

10	Transaction-Cost Aware Optimization	255
10.1	Market Impact	256
10.1.1	Temporary Market Impact	257
10.2	Multiperiod Optimization	262
10.3	Baldacci-Benveniste-Ritter	263
10.3.1	Comparison to Single-Period Optimization	265
10.3.2	The No-Market-Impact Limit	266
10.3.3	Optimal Liquidation	267
10.3.4	Deterministic Alpha	267
10.3.5	AR(1) Signal	268
10.3.6	Mixing Signals	269
10.3.7	Essential Statistics for AR(1) Processes	269
10.4	Further Reading	270
11	Hedging	271
11.1	Toy Story	271
11.2	Factor Hedging	273
11.2.1	The General Case	273
11.3	Hedging Tradable Factors with Time-Series Betas	276
11.4	Factor-Mimicking Portfolios of Time Series	279

11.5 ★Appendix	281
III After the Trade	283
12 Dynamic Risk Allocation	285
12.1 The Kelly Criterion	285
12.1.1 Kelly Portfolios: Mathematical properties	291
12.2 Log-Return Mean-Variance Optimization	294
12.3 Fractional Kelly and Drawdown Control	296
12.4 Variants of Fractional Kelly: Finite Horizon, Transaction Costs, and Heuristics	298
12.5 Further Reading*	300
13 Ex Post Performance Attribution	301
13.1 Performance Attribution: The Basics	302
13.2 Performance Attribution with Errors	303
13.2.1 Two Paradoxes	303
13.2.2 Estimating Attribution Errors	304
13.2.3 Paradox Resolution	305
13.3 Maximal Performance Attribution	307
13.4 Selection vs. Sizing Attribution	314
13.4.1 Connection to the Fundamental Law of Active Management	317
13.4.2 Long-Short Performance Attribution	317
13.5 Time-Series Performance Attribution	318
13.6 Appendix*	319
13.6.1 Proof of the Selection vs. Sizing Decomposition	319
13.7 Exercises	322
14 ★Appendix	323
14.1 Realized Variance of Minimum Variance Portfolios	323
14.2 Asymptotic Properties of Principal Component Analysis	325
14.3 The Linear-Quadratic Regulator	326
14.4 The Discounted Linear-Quadratic Regulator	326
14.5 Spiked Covariance Matrix: Basic Results	329
14.5.1 Some Useful Results from Linear Algebra	329
14.6 Optimal Trading: The Single-Signal Case	329
14.7 Conditioning	331
14.8 Three Papers on Backtesting Tests	333

14.8.1 White (2000)	333
14.8.2 Romano and Wolf (2005)	334
14.8.3 Hansen, Lunde and Nason (2011)	335
15 Bibliography	337
15.1 Bibliography	337
Index	359

Notation

\mathbb{R}^n	Field of real numbers
\mathbb{N}^n	Field of natural numbers
x, y, \dots	scalars
$\mathbf{x}, \mathbf{y}, \dots$	vectors
$\mathbf{X}, \mathbf{Y}, \dots$	matrices
\mathbf{x}', \mathbf{X}'	vector or matrix transpose
\mathbf{X}^+	Moore-Penrose pseudoinverse
Ω	covariance matrix
$\text{diag}(x_1, \dots, x_n)$	Diagonal matrix with scalars x_1, \dots, x_n on the main diagonal
$[\mathbf{x}]_i, x_i$	i -th element of a vector
$[\mathbf{X}]_{i,j}$	element of a matrix on i -th row and j -th column
$[\mathbf{X}]_{i,\cdot}$	i -th row of matrix X
$[\mathbf{X}]_{\cdot,j}$	j -th column of matrix X
$\text{trace}(\mathbf{X})$	Trace operator: $\text{trace}(\mathbf{X}) = \sum_i [\mathbf{X}]_{i,i}$
$\delta_{i,j}$	Kronecker's delta: $\delta_{i,j} = 1$ if $i = j$, 0 otherwise
$\delta(t)$	Dirac's delta function: $\delta(t) = 0$ for $t \neq 0$, and $\int_{-\infty}^{\infty} \delta(t) dt = 1$
$\mathbf{1}$	vector whose elements are all ones: $[\mathbf{1}]_i = 1$ for all i
$\mathbf{1}\{x\}$	indicator function, equal to 1 if x is true, 0 otherwise
\mathbf{q}	unit function in L^2 : $q(x) := 1$
\mathbf{e}_i	vector whose i -th element is 1 and the others are zero: $[\mathbf{e}_i]_j = \delta_{i,j}$
\mathbf{I}_m	identity matrix of size $m \times m$
$\ \mathbf{x}\ _n$	n -norm, for $n \geq 1$: $\sum_i x_i ^n$
$\ \mathbf{x}\ _{\mathbf{Q}}$	$\sqrt{\mathbf{x}' \mathbf{Q} \mathbf{x}}$, for a symmetric positive definite matrix \mathbf{Q} .
$\ \mathbf{H}\ _2$	operator norm for a symmetric positive definite matrix \mathbf{H} : $\max_{\mathbf{x}} \ \mathbf{H}\mathbf{x}\ / \ \mathbf{x}\ $.
$\lfloor x \rfloor$	largest integer that is less than, or equal to, x

$\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{R}$	scalar product of two vectors, i.e. $\mathbf{x}'\mathbf{y}$
$\langle \mathbf{A}, \mathbf{B} \rangle \in \mathbb{R}$	scalar product of two matrices, i.e. trace $(\mathbf{A}'\mathbf{B})$
$\mathbf{x} \circ \mathbf{y} \in \mathbb{R}^n$	Hadamard product of two vectors and matrices: $[\mathbf{x} \circ \mathbf{y}]_i := x_i y_i$
$\mathbf{x} \perp \mathbf{y}$	random variables \mathbf{x} and \mathbf{y} are independent
$E(x), E(\mathbf{x})$	expectation of random variables or random vectors
$E_\xi(f(\xi))$	expectation of a function f of random variable ξ
$\hat{E}(\mathbf{x})$	average of a vector \mathbf{x} : $\hat{E}(\mathbf{x}) := n^{-1} \sum_i x_i$
$\text{var}(x)$	variance of arandom variable x
$\mathbf{L}_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$	shift operator: $L_1 \mathbf{x} := (x_2, x_3, \dots, x_n, 0)'$, i.e., $[\mathbf{L}_1]_{i,j} := \delta_{i,j-1}$
rv	random variables
iid rv	independent identically distributed random variables
pdf	positive definite
$\stackrel{d}{=}$	equality in distribution of two random variables
\sim	rv distributed according to a distribution, e.g. $\xi \sim N(0, 1)$
\Rightarrow	convergence in distribution

Part I

Before the Trade

Chapter 1

The Map and the Territory

The Questions

1. What are the instruments we are trading?
2. Who are the players?
3. What are the sources of excess returns?

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

This chapter is a guide to the essential components of quantitative investing. When considering the meaning of a word, it’s often instructive to go back to its etymology, so let’s play this game. Despite being a Germanic language, English adopted many words from Latin, sometimes by way of French. “Investing” comes from “Investire”, which in Latin meant “to cover with a vest”, or “to put in a vest”. So it should be hardly surprising that two thousand years later, vests would become the favorite garment of hedge fund managers. In the Middle Ages, the verb took on the additional meaning “to surround, to have ownership of”. It is also possible that the modern meaning overtook the old because, in ceremonies in which ownership was transferred, the new owner was “invested” with a cloak and other regalia. In Italian—the direct successor to Latin—the old meaning is gone, and “investire” only means “to receive possession of something”. As for “quantitative”, that is Latin too: “quantum”, a noun denoting something that

can be measured, increased, and decreased. We will deal with ownership, sold and bought in units that can be measured, increased and decreased. This is, unfortunately, the whole of finance. You can own a house, a painting, a bet on the survival of humankind, or even an idea. Each one of these investment topics deserves its own book, written by a competent author. In writing this book, I have chosen to trade off generality in favor of detail. I have covered each subject with the goal in mind that you would have sufficient information to understand it, implement it, and critique it. However, even an analytical book needs an introduction that puts things in their proper context. In this chapter, I aim to provide that context. You will have a broad understanding of the classes of securities to which these methods apply; and of the way these securities are traded, and by whom. This is a necessary prerequisite to explore fundamental questions: where are excess returns coming from? What causes these trading opportunities? Finally, I will present the essential components that make up the analytical framework of a quantitative portfolio manager. The underlying message is that to be successful, an investor must understand how things work. A seminal early book on investing is titled “The Intelligent Investor”. To double down on Latin, the original meaning of “intelligent” is “to read into something”, similarly to “insightful” in the English language. Your success will come from reasoning about the behavior of your counterparties, the rules governing the trading of your assets, and the functioning of exchanges. Many budding quants focus on quantitative methods. The fact is that theory is cheap and is often not hard. What is hard is putting the right tool at the service of the right insight.

Finally, this is the only chapter without mathematics. You should enjoy while it lasts.

1.1 The Securities

We will be concerned with *standardized products* that are *liquid*. We explain these concepts in more detail.

To “own” an object is effectively to own *claims* on that object in the future. If you own a house, you can live in it or rent it out (your claim) and it is yours. This claim is not absolute, however. In most countries, the local or central government may need your property for reasons of public welfare and can require you to exchange your claim for cash at a fair price. If you own a painting, you may enjoy it in the confines of your house, but may not necessarily own its reproduction rights. If you own a bet on the future of humanity, your counterparty may have some *force majeure* clauses that prevent it from paying

(e.g., consider a zombie apocalypse scenario). Defining ownership of an “idea” is especially challenging and prone to be treated on *ad hoc* basis. Compared to the infinite and ever-changing nature of the meaning of property rights, our coverage is very narrow. Specifically, we focus on the subset of contracts that are standardized and liquid. We buy and sell *standardized* claims. These claims come in a few varieties, and their attributes are clearly defined and known to all potential buyers and sellers. Examples are:

- *Equities and ETFs.* These give us partial ownership in companies, or groups of companies, and entitle us to receive future cash payments generated by the economic activities of these companies.
- *Futures.* These contracts deliver a physical commodity or a cash payment contingent on the state of the world at a future date, at a price determined today.
- *Bonds.* These are contracts that allow the transfer of debt claims among parties. An investor lends money to a borrower, in exchange for a fixed cash flow in the future (for example, periodic interest payments and a final payment). A bond makes this claim transferable to other lenders.
- *Vanilla options.* These are claims that depend on the future value of some underlying asset; for example, you may receive the right (but not the obligation) to buy a stock at a future, at a price determined today. The nature of these claims is standardized, hence the term “vanilla”.
- *Interest Rate Swaps (IRS).* These contracts allow the exchange of a certain, deterministic cash flow stream for an uncertain one, which depends on interest rates at future dates.
- *Credit Default Swaps (CDS).* These contracts insure the buyer against the failure of a company at a future date, in exchange for recurring fixed payments.

Secondly, these contracts are *liquid*. For our purposes, a liquid contract is one that can be bought and sold at large enough sizes, and at sufficiently short time horizons, to enable quantitative strategies to be implemented. This means that if we plan to buy or sell a contract, we should be able to do so without incurring a transaction cost so high that our strategy is not economically attractive even for small trading sizes, and that the waiting time due to searching a counterparty should not be so long as to make the transaction economically unattractive.

The properties of standardization and liquidity are closely intertwined. Increased standardization tends to enhance liquidity by consolidating demand, as it aggregates dispersed demand from bespoke products towards a smaller set of standardized ones. Furthermore, standardization streamlines the trading process, reducing transaction costs and bolstering robustness, thereby fostering investor trust and attracting greater participation, thus enhancing liquidity. However, the downside is that customers may sacrifice the ability to trade certain useful product characteristics. Determining the optimal level of customization, even at the expense of liquidity, remains an ongoing process of learning and adaptation. For instance, prior to the 2008 financial crisis, Credit Default Swaps (CDSs) exhibited greater variety. However, the ‘Big Bang’ initiated by the International Swaps and Derivatives Association (ISDA) on April 8, 2009, simplified contract terms, including standardizing coupon rates (100bps and 500bps) and introducing a standard upfront payment, which played a pivotal role in restoring confidence in this asset class ([Vause, 2010](#)).

Trading and liquidity are at the core of the book. In order to better understand the trading process and the nature of liquidity, we should describe in some detail how trading on exchange and over-the-counter happens.

1.2 Modes of Exchange

At any given time, economic agents want to buy or sell contracts. They want to do so quickly, securely, and cheaply. The three options around which trading is currently organized are exchanges, over-the-counter, and dark pools. Exchanges are venues in which the orders of buyers and sellers are anonymized and matched against each other. Orders are characterized by size, the number of contracts, direction (buy or sell), and price. They represent requests to buy or sell a number of contracts. The exchange records such active orders on a ledger, known as the limit-order book (LOB), and employs a set of priority rules to match buy and sell orders in the exchanges aptly named matching engine. In order to trade on an exchange, one must be a member of that exchange. Membership entails apparent benefits and less-apparent responsibilities. Market participants must maintain sound governance, risk processes, and capital structure.

Exchanges evolve continuously due to two driving forces. On one side, there’s a push toward consolidation, which reduces operating costs and gives the owner pricing power. On the other side, technical and process innovations introduce new competitors into the market. In the US alone, there are more than a dozen equity stock exchanges. Exchange-traded assets, such as stocks, options, and

futures (including Forex futures), are often liquid, although this condition is neither necessary nor sufficient: some exchange-traded assets are traded in minimal volumes and, therefore, are not liquid, and some very liquid products are traded off-exchange.

Other assets are not traded on exchanges, but *over-the-counter (OTC)*. In this case, the buyer or seller transacts through an institutional market participant, the broker-dealer, which is connected to other broker-dealers and facilitates the matching of orders. Bonds, Interest Rate Swaps, Forex currencies, Forex Futures, CDS are examples of contracts traded OTC. Some of these, like currencies, are among the world's most liquid contracts. A precondition for liquidity is standardization. Think of a house. "The New York housing market" is very different from the stock market in that each of the 16 billion outstanding Apple shares (as of June 2022) is indistinguishable from the other and sells in a matter of seconds. In contrast, a house has many attributes that make it unique: location, size, age, blueprint, and condition. Another characteristic of liquid markets is the large number of participants. When numerous participants are involved in a market, competing for a relatively low number of contracts, transactions become more frequent, the necessity for bilateral bargaining diminishes. The ability of any individual participant to influence the price significantly is reduced. To illustrate, consider the housing market as a counterpoint: when selling a house, you typically negotiate with one specific buyer (out of a few eligible ones), who may spend many hours searching for the right property and may engage in intense bargaining, sometimes to the point of contention, to secure the best possible price.

Finally, *Dark Pools* (a type of ATS, or Alternative Trading System, that does not make its limit order book transparent) are additional venues that are distinct from exchanges (although sometimes owned by them). Dark Pools address the needs of certain institutional investors to execute orders without displaying their trading intentions. By design, dark pools hide order details and only make trades details available after execution. As of 2019, approximately xx of flow is traded on Dark Pools.

1.3 Who Are the Market Participants?

It is convention to partition traders into the *Sell Side* and *Buy Side*. The former facilitates trading by providing services; the latter receives trade for their own benefit. Below I describe the participant types. For a more detailed description, see [Harris \(2003\)](#).

1.3.1 The Sell Side

The sell side comprises brokers, dealers and broker-dealers.

- *Dealers*¹ fulfill their clients' demand, thus providing liquidity. They take the opposite side of the trade; they are profitable if, on average, they sell (buy) at a price higher (lower) than what they initially paid to buy (sell) the asset. The difference between buy and sell prices is the *spread*. When dealers interact with clients, they quote the buy price (the *bid price*) and the sell price (or *ask price*) for a contract. They are effectively *making a market*, since these quotes make transactions possible. In OTC markets, dealers are the primary liquidity providers. The most sophisticated among such markets allow the dealers to quote prices, quantities, and other attributes continuously. For example, fixed-income products can be traded on Dealerweb or Bloomberg. In order markets or for highly bespoke products, the dealers quote on request, possibly one-sided only, for a specific quantity and with an expiration time. The quote, or the spread if the quote is two-sided, depends on the quantity. Similarly to speculators, dealers trade on their own behalf. Like speculators, they hold a portfolio (or an *inventory* of positions) and face the issue of facing trading counterparties that may be more informed than themselves. Unlike speculators, dealers are passive traders, in that they respond to their clients. Also, unlike speculators, dealers enjoy special regulatory status. Because dealers observe the demand flow of their clients, they are informed agents, often serving the needs of informed clients. The dealers' profit originates from the realized spread of their trade (which is usually lower than the quoted spread) but also from the specific information the dealer derives from the order flow. One specific type of flow originates from retail investors (who we introduce later in the chapter). These investors access the market indirectly through brokers. Brokers have special arrangements to direct market orders to dealers, who commit to execute them while offering certain price guarantees on the trade.

In summary, dealers are liquidity providers, and they are compensated for services through trading profits.

- *Brokers*² trade on behalf of their clients. When the broker receives an

¹Ch. 13 in [Harris \(2003\)](#).

²Ch.7 in [Harris \(2003\)](#).

order from a client, together with information about the client's time and price preferences, it searches for the most effective channel to execute it in accordance with these preferences. For example, a client sends a broker an order to buy a certain number of shares of a company. The broker is a member of all major exchanges. It splits the large order into smaller orders and routes them to the various markets at times that meet the execution horizon of the client or its expected cost. Unlike dealers, brokers are intermediaries who take no risk by holding contract positions at any given time. The intermediation service they provide is beneficial, however, and it comes with its own risks. First, brokers provide exchange access to non-member clients and they provide OTC dealer access to non-institutional clients. Institutional clients, too, may want to enlist a broker when interacting with dealers, since the broker anonymizes the clients. Secondly, broker intermediation solves bilateral settlement risk: money is exchanged for contracts after the trade occurs. Clients need to know and trust their counterparty to protect themselves from insolvency, renegeing, or non-compliance. There is a small number of brokers compared to the number of traders, so that clients need to approve (and be approved by) only by few counterparties; a reduction in time, cost, and risk. This, of course, does not eliminate counterparty risk. It transfers it to the brokers. The brokers manage it by vetting the clients, and by requiring that clients deposit capital at the broker, which the broker uses in case of client insolvency. The brokers also *clear* and *settle* trades on behalf of the client. In addition to these services, brokers, and especially *prime brokers*, the subclass of brokers servicing hedge funds and other sophisticated investors, offer their clients other services:

- *Custodial services.* Brokers ensure receipt, recording, and safekeeping of securities.
- *Rehypothecation.* Clients may allow the brokers to use their securities for the brokers' own needs in exchange for fees or rebates. For example, brokers may use client securities as collateral for their own transaction or lend them to other clients.
- *Margin loans.* Brokers lend clients short-term capital to buy securities. They charge them SOFR³ plus a spread.

³Secured Overnight Financing Rate (SOFR) is a measure of the interest rate for overnight cash loans.

- *Location of short positions.* Clients may want to *short* stocks, i.e., sell shares first, and buy them back at a later time, with the expectation that future prices will be lower than current ones. Brokers enable these transactions by lending shares from a third party and making them available to the clients. The clients then sell them in the open market, buy them back at a later time and return them to the broker. After the initial sale, but before the buyback, the broker invests the cash proceeds at a rate SOFR plus a spread. The client receives from the broker SOFR minus a spread.
- *Research reports and services*, as well as broker specific data. These services used to be bundled in broker commissions but after the implementation of MIFID II regulation, they are now charged separately.
- *Capital introductions*, in which brokers facilitate the connection between hedge funds and potential investors.

In summary, brokers offer diversified services, the most important of which is to facilitate clients' transactions. They are compensated by commissions, interest on cash balances, interest on lending, and payment for order flow (PFOF), which is the compensation brokers receive from market makers for routing orders to them.

Broker-Dealers, also called *dual traders*, combine the previous two functions in a single entity. They act both on behalf of the client and on their own behalf. This introduces a tension. The dealer's arm is incentivized to use the broker's information in trading to its advantage. Maybe the simplest action is *front-running*: the dealer is aware of incoming buying or selling demand for a security, and buys it in advance before this demand manifests in the market and is reflected in prices. To mitigate this type of behavior, regulations are in place to safeguard the interest of the client. The most important law regulating brokers, dealers, and broker-dealers is the Securities Exchange Act of 1934 (or '1934 Act').

1.3.2 The Buy Side

The buy side usually trades with the sell side. You (the reader of this book) are likely to be a member of this group, even though certain dealers face quantitative challenges similar to yours. It is important to understand who the actors in the buy-side drama (occasionally, tragedy) are, because you will continuously interact with them, and your excess returns will be the outcome

of this interaction. We could classify the buy-side actors according to several criteria. For example, the sub-industry to which they belong: life insurers, mutual funds, hedge funds, and so on. I opt to classify them (subjectively!) based on the type of investing they perform.

- *Indexers* are passive investors. Their portfolios replicate the compositions of the benchmarks, or indices, generated by data providers like MSCI, S&P, Russell, CRSP, or from exchanges like FTSE 100, TOPIX, and Deutsche Börse. These indices are updated on a quarterly or biannual basis, and they comprise bond indices as well, like the Bloomberg Agg (until 2016 owned by Barclays). Several investment vehicles track indices; mutual funds and exchange-traded funds are the largest in terms of size. Large firms in this group are Blackrock, Vanguard, and State Street. Indexers make up a large and growing share of the total asset base. According to estimates by [Chinco and Sammon \(2023\)](#), they represent over 37% of the US stock market capitalization as of 2020.
- *Hedgers* are firms participating in markets with the primary objective of reducing financial risk originating from their core businesses. For example, currency risk is faced by any firm doing business internationally. Firms such as airlines and manufacturing companies purchase fossil fuels (gas, Brent, and West Texas Intermediate), whose price variability can be very disruptive. Hedgers primarily participate in derivative markets: futures, swaps, and options. Hedgers differ from other participants that also hedge, such as dealers or hedge funds, in that hedging is the primary activity they perform.
- *Institutional Active Managers* are firms investing on behalf of their clients. They run strategies that are sometimes benchmarked to commercial indices and hope to beat them. There is some evidence of underperformance of funds serving retail investors; see S&P SPIVA report⁴ or the Refinitiv study⁵. Both show that over 60% of funds underperform their benchmarks over a one-year trailing basis. The outperformance of funds over one year is not persistent: as of January 2024, 91% of funds trail the performance of the S&P500 over the previous 15 years. On the other side, funds serving institutions seem to beat their benchmarks ([Gerakos and Linnainmaa, 2021](#)). The *tracking error* is a measure of the risk they can take when

⁴<https://www.spglobal.com/spdji/en/research-insights/spiva/>.

⁵<https://lipperalpha.refinitiv.com/reports/2024/01/monday-morning-memo-performance-review-relative-performance-equity-funds-2023>.

differing from their reference benchmarks. Otherwise stated, their portfolios can be expressed as the sum of the positions in a benchmark, and of discretionary positions of a “tracking portfolio”. A large tracking error gives the funds much discretion; a low one makes them close to the indices, and makes them “index huggers,” i.e., index funds in disguise.

- *Asset Allocators* manage portfolios composed of securities in multiple asset classes. Within an asset class, the portfolio closely follows a representative benchmark. One can view asset allocators as managers of a portfolio of asset classes. The relative weight of these asset classes in the portfolio is either constant or changes slowly. Common asset classes are equities, bonds, commodities, and cash equivalents⁶. In addition, asset allocators invest in alternative asset managers like private equity firms, venture capital, hedge funds, and real estate.
- *Informed Traders* include primarily hedge funds and principal trading firms. These firms are usually organized as partnerships, although a few are public companies. They face fewer constraints than institutional managers. Whereas principal trading firms only have general partners (GPs, the principals) investing their own money, hedge funds also have limited partners (LPs) who do not invest actively⁷. These firms pursue absolute returns (i.e., not tracking a benchmark), which exhibit low correlation to the indices of major asset classes⁸. Informed traders invest heavily in human capital, technology, and data to achieve this goal. They fulfill two major functions. The first one is *price discovery*. By using all information available to them, they generate estimates of the true value of securities. If the security prices differ from their estimates, they trade to exploit the mispricing. If the price is lower than their estimate, they buy the security. In the process, they increase its price and bring it closer to equilibrium. Mispricing can take many forms. If the same security is offered at different prices on different exchanges, *arbitrageurs* (a subset of informed traders) will try to exploit the difference; of course, this may not be easy to do, so the difference either persists, or disappears very

⁶Cash Equivalents are highly liquid assets with low returns, like bank certificates of deposit, short-dated treasuries, or commercial paper.

⁷GPs in hedge funds are both principals (since they invest their own money) and agents (on behalf of the LPs). To resolve this conflict of interest, the SEC regulates the class of investors who can be LPs (high net-worth, sophisticated investors), and gives LPs special privileges within the fund.

⁸This is not always true in practice because a) some hedge funds have an explicit market exposure; b) some hedge funds have an asymmetric exposure to the market ([Agarwal and Naik, 2004](#)).

quickly due to technology investment in low-latency trading. The second role of informed traders is liquidity provisioning. Supply and demand of certain assets is predictable to a certain degree. I provide examples in Section 1.4; hedge funds and market makers have developed specialized strategies that predict imbalances, hold (or short) securities before the liquidity need materializes, and meet the liquidity needs at the event. The range of possible intervals between prediction and event can be vast – from sub-second for high-frequency market makers to weeks or months for hedge funds.

- *Retail Investors* trade for their own account via retail brokers. In 2020, retail investors made approximately 20% of total volume; the share was slightly more than 10% in 2011⁹. Several studies, across different national markets and periods, have shown that retail traders are consistently unprofitable ([Barber and Odean, 2013](#)); retail trader flow is uninformed. This is one of the reasons why it is highly sought after by dealers, who will pay the retail brokers for routing it to them (payment for order flow).

1.4 Where Do Excess Return Come From?

Now that we have introduced the main actors in the play (usually tragedy, rarely comedy, and occasionally farce) of investing, we can discuss the sources of excess returns. The “excess” qualifier means “in excess of portfolio invested in risk-free assets, such as short-dated US Treasurys.” This topic is central both to academic financial research and to practitioners. Academic finance is primarily concerned with the question of *efficiency*. In the words of [Malkiel \(1987\)](#):

A capital market is said to be efficient if it fully and correctly reveals all available information in determining security prices. Formally, the market is said to be efficient with respect to some information set, ϕ , if security prices would be unaffected by revealing that information to all participants. Moreover, efficiency with respect to an information set, ϕ , implies that it is impossible to make economic profits by trading on the basis of ϕ .

An exceptionally concise definition, if there ever was one. At its core is the “information set ϕ ”. This could be, for example, the set of all historical prices

⁹[BNY Mellon Insights: The Rise of Retail Traders](#)

of the traded securities. Nowadays, this information can be obtained with *relatively*¹⁰ little effort. A different type of information set is publicly available information¹¹, defined in the US as “any information that you reasonably believe is lawfully made available to the general public from: (i) Federal, state or local government records; (ii) Widely distributed media; or (iii) Disclosures to the general public that are required to be made by federal, state or local law.” An even finer information set is the set of *all* information available to *any* investor. Academic research tries to determine the validity of the statement that “security prices would be unaffected by revealing that information to all participants.” Note that this does not mean that ϕ is not helpful to predict future prices. Indeed, there is empirical evidence that asset prices are predictable. However, the hypothesis is that current prices may not be affected. We do not trade in the direction of returns, up to the point that the investing opportunity disappears. This is unintuitive. Why would we not take advantage of an informative prediction? One reason is *risk*. Even if we have some information about the future return of an asset, the uncertainty around the prediction is too high for us to take advantage of it. For example, say that, to the best of our knowledge of ϕ , we expect the US market to appreciate 8% next year, while our cash custodied at the broker will return a measly 2%. Does this imply that we will rebalance our portfolio to 100% a market-tracking asset like SPY? Hardly. The reason is that the standard deviation of market returns is 20%, a little too high for comfort. Risk, however is not the only reason. Another one is *liquidity*. Indeed, the road to hell of an investor is littered with quite accurate predictions of assets that barely trade or do not trade at all. A famous example is the spin-off of Palm (a now-defunct mobile device company) by 3Com (a telecom equipment maker, also defunct) in 2000. 3Com floated on the public market 5% of the shares of Palm, while retaining the other 95%. Right after the IPO, Palm had a market value of \$54B, while 3Com had a market capitalization of \$28B. The implied value of 3Com assets was \$-22B, even though the company had no debt, \$1B in cash, and positive cash flow. Either 3Com was dramatically undervalued, or Palm was dramatically overvalued. An investor could have therefore bought 3Com shares and shorted Palm for an equal amount. The portfolio comprised of these two assets was a synthetic asset whose return could be predicted. There was a problem, however. Palm shares were in short (pun intended) supply. In order to short a share, the investor must first borrow it, at a rate decided by

¹⁰The “relatively” here denotes a bit of sarcasm. Collecting long time series historical prices of good quality, accounting for corporate actions, is hard, expensive and requires skill.

¹¹PRIVACY OF CONSUMER FINANCIAL INFORMATION UNDER TITLE V OF THE GRAMM-LEACH-BLILEY ACT. §163.3 (w) (1).

the lender. If quoted at all, these rates were so high as to make the trade either unattractive or impossible. Risk and Liquidity are not the only two factors limiting the exploitation of information. We list two more. The first one is *funding*. Consider a scenario in which the certain assets, or certain portfolios¹², have lost much of their value due to market distress. We are managing a small hedge fund, which has also lost money in this environment. Based on historical examples, we have a strong belief that such assets will rebound. Such scenarios occur quite regularly, especially in “deleveraging spirals”. However, we do not have much capital available to post as margin. In addition, we need a capital buffer in order to withstand a possible additional loss in the very short term. Funding constraints prevent us from buying the asset, in spite of our accurate forecast.

A significant source of excess returns arises from *flow predictability*. Some agents, notably institutional investors and market makers, but not only, will trade known securities on known dates. Speculators can then take advantage of this information by providing liquidity beforehand¹³. One of the most important instances of this is *index rebalancing*. Several index providers update the weights of their indices on predetermined dates using well-defined rules. Some securities are added to the index, others are removed, and finally most of the remaining ones have an updated weight. The term used for this process is *index reconstitution*. For example, TSLA was added to the NASDAQ 100, effective July 15, 2013. The announcement was made on July 10, 2013; but several investors could have forecasted the event well before that date. These investors would then purchase TSLA shares and sell them at the closing auction of July 15, 2013. The ETF, mutual funds and bespoke products that track the index have an obligation to buy TSLA on the close of that day, and the resulting demand is likely to push up the stock price. The informed investors providing liquidity do not do so risk-free. They hold the stocks until the effective date, and over these days are exposed to the risk that TSLA may suffer from company-specific or industry-specific losses. Moreover, there is the remote risk that the rebalance be cancelled or postponed. The size of passive investing is large and its estimate ranges from 17.5% (Novick, 2017) to 38% (Chinco and Sammon, 2023) of total assets under management. The buyer of index products bears the indirect cost of such rebalancing (Li, 2021). This is just one prominent example of predictable flows, but several others exist, usually smaller in size, but also not as widely known as index

¹²As in the case of the pair 3Com-Palm, we can interpret portfolios as synthetic assets.

¹³Note that this is not the same as *front running*. The latter also consists of buying or selling a security based on a demand forecast of said security; but in the latter case, the forecast relies on *non-public* information.

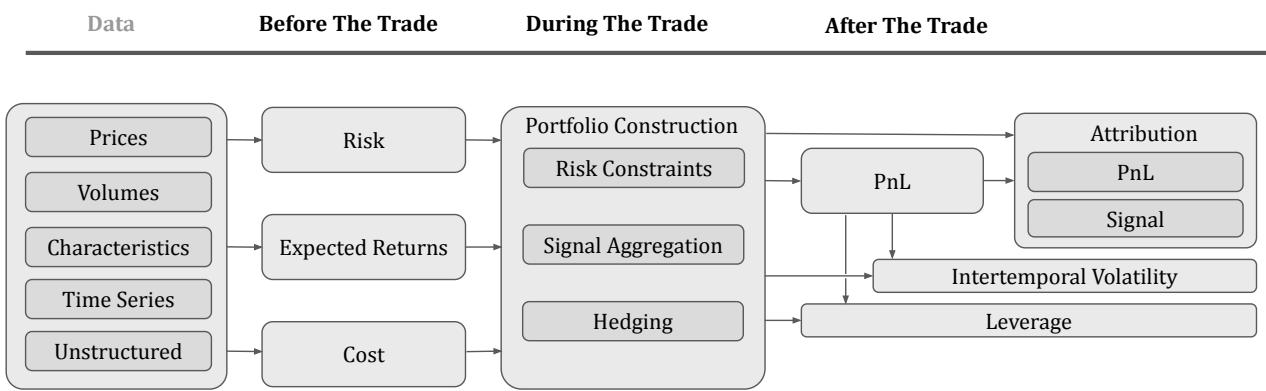


Figure 1.1: The components of the investment process.

rebalancing. Their common feature is the existence of institutional or procedural constraints (sometimes driven by internal processes, other times by regulatory requirements) that introduce predictability in the demand of securities.

Finally, we consider a last source of excess returns: *informational advantage*. This means that the investors do not only differ by their risk attitude, their tolerance to illiquidity, their funding level, but also by their information sets. This is what is often meant by “statistical arbitrage,” the ability to predict returns accurately based on insights our competitors do not have.

In summary, even assuming accurate information owned by participants, some of them cannot exploit this information. We have listed several possible causes of return predictability:

- pure arbitrage;
- heterogeneous risk preferences, liquidity, funding, flow predictability;
- informational advantages.

These categories are not exclusive. For example, flow predictability and liquidity are related, albeit not identical; and the distinction between being compensated for risk-taking and holding an actual information advantage is usually unclear. However, these broad classes can help us reasoning about one strategy’s edge.

1.5 The Elements of Quantitative Investing

The investment processes is usually viewed as a highly structured process. There are separate components, the development of which is the responsibility of

separate teams. In Figure 1.1 I show a possible organization of these components, which I follow in the organization of the book. I review them below.

- *Data*. The essential inputs to investing are under the “Data” section to the extreme left.
 - *Prices and trading volumes* are often collected at regular intervals, e.g., minutely, every 5, 10, 15 minutes, or daily. For high-frequency strategies it may be necessary to use order-level exchange data.
 - *Characteristics* are numerical vectors associated with a security and a time stamp. Consider them as descriptors of the security. For a stock, a characteristic may be a measure of “quality”, like the free cash flow generated by the firm in the most recent quarter, divided by the market cap of the firm. Another widely used characteristic is the realized return of the stock over a certain interval, for example the past six months.
 - *Time Series* differ from characteristics (which are multivariate time series) in that they are not associated to individual securities, but rather provide additional information entering the investment process. Examples of time series are the Consumer Price Index (CPI), which can be used to estimate inflation rate in the U.S.; the yield of the 10-year Treasury bond; the Federal Funds Effective Rate, the overnight rate for unsecured lending of reserves among commercial banks; and the VIX, a forward-looking measure of US equities market volatility.
 - *Time Series of Unstructured Data* are the dark matter of financial data. Prices, characteristics and times series are structured data i.e. numerical, categorical or ordinal (i.e., rankings), and in tabular form. Unstructured data are usually character sequences representing natural language (examples are earnings transcripts and firm news), or images (for example, satellite images), video/audio files, or multimodal data, i.e., a combination of all of the above formats.
- *Before the Trade*. Data are used to develop three components that enter the portfolio construction during the trade. Because they precede the trading process, I classify their development as being “before the trade”.
 - *Risk*. The word “risk” can mean many things. In the context of this book, we will use portfolio volatility as a proxy for risk. Estimating this volatility for an arbitrary portfolio is a challenging task.

- *Expected Returns.* In order to be profitable, a trading strategy needs to have informative predictions of future returns. This is often viewed as the paramount concern of a quantitative investor, and to a large extent it is. In nearly all modern trading systems there is more than one estimate of expected returns. The number of estimates can run in the thousands or millions.
- *Transaction Costs and Market Impact.* Trading securities is expensive, and these costs are unavoidable when deploying a strategy. Among other things, transaction costs determine whether a predicted return at some horizon can be turned into a profitable strategy or not; and what is the maximum profit that can be extracted from such a strategy.
- *During the Trade.* The three components developed in the previous stage are combined in the portfolio construction phase. This happens “during the trade”, because the portfolio construction procedure results in real-time trading decisions, and these decisions determine the Profits and Loss of the strategy. The decisions taken in the portfolio construction process are:
 - *Incorporation of Risk Constraints.* A strategy’s PnL is a function of the maximum risk that it can take. This is usually represented in the portfolio construction problem either in the form of constraints, or in the form of penalties added to the objective function. Risk constraints and penalties can have a very material impact on the performance of the strategy.
 - *Signal Aggregation.* We use the term *signal* for a model of expected returns. As I mentioned above, there can be many signals for the returns of a single asset. A problem encountered in practice is combining such signals into a single signal.
 - *Hedging Decisions.* Certain trading strategies have *exposure* to systematic risk. In layman’s terms: they can lose money because their returns are correlated to market-wide sources of risk. Some of these sources can be *hedged*, which means that such risk can be counterbalanced. This is an important concept in portfolio construction.
- *After the Trade.* The trading process generates a time series of Profit and Loss (PnL), both for the overall portfolio, and for its constituents.

- *Performance Attribution.* In the portfolio construction phase we estimate expected returns of individual signals. This is an *ex ante* exercise. *Ex post*, we observe the actual PnL of the portfolio. Performance attribution is the practice to tracing back performance to its possible sources, to see what worked and what did not, so that we can learn from the experience. Moreover, performance attribution can also be employed to assess whether we have skill in sizing our bets.
- *Intertemporal Volatility Allocation and Leverage.* How should we allocate risk across periods? This decision is very consequential to capital growth. A closely related question is that of leverage, defined as the ratio between Gross Market Value (GMV) of the portfolio and Assets under Management (AUM).

I stress that different arrangements are possible; for example, [Narang \(1990\)](#) employs a simpler scheme, and [Pedersen \(2015\)](#) has yet another one. Some of the individual elements I introduce are present in these models of quantitative investing.

The Takeaways

1. Market participants can be broadly classified into sell-side and buy-side participants. The sell-side participants are:

- Dealers;
- Brokers;
- Broker-Dealers.

The buy-side participants are:

- Indexers;
- Hedgers;
- Institutional Active Managers;
- Asset Allocators;
- Informed Traders (e.g., hedge funds, principal trading firms);
- Retail Investors.

2. Excess returns arise from five major sources:

- a) Risk;
- b) Liquidity;
- c) Funding constraints;
- d) Predictable flows;
- e) Informational advantage in predicting future returns.

3. Quantitative investing employs elementary analytical models that can be partitioned in three broad groups:

- a) Risk measurement models and data;
- b) Market Impact models;
- c) Models of expected returns of assets.

Chapter 2

Returns: Properties and Models

The Questions

1. How do we define returns for equities, bonds, credit instruments, futures?
2. What are the stylized properties of returns?
3. Why is volatility an important measure for risk and portfolio construction?
4. What is GARCH? How do we use it?
5. How do we use Kalman Filtering to estimate volatility?
6. How do we model multivariate returns?

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

We start with models of univariate returns for two reasons. First, single-asset returns are the basic constituents of portfolios. We cannot hope to understand portfolio behavior without a solid understanding of their building blocks. Therefore, it is necessary to summarize the salient empirical properties of stock returns and the most common processes employed to model them, specifically to model volatility effectively. These models have general applicability

and are even more useful when combined with other families of models for multivariate returns. GARCH and exponential moving averages are essential tools for the working modeler. In the process, I introduce models that justify their use. Exponential moving averages find their motivation in linear state-space models, while GARCH is an instance of a nonlinear state-space model. These models will be your friends for life. The chapter has five parts. First, we lay out definitions of returns. Second, we summarize some “stylized facts” (empirical features of returns that are ubiquitous and relevant to risk management). Third, we skim GARCH models and realized volatility models. Because both topics have been covered extensively in textbooks, my goal is to introduce the essentials, their associated insights, and provide a jump-off point for the reader. Lastly, I touch on the State-Space Model for Variance Estimation.

2.1 Returns

2.1.1 Definitions

We have a set of n assets and a currency, also called the numeraire¹. We will use dollars throughout as currency. It is customary to assume that each of these assets is infinitely divisible. We buy the equivalent of a unit of currency for asset i . We denote the value of the asset tomorrow R_i . An equivalent way to define returns is from the closing price of security i on days 0 and 1, $P_i(0)$ and $P_i(1)$, respectively. The *return*² is defined as

$$r_i(1) := \frac{P_i(1) - P_i(0)}{P_i(0)}$$

We extend this definition to the case in which the security pays a dividend. The holder of the asset receives an amount $D_i(1)$, and the return is then defined as $P_i(0)$ and $P_i(1)$ respectively. The *dividend-adjusted return* is defined as

$$r_i(1) := \frac{P_i(1) + D_i(1) - P_i(0)}{P_i(0)}$$

We denote the vector of daily returns at time t as $(R_1(t), \dots, R_n(t))$. A great deal of equity risk management deals with the properties of this vector. For a

¹This word comes to English from the Latin *numerarius*, or ”a number”, ”a unit”, through the French *numéraire*.

²Definitions of returns, log returns, dividend-adjusted returns are in [Ruppert and Matteson \(2015\)](#) and [Connor et al. \(2010\)](#).

portfolio $\mathbf{w} \in \mathbb{R}^n$, where w_i is a monetary amount invested in asset i , the Profit and Loss (PnL) in a single period is given by the change in the value of the portfolio. The number of shares owned in asset i is given by $w_i/P_i(0)$. The value of the portfolio in period one is $\sum_i (w_i/P_i(0))P_i(1)$, and the change in value is $\sum_i (w_i/P_i(0))P_i(1) - \sum_i w_i P_i(0)$. In vector form, this equals $\mathbf{w}'\mathbf{r}$.

2.1.2 Excess Returns

In the rest of the book, we will not use security returns, but returns minus the *risk-free rate*. If, for example, we model daily returns, the risk-free rate r_f is the interest rate paid by the investor for borrowing cash over the same period, or paid to the investor for cash held in their account³. If we hold a security, we pay interest on the cash position of that security. If we are short, we receive interest. Cash is to all effects a security, but a special one, in the sense that it has much lower volatility (for modeling purposes, negligible volatility) than the other risky assets. We borrow or lend an amount equal to the Net Market Value (NMV) of our portfolio, i.e., the sum of the values of each position. The return of a portfolio is

$$\sum_i w_i r_i - (\sum_i w_i) r_f = \sum_i w_i (r_i - r_f)$$

The formula allows us to eliminate the risk-free asset from the portfolio and provides a natural interpretation of security returns as returns in excess of a rate received in the absence of investing. In the U.S.A., the reference rate is a reference overnight lending rate, like the Secured Overnight Financing Rate (SOFR)⁴.

2.1.3 Log Returns

If \mathbf{r} follows a multivariate Gaussian distribution, then so does the portfolio. The variance of this portfolio can be computed by using just two pieces of information: the portfolio weights, and the covariance matrix of the returns.

The question of whether net returns are Gaussian is an empirical one. We at least know that *if* net returns are Gaussian, they are very tractable for analysis

³The two rates are not exactly the same: when borrowing, the effective rate charged to the borrower by the lending institution is risk-free plus a small spread, and the rate paid by the same institution to a lender is risk-free minus a spread. For modeling purposes, we consider them identical.

⁴<https://www.newyorkfed.org/markets/reference-rates/sofr>.

at a given point in time. However, they are not easily tractable in time series analysis. For example, define the cumulative total return over periods $1, \dots, T$.

$$\begin{aligned} r_i(1 : T) &:= \frac{P_i(T)}{P_i(0)} - 1 \\ &= \frac{P_i(T)}{P_i(T-1)} \frac{P_i(T-1)}{P_i(T-2)} \cdots \frac{P_i(1)}{P_i(0)} - 1 \\ &= (r_i(T) + 1) \times (r_i(T-1) + 1) \times \dots \times (r_i(1) + 1) - 1 \end{aligned}$$

If $r_i(t)$ are normally distributed, the cumulative total return is not normal distributed, and its distribution rapidly diverges from the normal distribution.

The variance of the cumulative returns is not a simple function of the single-period variances.

On the other side, log returns compound under multiplication. Let $\tilde{r}(t) := \log(1 + r_i(t))$. Then, the log of the compound return is equal to the sum of the log returns in the same period, and if the log return is normal, so is the log of the compound returns. If the returns are independent, the variance of the log of compound log return is equal to the sum of the variances. We can reconcile the two view of returns – raw and log – if the approximation $\log(x) = x - 1 + o(|x-1|)$ is sufficiently accurate, i.e., if net returns are small. In this case, we can make the approximation $\tilde{r}_i \simeq r_i$, which is sufficiently accurate provided the returns are not too large.

A common approximation for the compounded net return of an asset over time is given by

$$\begin{aligned} \prod_t (r(t) + 1) - 1 &= \exp \left(\sum_t \tilde{r}(t) \right) - 1 \\ &\simeq 1 + \sum_t \tilde{r}(t) - 1 \\ &\simeq \sum_t r(t). \end{aligned}$$

Always verify the accuracy of the approximation, for example comparing the estimate of models developed using r and \tilde{r} . When the assets are equities, the approximation is usually considered adequate for daily interval measurements or shorter.

2.1.4 Estimating Prices and Returns

To estimate return, we need prices. Prices, however, depend crucially on the way a market is designed. Over-the-counter markets (Harris, 2003) differ from exchanges that employ limit-order books (Bouchaud et al., 2018). Within a single-exchange, the trading mechanism can change over the course of the day, with auctions often taking place at the beginning and at the close of the trading day. As a result of market design, the observation of prices exhibits measurement error. The most conspicuous example of such an error is the bid-ask spread. In limit order books, the buy orders have a price attribute (the “bidding” price per share the buyer is willing to pay) and a quantity. Similarly, the sell orders have a price attribute, or “asking price” and a quantity. Asking prices are higher than bidding prices, and the difference is called the bid-ask spread. This spread is a multiple of the minimum tick size⁵. For a transaction to occur, a buy order or a sell order must cross the spread; either event can occur. As a result, the transaction price will be either at the top or the bottom of the bid-ask spread interval. Successive transactions will have different price marks due to the partial randomness of buying and selling transaction. The bid-ask spread bounce is not the only source of measurement error. For example, prices can differ by exchanges, and the selection of price by timestamp depends on the choice of data integration. Then, there may be outright measurement errors. It is important to consider the fact that prices are imperfectly observed early on, rather than ignore them and their impact and face unintended consequences. Perhaps the simple model is the Roll model (Roll, 1984). Model for asset prices. In this model, the “true” price m_t of an asset evolves as an arithmetic random walk, and we imperfectly observe the price p_t . In formulas:

$$\begin{aligned} m_{t+1} &= m_t + \sigma_\epsilon \epsilon_{t+1} && \text{(evolution)} \\ p_{t+1} &= m_{t+1} + \sigma_\eta \eta_{t+1} && \text{(observation)} \end{aligned}$$

with ϵ_t, η_t independent random variables (serially and from each other) distributed according to a standard normal.

Before we try to estimate prices, the model has an immediate and testable consequence: consecutive price differences are negatively correlated. The price difference is $\Delta p_{t+1} := \sigma_\epsilon \epsilon_{t+1} + \sigma_\eta (\eta_{t+1} - \eta_t)$, which is zero in expectation. However,

$$\begin{aligned} E(\Delta p_{t+1} \Delta p_t) &= -\sigma_\eta^2 \\ E(\Delta p_{t+1} \Delta p_s) &= 0, \quad s < t \end{aligned}$$

⁵As of publication time, the minimum tick size is \$0.01 in US exchanges for shares trading above \$1.

The lag-one autocorrelation can also be used to estimate the measurement error. The presence of large non-zero autocorrelations beyond lag one may point to model inadequacy, in the sense that there are actual long-term dependencies in the price process m_t . The model can be extended; see Section 2.4. An optimal estimator for m_t is provided by the Kalman filter. The filter is covered in the Appendix, Section 2.5.1, and specifically in Example 1 of Subsection 2.5.2. The estimator is given by

$$\hat{m}_{t+1|t} = (1 - K)\hat{m}_{t|t-1} + Kp_t$$

Where the explicit formula for $K \in (0, 1)$ is given in the Appendix. The smaller the ratio $\sigma_\eta/\sigma_\epsilon$, the higher the K , which makes sense: we do not need to average observation if the price observations are accurate. The gist of the model is that an exponential moving averages of prices is preferable to just taking the last price in the measurement period. If we want the daily closing price, for example, we may want to use a weighted average of 5-minute interval prices in the preceding interval. There is a caveat, however. Suppose we have estimates \hat{m}_t , and we use these estimates to compute returns at intervals T ; i.e. $r_T := \hat{m}_{nT}/\hat{m}_{(n-1)T} - 1$. Because we employ the same observed prices p both in $\hat{m}_{(n-1)T}$ and \hat{m}_{nT} the two estimates are positively correlated. One should always check that $(1 - K)^T \ll 1$ to alleviate this spurious correlation.

2.1.5 Stylized Facts

Before building the house, we need to look at the bricks, namely, the statistical properties of the single-stock returns. Below I list some “stylized facts” about stock returns, and discuss their relevance to risk modeling and management. Returns have a lower bound at -1. We usually characterize the properties of $r(t) := \log(R(t))$. We focus on the properties of $r(t)$, but also $|r(t)|$ and $r^2(t)$, the volatility of the log returns. Here are some properties. See (Cont, 2001; Taylor, 2007; Ratliff-Crain et al., 2023).

1. *Absence of autocorrelations.* Lagged autocorrelations are small unless you observe prices and returns at time scales in which the market microstructure becomes relevant (say, intraday). See Fig. 2.1.
2. *Heavy tails.* The *unconditional* distribution of returns shows heavy tail behavior. This will be made more precise in the following section, but the probability of a large return is higher than what would be consistent with any “thin-tailed” distribution with infinite moments. Examples of sample

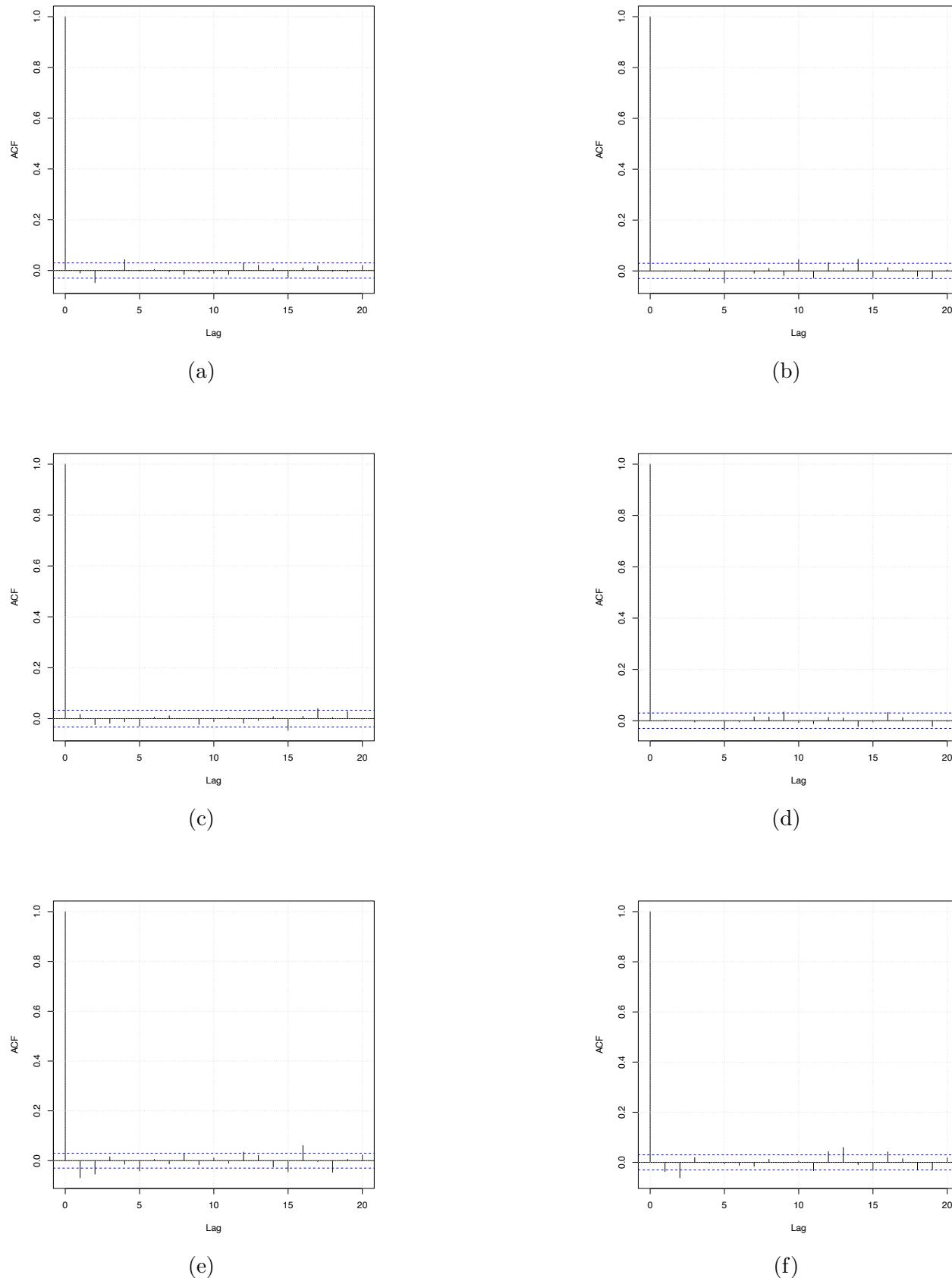


Figure 2.1: Autocorrelation plot of daily log returns (range: 1/3/2000-12/8/2017) for (a) AAPL, (b) IBM, (c) NRG, (d) WAT, (e) SPY, (f) XLK.

Stock	Skewness			Kurtosis		
	Mean	Left	Right	Mean	Left	Right
AAPL	-0.2	-0.5	0.2	5.7	3.6	7.8
IBM	0.1	-0.2	0.5	7.1	5.4	8.7
NRG	0.4	-0.5	1.2	14.3	7.9	20.0
WAT	-2.0	-3.3	-0.6	29.8	12.8	48.1
SPY	-0.1	-0.7	0.6	11.4	6.5	16.0

Table 2.1: Sample skewness and kurtosis of daily log returns and $p = 0.01$ confidence intervals estimated using nonparametric bootstrap with replacement (5000 variates). Range: 1/3/2001-12/8/2017.

kurtosis are in Table 2.1. The *conditional* (say, conditional on the return’s entire history up to time t) distribution of returns may show heavy tail behavior as well, but with lighter tails than the unconditional one.

3. *Autocorrelation of absolute returns and second moments.* The time series $|R(t)|$ and R_i^2 show strong autocorrelation. The autocorrelation of absolute values is greatest and is called the “Taylor Effect” in the literature (Taylor, 1986; Granger and Ding, 1995).
4. *Aggregational Gaussianity.* At longer time scales (say, weekly or monthly returns, as opposed to daily or intraday returns), the distribution of returns becomes closer to a Gaussian distribution.

Reality⁶ is in stark contrast with simple models of univariate price dynamics like the geometric diffusion process at the core of simple derivative pricing models:

$$dP(t) = \mu P(t)dt + \sigma P(t)dW(t) \quad (2.1)$$

This model predicts Gaussian, independent log returns, which are inconsistent with the empirical evidence. First, returns show little serial autocorrelation. This *does not mean* that returns are independent, nor that returns are unpredictable based on the history of returns or some additional explanatory variables. Regarding the former point: zero-autocorrelation does not imply independence.

Regarding the latter, *returns are predictable*. This is not only an article of faith of active investors, who usually do a terrible job at it, but also a relatively

⁶Note, however, that I am not including the Leverage Effect among the stylized facts. In the words of Cont (2001), “most measures of volatility of an asset are negatively correlated with the returns of that asset”. This effect is not sufficiently strong in recent data, as shown by Ratliff-Crain et al. (2023).

uncontroversial empirical finding among academics⁷. Nevertheless, even though they are predictable, they are not so trivially predictable.

Regarding heavy tails: for asset returns, we restrict our attention to power-tailed distributions: the complement of the cumulative distribution function follows a power law: $\bar{F}(x) := P(r > x) = Cx^{-\alpha}$, with $\alpha > 0$. Compare this to Gaussian returns: if $r \sim N(0, 1)$, then a common approximation Wasserman (2004) for the tail probability is

$$\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \left(\frac{1}{x} - \frac{1}{x^3} \right) \leq \bar{F}(x) \leq \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \frac{1}{x} \quad (2.2)$$

For the case $|x| \geq 1$, the right-side inequality can be used to bound quantiles and the symmetric inequality of the left tail:

$$\bar{F}(x) \leq \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \Rightarrow \bar{F}^{-1}(1 - \delta) \leq \sqrt{2 \log[1/((\sqrt{2\pi}(1 - \delta)))]} \quad (2.3)$$

$$F(x) \geq \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \Rightarrow F^{-1}(\delta) \geq -\sqrt{2 \log[1/(\sqrt{2\pi}\delta)]} \quad (2.4)$$

The approximation is quite accurate: $-\sqrt{2 \log[1/(\sqrt{2\pi}\delta)]} \leq F^{-1}(\delta) \leq -0.965 \times \sqrt{2 \log[1/(\sqrt{2\pi}\delta)]}$ for $10^{-10} < \delta < 1$. A Gaussian random variable has finite moments of any order. A power-tail random variable with exponent α has finite moment only up to α . A Gaussian random variable has quantiles bounded, up to constants, by $\sqrt{\log(1/\delta)}$, while a power-tail one has a quantile of the form $-(1/\delta)^{1/\alpha}$. It is not controversial that the unconditional log returns have heavy tails. It is still not settled what the exponent α associated with the distribution is. it seems however that $\alpha \simeq 4$. This is important for estimation purposes. A sufficient condition for the estimability of the volatility of returns is that their fourth moment is finite. To see this, recall that the Central Limit Theorem says that, if x_t are iid random variables with mean μ and variance σ^2 , then $T^{-1/2} \sum_{t=1}^T x_t$ converges in distribution to a Gaussian random variable with mean μ and variance σ^2 . The theorem allows us to establish an asymptotic result on $E(r^2)$: assume that r_i are iid. Set $x_t := r_t^2$. If we want to estimate $E(r_t^2)$ using the CLT, then we need finiteness of $E(r_t^4)$. This seems to be the case. However, a related question is whether the *conditional* return distribution is heavy-tailed. If the heavy tailed characteristic of conditional returns is ignored or considered

⁷John Cochrane has written extensively on this, e. g., Cochrane (2008) and the blog entry “Predictability and correlation” (<http://johnhcochrane.blogspot.com/2014/01/predictability-and-correlation.html>)

inessential, then it is possible to model returns as a process with conditional Gaussian returns and heavy-tailed unconditional ones. This family, denoted Conditional Heteroscedastic Models, is rich and the subject of the following subsection. We won't cover models with long-range dependence and/or heavy tailed conditional and unconditional returns, like Lévy processes and FARIMA models. No model covers all the empirical features observed in stock returns. GARCH models (and mixture models in general) have the benefit of being easy to interpret, simulate, and estimate.

2.2 Conditional Heteroscedastic Models (CHM)

This family of models was first proposed in the early 1980s by [Engle \(1982\)](#); [Engle and Bollerslev \(1986\)](#). By the next decade they had been generalized and applied to several economic domains. They are extensively covered in any Econometrics book.

The most popular and studied model in this family is the GARCH(1,1) model. It has good empirical properties, its theoretical properties have been characterized, and can be estimated efficiently. It also conveys the gist of the large set of models in this family. The fundamental insight of the model is to make the *parameters* in the model a part of the state of the stochastic process. The laws for GARCH(1,1) are

$$r_t = h_t \epsilon_t \tag{2.5}$$

$$h_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 h_{t-1}^2 \tag{2.6}$$

$$\epsilon_t \sim N(0, 1) \tag{2.7}$$

To gain some intuition, let us look at the second equation of the GARCH process when we remove the term $\alpha_1 r_{t-1}^2$. The equation

$$h_t^2 = \alpha_0 + \beta_1 h_{t-1}^2 \tag{2.8}$$

can be rewritten as

$$h_t^2 - h^2 = \beta_1(h_{t-1}^2 - h^2)$$

where

$$h^2 := \frac{\alpha_0}{1 - \beta_1}$$

The value of h_t^2 converges to h^2 at a geometric rate, so long as $|\beta_1| < 1$. High values of the squared return r_t^2 shock the volatility upward, provided that $\alpha_1 > 0$. This in turn increases the probability of large squared returns in the following period, giving rise to a rich dynamic behavior. The increase in volatility cannot continue unabated, because the term $\beta_1(h_{t-1}^2 - h^2)$ will dampen variances that are much greater than the “equilibrium level” h^2 . This can be seen through substitution in the second equation of the model:

$$h_t^2 = h^2 + \alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} r_{t-i}^2 \quad (2.9)$$

One could replace the true values of $\alpha_0, \alpha_1, \beta_1$ with estimates, and interpret the formula by saying that the variance estimate is an exponential moving average of non-iid returns, since they are modulated by h_t , in light of Equation (2.5).

2.2.1 GARCH as random recursive equations*

We now look at GARCH(1, 1) through different modeling approaches. First, we could reformulate it as a random iterated function. Rewrite Equation (2.6) as

$$h_t^2 = \alpha_0 + \alpha_1 h_{t-1}^2 \epsilon_{t-1}^2 + \beta_1 h_{t-1}^2$$

Set

$$a_t := \beta_1 + \alpha_1 \epsilon_{t-1}^2$$

The random variables a_t are iid. Then

$$h_t^2 = a_t h_{t-1}^2 + \alpha_0$$

This formulation shows that the process is Markovian, and that the variance process is governed by an autoregressive equation with random coefficients. This allows us to study the process using the toolkit of random recursive equations. By recursion (Lindner, 2009), we can rewrite the equations as

$$h_t^2 = \left(\prod_{i=0}^k a_{t-i} \right) h_{t-k-1}^2 + \alpha_0 \sum_{i=0}^k \prod_{j=0}^{i-1} a_{t-j}$$

The product $x_t := \prod_{i=0}^t a_i$ plays an important role (Nelson, 1990). If we can identify the conditions under which it converges to zero almost surely (a.s.)

and fast enough to guarantee that $\sum_{i=0}^k \prod_{j=0}^{i-1} a_{t-j}$ is finite a.s., then we have proven the existence of an asymptotic limit for h_t^2 . Let us consider the process $\{x_t : t > 0\}$. First, assume $x_t \geq 0$; it diverges if and only if $\log x_t \rightarrow \infty$. We then have to find the conditions under which

$$\sum_{i=0}^t \log(\beta_1 + \alpha_1 \epsilon_{i-1}^2) \rightarrow \infty \quad \text{a.s.}$$

Since this is the sum of iid random variables, a necessary and sufficient condition for this is that $\mu := E[\log(\beta_1 + \alpha_1 \epsilon_0^2)] > 0$, provided that the variance of $\log(\beta_1 + \alpha_1 \epsilon_0^2)$ is finite. If that is the case, then we can apply the Strong Law of Large Numbers:

$$\frac{1}{t} \sum_{i=0}^t \log(\beta_1 + \alpha_1 \epsilon_{i-1}^2) \rightarrow \mu \quad \text{a.s.}$$

Conversely, assume that $E[\log(\beta_1 + \alpha_1 \epsilon_0^2)] < 0$. Then $\log x_t \rightarrow -\infty$ a.s., and $x_t \rightarrow 0$ a.s. Under this condition, the unconditional variance is

$$h_t^2 = \alpha_0 \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} a_{t-j} \tag{2.10}$$

The kurtosis of the process is

$$k = \frac{3(1 + \alpha_1 + \beta_1)(1 - \alpha_1 - \beta_1)}{1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2} = 3 \frac{1 - (\alpha_1 + \beta_1)^2}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3$$

so the process is leptokurtic as long as $\alpha_1 > 0$. How about skewness? The the unconditional returns are not skewed, because

$$E((r_\infty - Er_\infty)^3) = E(h_\infty^3)E(\epsilon_t^3) = 0$$

Finally, we point out that not only are the unconditional returns leptokurtic, but do in fact have Pareto tails, provided the process is stationary: $P(r_t > x) \sim x^{-\alpha}$, for some $\alpha > 0$; see [Mikosch and Stărică \(2000\)](#); [Buraczewski et al. \(2016\)](#).

2.2.2 GARCH(1,1) and Return Stylized Facts

The GARCH improves on the distributional properties of returns of returns r_t , by making them closer to normal; see Figure 2.2. How does the GARCH(1, 1) model stack up against the stylized facts?

1. *Absence of autocorrelations.* This property is satisfied (not hard to verify directly).
2. *Heavy Tails.* The unconditional returns are leptokurtic. Moreover ([Mikosch and Stărică, 2000](#)), the tails of the unconditional returns are heavy tailed. So, this checks out. *However*, wait until point 4 below (autocorrelation of absolute return) before you celebrate.
3. *Autocorrelation of absolute and squared returns.* The ACF for GARCH(1, 1) is positive for both absolute and squared returns. For squared returns, it has the form ([He and Teräsvirta, 1999](#); [Ruppert and Matteson, 2015](#))

$$\rho_n = \begin{cases} \frac{\alpha_1(1 - \alpha_1\beta_1 - \beta_1^2)}{1 - 2\alpha_1\beta_1 - \beta_1^2} & \text{if } n = 1 \\ \rho_1(\alpha_1 + \beta_1)^{n-1} & \text{if } n > 1 \end{cases}$$

However, if we look at kurtosis and lag-1 autocorrelation for common stock indices, it appears that the autocorrelation is *too high* for a given observed kurtosis level. See [Teräsvirta \(2009a\)](#).

4. *Aggregational Gaussianity.* Although there are no known results on this properties, to the best my knowledge, it is satisfied empirically.

Table 2.2: Kolmogorov-Smirnov distances between the theoretical normal distribution and the empirical distribution of log returns and residuals of GARCH(1, 1) of log returns. The distance is reduced in all instances, with the largest improvements for the two proxies for the market (SPY) and the technology sector (XLK). For background on the Kolmogorov-Smirnov distance, see [DeGroot and Schervish \(2012\)](#), Ch. 10.

Stock	Unconditional	GARCH(1, 1)
AAPL	0.067	0.044
IBM	0.078	0.047
NRG	0.088	0.060
WAT	0.109	0.091
SPY	0.098	0.040
XLK	0.091	0.043

Summing up, some but not all of the stylized facts about log-returns are captured by GARCH(1, 1).

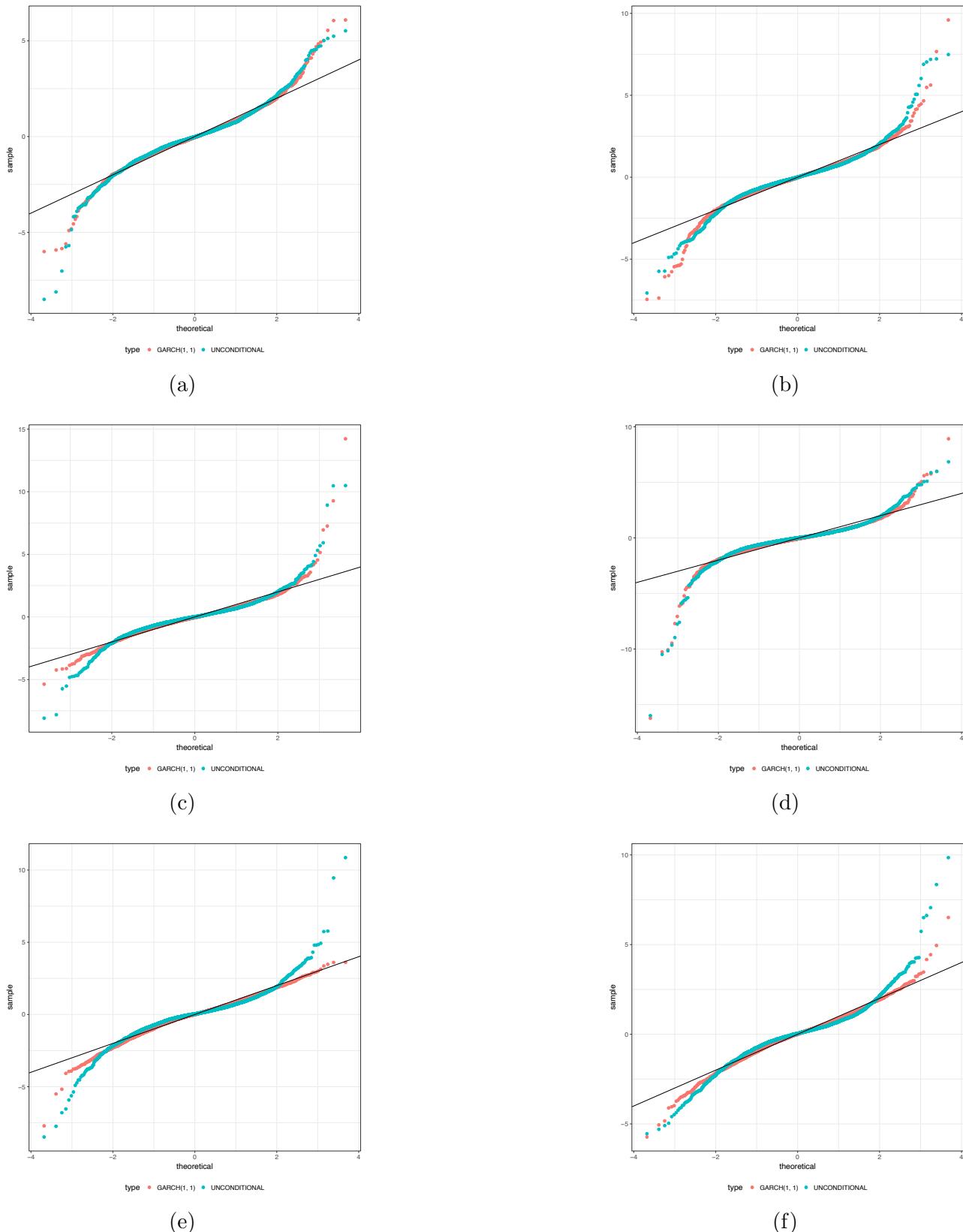


Figure 2.2: Quantile-Quantile plot for daily log returns (blue dots) and GARCH(1,1) residuals (orange dots) of log returns against the theoretical normal distribution for (a) AAPL, (b) IBM, (c) NRG, (d) WAT, (e) SPY, (f) XLK. Return range: 1/3/2001-12/8/2017.

Table 2.3: Estimated α for left and right tail of probability density function $p(x) \propto x^{-\alpha}$. We use the MLE estimator $\hat{\alpha} = 1 + n[\sum_i \log(x_i/x_{\min})]^{-1}$, where n is the number of observations above a cut-off value x_{\min} . The value of x_{\min} is set to -2.5 and 2.5 respectively. The values of α increases sizably for the two indices SPY and XLK.

Stock	Left Tail		Right Tail	
	Unconditional	GARCH(1, 1)	Unconditional	GARCH(1, 1)
AAPL	4.8	4.6	4.8	4.9
IBM	4.3	3.9	4.2	4.6
NRG	4.0	5.9	3.8	4.1
WAT	3.4	3.2	4.3	4.0
SPY	4.1	5.9	4.2	8.5
XLK	5.0	6.3	4.4	5.9

2.2.3 ★GARCH(1,1) Estimation

The vast majority of CHM applications are of order $p = q = 1$, so we restrict our analysis to this case for simplicity. Generalization to finite-order processes is straightforward. Define $\boldsymbol{\theta} := (\alpha_0, \alpha_1, \beta_1)$, and let f be the log density function of the standard normal distribution.

$$\begin{aligned} r_t &= h_t \epsilon_t \\ h_t^2 &= \phi(h_{t-1}^2, r_{t-1}^2, \boldsymbol{\theta}) \end{aligned}$$

By repeated substitution, we can express the unobserved variance h_t^2 as a function of the sequence r_1, \dots, r_{t-1} and θ . The log-likelihood of the sequence $\epsilon_{t-1} = r_t/h_t$ is given by

$$L(\theta) = \sum_{t=1}^T f \left(\frac{r_t}{h_t(r_1, \dots, r_{t-1}, \theta)} \right)$$

We can then estimate the parameters θ of the model by maximizing the log-likelihood. As an example, consider the GARCH(1,1) model. The recursive

equation for h_t^2 is given by Equation (2.9), so we solve

$$\begin{aligned} \min & \sum_{t=1}^T \left(\log h_t^2 + \frac{r_t^2}{h_t^2} \right) \\ \text{s.t. } & h_t = \left(\alpha_0 \frac{1 - \beta_1^{t-1}}{1 - \beta_1} + \alpha_1 \sum_{i=1}^t \beta_1^{i-1} r_{t-i}^2 \right)^{1/2} \quad t = 1, \dots, T \end{aligned}$$

2.2.4 Realized Volatility

CHMs model the asset volatility as an (unobserved) state of the return stochastic process. Once we have an estimate of the volatility at time t of returns, the rest is trivial. An alternative route would be to estimate directly the volatility from the data, for example with a simple moving-window estimator of the empirical volatility. This approach would not work if the epochs for which we need the estimates are days, and we only have daily data. In recent years, tick-level price data have become widely available; indeed, order-book level data are also available (with the entire process of order arrivals, fillings, and cancellations). It is now possible to compute 1-minute returns, enabling us to estimate the volatility of returns for daily predictions by using these high-frequency data. Below we review some of the statistical properties of realized volatility measurements. The starting point is Equation (2.1), i.e., a diffusion process for the log price $p(t) = \log P(t)$:

$$dp = \alpha dt + \sigma dW$$

where $W(t)$ is a Brownian process. $\alpha \in \mathbb{R}$ (the drift) $\sigma > 0$ (the volatility) are constants. In all applications of interest, the drift is much smaller than the volatility: $|\alpha| \ll \sigma$. The quantity α/σ is termed the (daily) *Sharpe Ratio* and will figure prominently in the rest of the book⁸. We observe the process in the interval $[0, 1]$ and measure the state variable p at intervals of length $1/n$. The measured return is $r(j) := p(j/n) - p((j-1)/n)$. Clearly, $r(j)$ are iid random variables, and $r(j) \sim N(\alpha/n, \sigma^2/n)$. The maximum likelihood estimators for

⁸This is the Sharpe Ratio of log returns, which is to a first approximation close to the daily Sharpe Ratio computed on returns.

drift and moments are

$$\hat{\alpha} = \sum_j r(j) = p(1) - p(0)$$

$$\hat{\sigma}_1^2 = \sum_j [r(j) - \hat{\alpha}/n]^2$$

We also consider the *uncentered* estimator of the volatility⁹.

$$\hat{\sigma}_2^2 = \sum_j r^2(j) \quad (2.11)$$

The first remarkable phenomenon is that the MLE estimator for the drift does not depend on the number of intervals n . Moreover, one can show that $\text{var}(\hat{\alpha}) = \text{var}(p(1) - p(0))$, and $p(1) - p(0) \sim N(\alpha, \sigma)$, so that $\text{var}(\hat{\alpha}) = \sigma^2$. The estimation error does not depend on the number of intervals either. To estimate the variance of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ we need a few formulas. The moments of $r(j)$ are those of a Gaussian random variable with mean α/n and variance σ^2/n :

$$E[r(j)] = \frac{\alpha}{n} \quad (2.12)$$

$$E[r^2(j)] = \left(\frac{\alpha}{n}\right)^2 + \frac{\sigma^2}{n} \quad (2.13)$$

$$E[r^4(j)] = \left(\frac{\alpha}{n}\right)^4 + 6\left(\frac{\alpha}{n}\right)^2 \frac{\sigma^2}{n} + 3\left(\frac{\sigma^2}{n}\right)^2 \quad (2.14)$$

so that

$$\text{var}(r^2(j)) = 2\left(\frac{\sigma^2}{n}\right)^2 + 4\left(\frac{\alpha}{n}\right)^2 \frac{\sigma^2}{n} \quad (2.15)$$

⁹An early analysis of the “vanilla” Realized Variance estimator is Barndorff-Nielsen and Shephard (2002) and a survey is Andersen and Benzoni (2009). Also useful are the surveys of Andersen et al. (2006, 2013), which situate realized volatility in the context of risk management techniques. Essential readings on realized volatility estimators are Zhang et al. (2005), which presents several estimators and introduces the idea of subsampling for RV; the series of papers Barndorff-Nielsen et al. (2008, 2009) on kernel-based estimators; the empirical paper by Liu et al. (2015), comparing several estimators, which includes subsampling and kernel. This list of estimators is not exhaustive. For example, Hansen and Lunde Hansen and Lunde (2006b) analyze an autocorrelation-adjusted estimator introduced in French et al. (1987). Bipower estimators are studied by Podolskij and Vetter (2009) and maximum likelihood ones by Aït-Sahalia et al. (2005). Moreover, these estimators depend on several parameters, like sampling and subsampling intervals, or the choice of kernel.

and

$$E(\hat{\sigma}_2^2) = \sigma^2 + \frac{\alpha^2}{n} \quad \text{from Equation (2.13)}$$

$$\text{var}(\hat{\sigma}_2^2) = 2\frac{\sigma^4}{n} + 4\left(\frac{\alpha}{n}\right)^2\sigma^2 \quad \text{from Equation (2.15)}$$

The estimator $\hat{\sigma}_2^2$ has a small finite-sample bias and is asymptotically consistent.

Insight 2.1: *Estimating Variance*

Based on Equations (2.13) and (2.15), you can use uncentered returns for variance estimation, since the bias is inversely proportional to n , and the estimator is consistent.

Let us reflect on the steps we took. We discretized the interval over which the price process occurs into n subintervals, and retained only the last price within an interval of length $1/n$, assuming the price had no measurement error. We saw that the drift estimator is unbiased, but its variance does not depend on the discretization: we have more estimates of the drift, but they are noisier. Unfortunately, there is no easy way to measure the drift, i.e., the expected return, of a security; otherwise, all Statistics undergraduates would be rich. Conversely, we have identified an uncentered estimator of the true variance σ^2 . As the number n of intervals approaches infinity, the estimator is unbiased. Its variance decreases like $2\sigma^2/n$, which is good news: in principle, we can estimate to arbitrary accuracy the volatility of the returns at time t ; and provided that the true volatility varies very little over time, we can use this estimate to predict variance at time $t+1$. The good news here is that if you need volatility estimates over a long time scale for your decisions (e.g., days), but have data over a shorter time scale (e.g., minutes), you do not have to devise a generative model like CHMs or others. What assumptions do not hold in this line of reasoning? Here is a list of issues to consider:

1. We ignored market microstructure. One source of noise is the bid-ask spread ([Harris, 2003](#)). When the seller initiates the transaction, she receives the bid price; when the buyer initiates it, he pays the ask price. There is an intrinsic error in the measurement of price, which is approximately equal to half the bid-ask spread. Measured log prices in interval t are

$p_t + \epsilon_t$, where the noise terms ϵ_t are iid random variables of the length of the measurement interval.

2. Another form of microstructure is thinly-traded securities, etc. If a stock trades less than once every five minutes on average, then using one-minute intervals is probably not a good modeling choice.
3. We assumed that volatility is changing slowly, or is ideally constant. This is not the case in practice. One approach is to impose a model on the time series of realized variances, so that we can produce an error estimate. E.g., a simple AR(1) model $\hat{\sigma}(t+1) = a + b\hat{\sigma}(t) + \tau\epsilon(t+1)$, with $\epsilon(t+1) \sim N(0, 1)$.
4. We ignored the distinction between open-to-close and close-to-open interval. Close-to-open returns are often fundamentally driven. Also, we are ignoring the large volatility and bid-ask spreads in the first minutes of the trading day.

For the rest of us, the question is: what to choose? Liu et al. (2015) compare a broad set of estimators, with several choices of parameters, for assets in different asset classes (equities, futures, indices). They use Romano and Wolf's procedure for multiple comparison (Romano and Wolf, 2005) and Hansen *et al.* "model confidence set" (Hansen et al., 2011). They find that the Vanilla RV at 5-minutes intervals performs competitively across various assets and asset classes. There are a few cases where this is not true. When higher-frequency measurements are available, this estimator is outperformed by a one-minute subsampled RV, by 1- and 5-second interval realized kernel. In addition, at lower frequencies, 5- and 15-minute truncated RV (Mancini, 2009, 2011) also outperforms vanilla RV. However, where available, 5-min nonoverlapping intervals seem to be a reasonable choice.

2.2.5 Combining CHM and Realized Volatility

Is it possible to have the best of both worlds, GARCH models and realized volatility? Hansen et al. (2012) present a model, RealGARCH(1,1), that combine both.

$$\begin{aligned} r_t &= h_t \epsilon_t \\ h_t^2 &= \alpha_0 + \beta_1 h_{t-1}^2 + \gamma x_{t-1} \\ x_t &= \xi + \phi h_t^2 + u_t \end{aligned} \tag{2.16}$$

The first two equations are similar to the standard GARCH(1, 1) model, with one difference: the term proportional to r_{t-1}^2 has been replaced by a term proportional to x_{t-1} . This variable is the observed estimate of the realized variance at time t ; when this estimator is more accurate than the rough estimate of variance r_{t-1}^2 , then the model will probably outperform GARCH(1, 1). The last Equation (2.16) models the dynamic behavior of the realized variance. It posits a linear dependence on h_t and on a stochastic term u_t . The random variables u_1, \dots, u_t are iid random variables, not necessarily with zero mean.

2.3 State-Space Estimation of Variance

2.3.1 Muth's Original Model: EWMA

A very popular estimator of the expected value of a time series $\{x_s\}$, based on data up to time t , is the *exponentially weighted moving average* (or EWMA). It takes the form

$$\hat{x}_t = (1 - K) \sum_{s=0}^{\infty} K^s x_{t-s}$$

for some $0 < K < 1$. We discount the past by giving its observations exponentially decreasing weights, which makes sense, and even more so when we write the estimate as a recursion:

$$\hat{x}_t = (1 - K)x_t + K\hat{x}_{t-1}$$

A low value of K forgets the past faster. The formula is computationally efficient both in terms of storage and computation. For uncentered variance estimation of a return, this takes the form

$$\hat{\sigma}_t^2 = (1 - K)r_t^2 + K\hat{\sigma}_{t-1}^2 \quad (2.17)$$

In academic journals, EWMA receives relatively low attention compared to GARCH models (for a rare example, see [Ding and Meade \(2010\)](#)); among practitioners, including major commercial risk model providers like RiskMetrics, Barra, and Axioma, it is the other way around. Aside from these practical considerations, is it possible to motivate the approach based on a model? We devote this section to understanding and extending this simple formula.

We will employ linear state-space models and Kalman Filters, which are briefly covered in the Appendix, Section 2.5.1. Rather than giving it a general

Insight 2.2: GARCH are EWMA with an offset

Recall Equation (2.9):

$$h_t^2 = \frac{\alpha_0}{1 - \beta_1} + \alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} r_{t-i}^2$$

This is, save for an offset, very similar to Equation (2.17):

$$\hat{\sigma}_t^2 = (1 - K) \sum_{i=1}^{\infty} K^{i-1} r_{t-i}^2$$

(we have changed the indexing convention to make it consistent with GARCH). The two are identical when $\alpha_0 = 0$, $\alpha_1 = 1 - K$, and $\beta_1 = K$.

treatment and then specializing to a specific model, we will jump right in the middle with a relevant example. As it happens, this example is also the simplest non-trivial example of a state-space model. The model (Muth, 1960) posits that there is a scalar *state* x_t that evolves randomly over time with the addition of a gaussian disturbance to its previous value. We observe the state imperfectly; the *observation* y_t is a noisy measure value x_t . In formulas:

$$\begin{aligned} x_{t+1} &= x_t + \tau_\epsilon \epsilon_{t+1} \\ y_{t+1} &= x_{t+1} + \tau_\eta \eta_{t+1} \\ \epsilon_t &\sim N(0, 1) \\ \eta_t &\sim N(0, 1) \end{aligned}$$

The innovations and the measurement noises are gaussian with mean zero, and their are independent of each other: $\epsilon_s \perp \epsilon_t$ $\eta_s \perp \eta_t$ for all $s \neq t$, and $\epsilon_s \perp \eta_t$ for all t and s . I skipped the derivation, which the interested reader can find in the Appendix. Define the ratio of measurement to innovation noise $\kappa := \tau_\eta / \tau_\epsilon$. The stationary $\hat{\sigma}_{t+1|t}$ standard deviation of the state estimate is given by:

$$\hat{\sigma}_{t+1|t}^2 = \tau_\epsilon^2 \frac{1 + \sqrt{(2\kappa)^2 + 1}}{2}$$

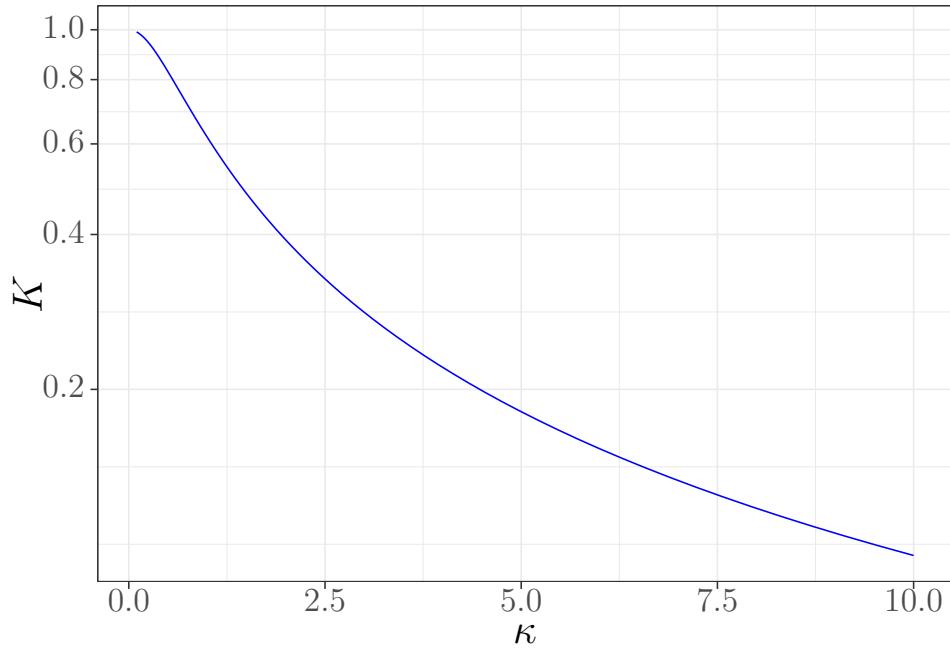


Figure 2.3: Relationship between K and $\kappa := \tau_\eta/\tau_\epsilon$.

and the optimal estimation recursion is

$$K := \frac{\hat{\sigma}_{t+1|t}^2}{\hat{\sigma}_{t+1|t}^2 + \tau_\eta^2}$$

$$\hat{x}_{t+1|t} = (1 - K)\hat{x}_{t|t-1} + Ky_t$$

For $\kappa \gg 1$ the formula simplifies:

$$\hat{x}_{t|t} = \frac{\kappa}{1 + \kappa}\hat{x}_{t|t-1} + \frac{1}{1 + \kappa}y_t$$

This is an exponential weighted average with a simple interpretation. Imagine that the state does not change at all. Then we want to use all the history we can, since old observations and new ones are drawn from the same distribution. The half-life of EWMA is indeed long. Conversely, when the state changes at a rapid pace, i.e., $\kappa \simeq 0$, then we want to discount the past very aggressively. According to Muth's original model applied to volatility estimation, the state is the instantaneous variance, and the observation y_t is r_t^2 , which is equal to σ_t^2 in expectation.

The model has obvious shortcomings. If returns are normally distributed, then the observation error is not normally distributed. More importantly, the model allows for negative values of the variance, and additionally models the

variance evolution as the sum of iid innovations. Over time, the distribution of the variance becomes more and more spread out: the standard deviation of the distribution grows as the square root of the number of periods. In practice, however, volatility appears to revert to a long-term average.

We cannot directly address the first problem. Kalman filters can work well with non-normal innovations and measurement errors, provided that these are not too heavy-tailed. As for the other shortcomings, we can refine the model to accommodate them. For example, we can introduce a mean-reverting model of variance, so that it behaves like an autoregressive process. We extend slightly the state equation by adding a mean-reversion term:

$$x_{t+1} = x_t - \lambda(x_t - \mu) + \tau_\epsilon \epsilon_{t+1}$$

The state reverts to value μ when it is away from this equilibrium value. The stationary distribution of x_t is gaussian, with the expected value equal to μ and standard deviation equal to $\tau_\epsilon^2 / (2\lambda - \lambda^2)$. The optimal variance estimator is still

$$\hat{x}_{t+1|t} = (1 - K)\hat{x}_{t|t-1} + Ky_t$$

However, compared to the first model, the value of K , when $\lambda > 0$ is *smaller*. Otherwise stated, the mean reversion term makes the distribution of the true variance more concentrated around its long-term mean. This means that we discount the past less. The detailed derivation of these formulas is in the Appendix, Section 2.5.2.

Insight 2.3: *On the Relative Merits of ECH and EWMA*

TODO

2.3.2 The Harvey-Shephard Model*

As a final example of the flexibility that linear state-space models can offer, I present the model by [Harvey and Shephard \(1996\)](#), which has several desirable features: it has a closed-form solution; the volatility is by design positive and the distribution of the volatility itself is log-normal, hence right-skewed, as we would expect; and the stock returns are locally lognormal.

The generating process for returns r_t is assumed to be

$$r_t = e^{\beta + \exp(x_t/2)\xi_t} - 1 \quad (2.18)$$

where β is a known constant, and $\xi_t \sim N(0, 1)$; hence returns are, at any point in time, lognormally distributed. Define

$$\begin{aligned} u_t &:= \log(1 + r_t) - \beta \\ \Rightarrow u_t &= \exp(x_t/2)\xi_t \\ \Rightarrow \log u_t^2 &= x_t + \log \xi_t^2 \\ &= x_t + \eta_t + \gamma \end{aligned}$$

where $\gamma := E(\log \xi_t^2) \simeq -1.27$, and η_t is a zero-mean random variable with standard deviation $\text{stdev}(\log \xi_t^2) \simeq 2.22$. Define

$$\begin{aligned} y_t &:= \log u_t^2 - \gamma \\ &= \log[(\log(1 + r_t) - \beta)^2] - \gamma \end{aligned}$$

so that we get an observation equation:

$$y_t = x_t + \eta_t$$

Now, we posit an evolution equation for x_t :

$$x_{t+1} = b + ax_t + \epsilon_t$$

This is the same model as AR(1), from which we obtain an estimate \hat{x}_t . If $\beta = 0$, then the formulas take a simple form: $u_t \simeq r_t$ and the state estimate is given by

$$\hat{x}_{t+1|t} = (1 - K)\hat{x}_{t|t-1} + K[\log[(\log(1 + r_t) - \beta)^2] - \gamma]$$

Since $R_t = \exp(\exp(x_t/2)\xi_t)$ is a lognormal random variable, the estimated standard deviation of R_t is

$$\hat{\sigma}_{t+1|t} = \sqrt{(e^{\exp(\hat{x}_{t+1|t})} - 1)e^{\exp(\hat{x}_{t+1|t})}}$$

A simplified Harvey-Shephard model starts with Equation (2.18), to which it applies the first-order approximation $e^x - 1 \simeq x$, and the parameter $\beta = 0$:

$$r_t = \exp(x_t/2)\xi_t$$

Define

$$\begin{aligned} \log r_t^2 &:= x_t + \log \xi_t^2 \\ &= x_t + \eta_t + \gamma \end{aligned}$$

where γ and η_t are defined as for the Harvey-Shephard model above. The model is completed by the Equations, also from the original model,

$$\begin{aligned}x_{t+1} &= b + ax_t + \epsilon_t \\y_t &= \log r_t^2 - \gamma\end{aligned}$$

The state estimate and volatility estimates are

$$\begin{aligned}\hat{x}_{t+1|t} &= (1 - K)\hat{x}_{t|t-1} + K[\log r_t^2 - \gamma] \\\hat{\sigma}_{t+1|t} &= e^{\hat{x}_{t+1|t}/2}\end{aligned}$$

2.4 Further Reading

A slightly dated reference on stylized facts for asset returns is [Cont \(2001\)](#); more recent ones are [Taylor \(2007\)](#); [Zivot \(2009\)](#); [Ratliff-Crain et al. \(2023\)](#). Gaussian bounds like Equation (2.2) are classic results; see [Wasserman \(2004\)](#). [Vershinin \(2018\)](#) on high-dimensional probability is a comprehensive reference for various finite-sample bounds. The literature on GARCH models alone is immense; [Tsay \(2010\)](#); [Zivot and Wang \(2003\)](#); [Cižek et al. \(2011\)](#); [Ruppert and Matteson \(2015\)](#); [Tsay \(2010\)](#); [Lütkepohl \(2005\)](#) are standard references, and survey are [Andersen et al. \(2006, 2013\)](#). The handbook [Andersen et al. \(2009\)](#) has dedicated chapters covering univariate ([Teräsvirta, 2009a](#)) and multivariate GARCH ([Teräsvirta, 2009b](#)), moments of GARCH models ([Lindner, 2009](#)), their detailed extremal properties ([Davis and Mikosch, 2009](#)), multivariate GARCH. For a recent empirical paper on the performance of GARCH, TARCH, EGARCH and a few other models, see [Hansen and Lunde \(2005\)](#); [Brownlees et al. \(2011\)](#).

The convergence properties of Random Recursive Equations (RREs) were studied first by [Kesten \(1973\)](#); [Diaconis and Freedman \(1999\)](#) survey the general recursive equations $x_t = f(x_{t-1}, \epsilon_{t+1})$, where $(\epsilon_t)_{t=1}^\infty$ is an iid random sequence, of which RREs are a special case. A monograph on RREs, covering both the univariate and multivariate case, is [Buraczewski et al. \(2016\)](#).

There are, to a first approximation, infinite references on Kalman filtering. Aside from the original [Kalman \(1960\)](#); [Kalman and Bucy \(1961\)](#), more modern textbook treatments are [Harvey \(1990\)](#); [Simon \(2006\)](#); [Whittle \(1996\)](#).

Roll introduced his model in [Roll \(1984\)](#) and a detailed discussion of the Roll model and its extensions is in [Hasbrouck \(2007\)](#).

2.5 ★Appendix

2.5.1 The Kalman Filter

This section contains a short treatment of the Kalman Filter (KF). The Kalman Filter predates Kalman's original articles in the early 1960's ([Kalman, 1960](#); [Kalman and Bucy, 1961](#)). At the time of their publication, computers had become available that made calculations feasible in real time. This made the (re)discovery of the filter by Kalman very timely. Rockets used by the Apollo program contained implementations of the Kalman Filter in 2KB of RAM. Since the 60s, the topic of linear control and filtering has flourished. Thousands of papers have been written on it, and there are several monographs covering the Kalman Filter in details from different perspectives: control ([Simon, 2006](#)), statistical ([Harvey, 1990](#)), econometric ([Hansen and Sargent, 2008](#)). I cover the KF for two reasons. First, because, for somewhat mysterious reasons, the derivation of the KF is often more complicated than it should be. A rigorous yet, I hope, intuitive proof essentially fits in half a page and should save the reader a few hours. Secondly, I wanted to present the problem under two different lens, and show its close connection to the Linear Quadratic Regulator (LQR). Both problems are essential tools in the arsenal of the quantitative finance researcher, so there is value in catching two birds with one stone¹⁰.

We need the following elementary fact. Let $\mathbf{Z} := [\mathbf{x}, \mathbf{y}]'$ be multivariate normal random vector with mean and covariance matrix

$$\boldsymbol{\mu}_{\mathbf{Z}} := \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix} \quad \text{cov}(\mathbf{Z}) = \begin{bmatrix} \Sigma_{\mathbf{x}, \mathbf{x}} & \Sigma_{\mathbf{x}, \mathbf{y}} \\ \Sigma_{\mathbf{y}, \mathbf{x}} & \Sigma_{\mathbf{y}, \mathbf{y}} \end{bmatrix}$$

The random vector \mathbf{x} , conditional on $\mathbf{y} = \mathbf{b}$ is still normally distributed, with conditional mean and covariance matrix equal to

$$\begin{aligned} E(\mathbf{x}|\mathbf{y} = \mathbf{b}) &= \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{y}} \boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{y}}^{-1} (\mathbf{b} - \boldsymbol{\mu}_{\mathbf{y}}) \\ \text{cov}(\mathbf{x}|\mathbf{y} = \mathbf{b}) &= \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{x}, \mathbf{y}} \boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{y}}^{-1} \boldsymbol{\Sigma}_{\mathbf{y}, \mathbf{x}} \end{aligned}$$

This can be verified directly by integration.

Our model has two components. The first is a *state*, represented by a random vector \mathbf{x}_t . This vector follows a simple evolution rule: $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_{t+1}$. The vector $\boldsymbol{\epsilon}_t$ is random, serially independent, and distributed according to a multivariate normal distribution. The state is not observable directly; the

¹⁰However, should you catch birds, please don't use stones, but nets, or food.

only thing we know is its probability distribution at time 1. We assume it is normal with known mean and covariance matrix. In addition, over time we observe is a vector \mathbf{y}_t , which is a linear transformation of \mathbf{x}_t , corrupted by noise: $\mathbf{y}_{t+1} = \mathbf{B}\mathbf{x}_{t+1} + \boldsymbol{\eta}_{t+1}$. Note the similarity with the factor model equation:

$$\begin{aligned} \text{state} &\leftrightarrow \text{factor return} \\ \text{observation} &\leftrightarrow \text{asset return} \end{aligned}$$

What is different is that factors returns are usually not modeled as being serially dependent.

The vector $\boldsymbol{\eta}_t$ is random, serially independent, independent of $(\boldsymbol{\epsilon}_t)_{t=1}^\infty$, and distributed according to a multivariate normal distribution.

Summing up, the distributions of $\mathbf{x}_1, \boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t$ are given by

$$\begin{aligned} \mathbf{x}_1 &\sim N(\hat{\mathbf{x}}_0, \hat{\Sigma}_0) \\ \boldsymbol{\epsilon}_t &\sim N(\mathbf{0}, \Sigma_\epsilon) & \boldsymbol{\epsilon}_t \perp \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_t \perp \boldsymbol{\eta}_{s+1} & s \leq t \\ \boldsymbol{\eta}_t &\sim N(\mathbf{0}, \Sigma_\eta) & \boldsymbol{\eta}_t \perp \boldsymbol{\eta}_s, \boldsymbol{\eta}_t \perp \boldsymbol{\epsilon}_{s+1} & s \leq t \end{aligned}$$

And the Linear State Space Model is given by

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_{t+1} \quad (2.19)$$

$$\mathbf{y}_{t+1} = \mathbf{B}\mathbf{x}_{t+1} + \boldsymbol{\eta}_{t+1} \quad (2.20)$$

I denote $\hat{\mathbf{x}}_{t|t-1}, \hat{\Sigma}_{t|t-1}$, the conditional estimates for the mean and covariance matrix of the state \mathbf{x}_t , based on the information $\mathbf{y}_0, \dots, \mathbf{y}_{t-1}$. And I denote $\hat{\mathbf{x}}_{t|t}, \hat{\Sigma}_{t|t}$ the estimates based on information $\mathbf{y}_0, \dots, \mathbf{y}_t$.

The vector \mathbf{Z}_t is defined as the combination of state and observation:

$$\mathbf{Z}_t := \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix}$$

Based on information up to time $t - 1$, the covariance of \mathbf{Z}_t is

$$\text{cov}(\mathbf{Z}_t) = \begin{bmatrix} \hat{\Sigma}_{t|t-1} & \Sigma_{\mathbf{x}_t} \mathbf{B}' \\ \mathbf{B} \Sigma_{\mathbf{x}_t} & \mathbf{B} \hat{\Sigma}_{t|t-1} \mathbf{B}' + \Sigma_\eta \end{bmatrix}$$

We observe \mathbf{y}_t . The vector \mathbf{x}_t is normally distributed. The conditional covariance of \mathbf{x}_t given \mathbf{y}_t is

$$\begin{aligned} \hat{\Sigma}_{t|t} &= \hat{\Sigma}_{t|t-1} - \hat{\Sigma}_{t|t-1} \mathbf{B}' (\mathbf{B} \hat{\Sigma}_{t|t-1} \mathbf{B}' + \Sigma_\eta)^{-1} \mathbf{B} \hat{\Sigma}_{t|t-1} & (update\ step) \\ &= [\mathbf{I} - \hat{\Sigma}_{t|t-1} \mathbf{B}' (\mathbf{B} \hat{\Sigma}_{t|t-1} \mathbf{B}' + \Sigma_\eta)^{-1} \mathbf{B}] \hat{\Sigma}_{t|t-1} \end{aligned} \quad (2.21)$$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \hat{\Sigma}_{t|t-1} \mathbf{B}' (\mathbf{B} \hat{\Sigma}_{t|t-1} \mathbf{B}' + \Sigma_\eta)^{-1} (\mathbf{y}_t - \mathbf{B} \hat{\mathbf{x}}_{t|t-1}) \quad (2.22)$$

Once we have the posterior distribution given the observation \mathbf{y}_t , the conditional distribution of \mathbf{x}_{t+1} follows from Equation (2.19). \mathbf{x}_{t+1} is Gaussian with the following conditional mean and covariance matrix:

$$\hat{\Sigma}_{t+1|t} = \mathbf{A}\hat{\Sigma}_{t|t}\mathbf{A}' + \Sigma_\epsilon \quad (prediction\ step) \quad (2.23)$$

$$\hat{\mathbf{x}}_{t+1|t} = \mathbf{A}\hat{\mathbf{x}}_{t|t-1} + \mathbf{A}\hat{\Sigma}_{t|t}\mathbf{B}'(\mathbf{B}\hat{\Sigma}_{t|t}\mathbf{B}' + \Sigma_\eta)^{-1}(\mathbf{y}_t - \mathbf{B}\hat{\mathbf{x}}_{t|t-1}) \quad (2.24)$$

The measurement and time update equations above are the whole of the Kalman Filter. If we combine Equations (2.21) and (2.23), the covariance matrix evolves according to the equation:

$$\hat{\Sigma}_{t+1|t} = \mathbf{A}(\hat{\Sigma}_{t|t-1} - \hat{\Sigma}_{t|t-1}\mathbf{B}'(\mathbf{B}\hat{\Sigma}_{t|t-1}\mathbf{B}' + \Sigma_\eta)^{-1}\mathbf{B}\hat{\Sigma}_{t|t-1})\mathbf{A}' + \Sigma_\epsilon$$

This is called a *Riccati recursion*. In steady state the covariance matrix does not change in consecutive periods: $\hat{\Sigma}_{t+1|t} = \hat{\Sigma}_{t|t-1}$. We can solve for the stationary matrix:

$$\mathbf{X} = \mathbf{A}\mathbf{X}\mathbf{A}' - \mathbf{A}\mathbf{X}\mathbf{B}'(\mathbf{B}\mathbf{X}\mathbf{B}' + \Sigma_\eta)^{-1}\mathbf{B}\mathbf{X}\mathbf{A}' + \Sigma_\epsilon$$

This is a *discrete time algebraic Riccati equation*.

The matrix

$$\mathbf{K}_t := \hat{\Sigma}_{t|t-1}\mathbf{B}'(\mathbf{B}\hat{\Sigma}_{t|t-1}\mathbf{B}' + \Sigma_\eta)^{-1}$$

is called the optimal Kalman gain. The equations become

$$\hat{\Sigma}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{B}]\hat{\Sigma}_{t|t-1} \quad (2.25)$$

$$\hat{\mathbf{x}}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{B})\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\mathbf{y}_t \quad (2.26)$$

$$\hat{\Sigma}_{t+1|t} = \mathbf{A}\hat{\Sigma}_{t|t}\mathbf{A}' + \Sigma_\eta \quad (2.27)$$

$$\hat{\mathbf{x}}_{t+1|t} = \mathbf{A}\hat{\mathbf{x}}_{t|t} \quad (2.28)$$

2.5.2 Kalman Filter Examples

Example 1 (Muth, 1960):

$$x_{t+1} = x_t + \tau_\epsilon \epsilon_{t+1} \quad (2.29)$$

$$y_{t+1} = x_{t+1} + \tau_\eta \eta_{t+1} \quad (2.30)$$

The stationary $\hat{\sigma}_{t+1|t}$ is given by the solution to the Riccati equation:

$$\begin{aligned} \frac{\hat{\sigma}_{t+1|t}^4}{\hat{\sigma}_{t+1|t}^2 + \tau_\eta^2} &= \tau_\epsilon^2 \Rightarrow \hat{\sigma}_{t+1|t}^2 = \frac{1}{2}\tau_\epsilon^2(1 + \sqrt{(2\kappa)^2 + 1}) \\ K &= \frac{\hat{\sigma}_{t+1|t}^2}{\hat{\sigma}_{t+1|t}^2 + \tau_\eta^2} \\ \hat{x}_{t+1|t} &= (1 - K)\hat{x}_{t|t-1} + Ky_t \end{aligned}$$

where we have introduced the parameter

$$\kappa := \frac{\tau_\eta}{\tau_\epsilon}$$

Loosely, this is a noise-to-signal ratio. It is high when the measurement error is high compared to the typical change of the state per period. For $\kappa \gg 1$ the formula simplifies: $K \simeq 1/(\kappa + 1)$.

$$\hat{x}_{t|t} = \frac{\kappa}{1 + \kappa}\hat{x}_{t|t-1} + \frac{1}{1 + \kappa}y_t$$

Example 2 (AR(1) model): In this model, the state equation is

$$x_{t+1} = b + ax_t + \tau_\epsilon \epsilon_t \quad (2.31)$$

To have a mean-reverting process, introduce a long-term mean value $\mu > 0$ and a relaxation constant $\lambda > 0$, and set

$$a := 1 - \lambda \quad (2.32)$$

$$b := \lambda\mu \quad (2.33)$$

Equation (2.31) becomes

$$x_{t+1} = x_t - \lambda(x_t - \mu) + \tau_\epsilon \epsilon_{t+1}$$

The state reverts to value μ when it is away from this equilibrium value. The stationary distribution of x_t is gaussian, with mean μ and standard deviation $\tau_\epsilon/\sqrt{2\lambda - \lambda^2}$.

Define:

$$u_t := x_t - \mu \quad (2.34)$$

$$v_t := y_t - \mu \quad (2.35)$$

We rewrite the equation as

$$\begin{aligned}x_{t+1} - \mu &= x_t - \mu + (a - 1)(x_t - \mu) + \tau_\epsilon \epsilon_{t+1} \\u_{t+1} &= \tilde{x}_t + (a - 1)u_t + \tau_\epsilon \epsilon_{t+1} \\u_{t+1} &= a\tilde{x}_t + \tau_\epsilon \epsilon_{t+1}\end{aligned}$$

The state space equations are

$$\begin{aligned}u_{t+1} &= au_t + \tau_\epsilon \epsilon_{t+1} \\v_{t+1} &= u_{t+1} + \tau_\eta \eta_{t+1}\end{aligned}$$

The Riccati equation is

$$\begin{aligned}(1 - a^2)\hat{\sigma}_{t+1|t}^2 + \frac{a^2\hat{\sigma}_{t+1|t}^4}{\hat{\sigma}_{t+1|t}^2 + \tau_\eta^2} &= \tau_\epsilon^2 \\ \Rightarrow \hat{\sigma}_{t+1|t}^2 &= \frac{1}{2} \left[(a^2 - 1)\tau_\eta^2 + \tau_\epsilon^2 + \sqrt{(a^2 - 1)\tau_\eta^4 + \tau_\epsilon^4 + 2(a^2 + 1)\tau_\eta^2\tau_\epsilon^2} \right] \\ &= \frac{1}{2} \left[(a^2 - 1)\tau_\eta^2 + \tau_\epsilon^2 + \sqrt{[(a^2 - 1)\tau_\eta^2 + \tau_\epsilon^2]^2 + 4\tau_\eta^2\tau_\epsilon^2} \right] \\ &= \frac{1}{2}\tau_\epsilon^2 \left[(a^2 - 1)\kappa^2 + 1 \right] \left[1 + \sqrt{1 + \left(\frac{2\kappa}{(a^2 - 1)\kappa^2 + 1} \right)^2} \right] \\ K &= \frac{\hat{\sigma}_{t+1|t}^2}{\hat{\sigma}_{t+1|t}^2 + \tau_\eta^2} \\ \hat{u}_{t+1|t} &= (1 - K)\hat{u}_{t|t-1} + Kv_t\end{aligned}$$

Now replace u, v using Equations (2.34) and (2.35):

$$\Rightarrow \hat{x}_{t+1|t} = (1 - K)\hat{x}_{t|t-1} + Ky_t$$

For $a = 1$ the formula is identical to that of Example 1; and it is straightforward to verify that $\hat{\sigma}_{t+1|t}^2$ is decreasing in a , and consequently also K is decreasing in a . There are two insights to be drawn from this:

1. The EWMA is still an optimal estimator for a mean-reverting model of volatility.
2. In the presence of mean reversion, K decreases, everything else being equal. We discount the past *less*, because mean-reversion causes volatility to be more concentrated. When the volatility is changing less from period to period, past observations become more informative.

Example 3 (Harvey and Shephard, 1996): The generating process for gross returns $R_t = P_t/P_{t-1}$ is assumed to be

$$R_t = e^{\beta + \exp(h_t/2)\xi_t}$$

where β is a known constant, and $\xi_t \sim N(0, 1)$. Define $u_t = \log R_t - \beta$. Then $u_t = \exp(h_t/2)\xi_t$. Square u_t and take the logarithm to linearize the equation:

$$\begin{aligned} \log u_t^2 &= h_t + \log \xi_t^2 \\ &= h_t + \eta_t + \kappa \end{aligned}$$

where $\kappa := E(\log \xi_t^2) \simeq -1.2703$, and η_t is a zero-mean random variable. Define

$$\begin{aligned} y_t &= \log u_t^2 - \kappa \\ &= 2 \log |\log R_t - \beta| - \kappa \end{aligned}$$

so that we get an observation equation:

$$y_t = h_t + \eta_t$$

Now, we posit an evolution equation for h_t :

$$h_{t+1} = b + ah_t + \epsilon_t$$

This is the same model as AR(1), from which we obtain an estimate \hat{h}_t . If $\beta = 1$, then the formulas take a simple form: $u_t \simeq r_t$ and the volatility estimate is given by

$$\hat{\sigma}_t \simeq \exp \left[\sum_{s=0}^{\infty} (1 - K)^{-s} (\log |\log R_t - 1| - \kappa) \right]$$

2.6 Exercises

Exercise 2.1. (15) Prove that $\left(\prod_{t=1}^T (1 + r(t)) \right)^{1/T} - 1 \leq T^{-1} \sum_{t=1}^T r(t)$.

Exercise 2.2. (20) Provide an example of two random variables that are uncorrelated but dependent.

Exercise 2.3. (25) Provide a second example, employing an entirely different rationale for the lack of correlation from the first one.

Exercise 2.4. (30) Let X, Y be two random variables taking values in \mathbb{R}_+ . Show that $\text{cor}(X^2, Y^2)$ if and only if $\text{cor}(X, Y) > 0$.

Exercise 2.5. (15) Derive the formula for $E(h_\infty^2)$ from Equation (2.10).

Exercise 2.6. (10) Prove that if $E(h_\infty^2)$ is finite, i.e. $\alpha_1 + \beta_1 < 1$, then a stationary distribution exists, i.e. $E[\log(\beta_1 + \alpha_1 \epsilon_0^2)] < 0$. (Hint: use Jensen's inequality)

The Takeaways

1. Security returns exhibit heavy tails, low autocorrelation but high autocorrelation in absolute value or square, and they are approaching log-normal returns over longer time intervals.
2. GARCH models capture most properties of returns, and can be used to estimate volatility.
3. The GARCH volatility estimates are exponential weighted averages of non-iid squared returns.
4. State-space models can be used to model a variety of
5. They will appear in factor models.

Chapter 3

Linear Models of Returns: The Basics

The Questions

1. What do we intend for linear models of returns?
2. How do we interpret them?
3. What are their applications?

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

Linear models of asset returns are a cornerstone of this book. They are flexible, interpretable, perform well in applications, and are supported by theory. Furthermore, they fit like a glove with mean-variance optimization and can be also used as a basis for a number of important tasks, like risk management and performance analysis. It is possible that you, the reader, will find this class of models inadequate in some way, at some point. But just because you have outgrown them, does not mean that you will find them useless. They will still enable you to reason about the entire investment process, and some of the theory will come in handy.

3.1 Factor Models

We saw in Chapter 2 how to model univariate returns. A direct extension to multivariate returns would be to model each security's return as an independent process. This would not be adequate, however, because returns are dependent. It is a natural step to model the common dependency among stock as being generated by a few common sources of randomness, and then to keep a random source of per-security randomness that is independent of the factors. The model for stock returns is called a *factor model* and takes the form:

$$\mathbf{r}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \boldsymbol{\epsilon}_t \quad (3.1)$$

where:

- $t \in \mathbb{N}$ denotes time;
- $\boldsymbol{\alpha}$ is an n -dimensional vector;
- \mathbf{r}_t is a random vector of n asset returns minus the risk-free rate (see Section 2.1.2);
- \mathbf{f}_t is the random vector of m factor returns;
- \mathbf{B} is a $n \times m$ loadings matrix;
- $\boldsymbol{\epsilon}_t$ is the random vector of n idiosyncratic (or specific) returns.

If the random vector $\boldsymbol{\epsilon}_t$ had a generic distribution, we would have gained nothing in tractability. Instead, we assume that i) they be independent from the factor returns; ii) have expected value equal to zero; iii) the covariance matrix $\text{var}(\boldsymbol{\epsilon}_t)$ be diagonal, or at least sparse in some sense. Often, models with a diagonal covariance matrix are called *strict* and models with a sparse covariance matrix are called *approximate*. The vector $\boldsymbol{\epsilon}_t$ is the *idiosyncratic component* of returns.

We usually refer to the term $\mathbf{B}\mathbf{f}_t$ as the *systematic component of asset returns*. We assume that the pair $(\mathbf{f}_t, \boldsymbol{\epsilon}_t)$ be identically distributed across periods or at least with a slowly-varying distribution, and that \mathbf{f}_t and $\boldsymbol{\epsilon}_t$ be independent for each t . I denote covariance matrices of \mathbf{f}_t and $\boldsymbol{\epsilon}_t$ with $\boldsymbol{\Omega}_{\mathbf{f}} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{n \times n}$ respectively. With this notation, the covariance matrix of assets is¹

$$\boldsymbol{\Omega}_{\mathbf{r}} = \mathbf{B}\boldsymbol{\Omega}_{\mathbf{f}}\mathbf{B}' + \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \quad (3.2)$$

FAQ 3.1: *Why is the covariance matrix $\Omega_r = \mathbf{B}\Omega_f\mathbf{B}' + \Omega_\epsilon$?*

The covariance matrix Ω_r does not depend on the intercept α , and the terms $\mathbf{B}f_t$ and ϵ_t are independent, so that $\Omega_r = \text{cov}(\mathbf{B}f_t) + \Omega_\epsilon$. The factor term is a linear transformation of f_t . For any random vector ξ with covariance Ω_ξ , $\text{cov}(\mathbf{B}\xi) = \mathbf{B}\Omega_\xi\mathbf{B}'$, because $[\text{cov}(\mathbf{B}\xi)]_{i,j} = \text{cov}(\sum_k [\mathbf{B}]_{i,k}\xi_k, \sum_\ell [\mathbf{B}]_{j,\ell}\xi_\ell) = \sum_{k,\ell} [\mathbf{B}]_{i,k}[\Omega_\xi]_{k,\ell}[\mathbf{B}]_{j,\ell} = [\mathbf{B}\Omega_\xi\mathbf{B}']_{i,j}$.

This decomposition is at the core of volatility modeling with linear returns and the subject of Chapters 8 and 9.

In this chapter we set aside the very important issue of estimating the parameters of Equation (3.1) from data, and focus instead on its usage and interpretation. Here is the plan for the next few sections. First, we review the *interpretations* of Equation (3.1), of which there are three:

1. as a graphical model;
2. as the superposition of low-dimensional cross-sectional return vectors;
3. as the overlap of the factor return vector with the asset loadings vector.

Secondly, we review the *transformations* that can be operated on factor models. There are three of those too:

1. Rotations;
2. Projections;
3. Push-outs.

These mathematical operations are versatile tools in the hand of the quantitative manager to reformulate, simplify or extend a model.

Lastly, we describe the uses of factor models. There are quite a few:

1. *Forecast* and *Decompose* volatility, so that we can separate wanted vs unwanted risk;
2. Be a fundamental input to *Portfolio Construction*;

¹If this formula is not familiar to you, derive it, maybe using a single-factor model first.

3. Understand performance and separate skill from luck;
4. Serve as a foundation for *alpha research*.

3.2 Interpretations of Factor Models

style	country				industry			
0.12	-1.91	...	0 0 1 ...	1 0 0 ...	0 1 0 0 ...	0 0 0 1 ...	1 0 0 0
-2.39	-3.00	...						
-1.52						
...						
...						

Figure 3.1: A typical loadings matrix. The style loadings comprise an “intercept” factor (sometimes termed “country” factor). The loadings are all ones, and the intercept factor contribution to total returns is the same for all assets. The other style loadings often are standardized. The country and industry loadings take values equal to 1 if the asset belongs to the country or industry.

Before we start interpreting, let us make factor model more concrete with an example. In Figure 3.1 you see a “typical” loading matrix used in a risk model. A few columns contain *style* loadings². Other columns consisting of dummy variables indicating whether the stock belongs to a particular industry. There may be a column for “energy explorers and producers”, a column for “biotechnology company”, and so on. A stock will have a “1” loading if it belongs to the industry, 0, otherwise. Finally there are columns consisting of dummy variables denoting country classification, similarly to industry. When the factor return for a country or an industry is high, it will move all the stocks in the

²The use of the term “style” will be clear later, when we associate to loadings investment styles.

industry. The factor structure captures comovement among stocks with certain obvious commonalities, as well as less obvious ones.

Even though Equation (3.1) is older than modern Statistics (having really originated in the unpublished work of Gauss), it is surprisingly rich in meaning, and possibly even richer when used in financial applications. First, let's review some interpretations of the equation.

3.2.1 Graphical Model

The first one is as a graphical model. Since $\mathbf{r} - \boldsymbol{\alpha} = \mathbf{B}\mathbf{f}$, for each asset i this equation holds³:

$$E(r_i - \alpha_i | f) = \sum_j [\mathbf{B}]_{i,j} f_j$$

Each of the many asset returns is dependent on all, or some of, the few factor

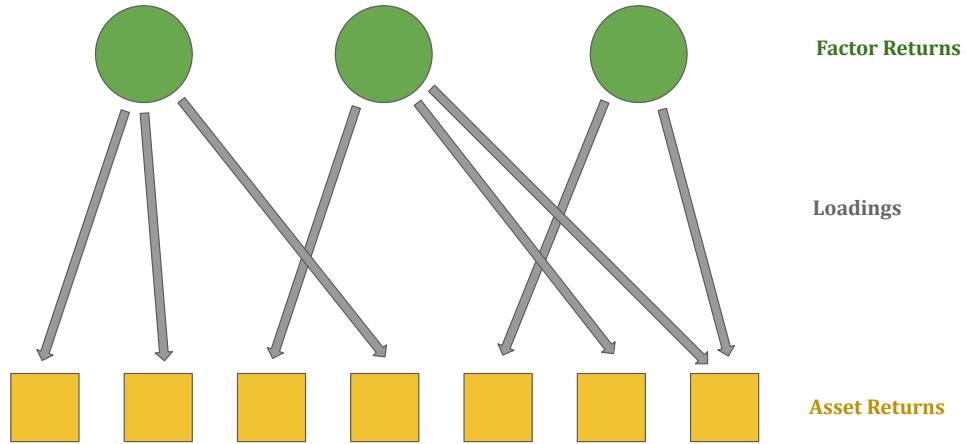


Figure 3.2: Factor models as graphical models.

returns. In a typical regional risk model (say, America, Asia, or Europe) we have up to 10,000 assets and up to 100 factors. In Figure 3.2 we show the relationship visually. A few factors (in green) determine the expected asset returns in excess of $\boldsymbol{\alpha}$ (in yellow), though the link provided by loadings \mathbf{B} (in grey). When the matrix \mathbf{B} is sparse, the corresponding graph is sparse.

³Here, and in the rest of the book, we use these notational conventions: $[\mathbf{B}]_{i,j}$ is element of \mathbf{B} on the i th row and j th column; $[\mathbf{B}]_{i,\cdot}$ and $[\mathbf{B}]_{\cdot,j}$ vectors corresponding to the i th row and j th columns of \mathbf{B} respectively.

3.2.2 Superposition of Effects

The second interpretation is as an overlap of influences on asset returns. a model for the cross-section of returns, i.e., the vector of returns at a given point in time. Let $[\mathbf{B}]_{\cdot,j}$ be the j th column of the matrix \mathbf{B} . We rewrite $E(\mathbf{r} - \boldsymbol{\alpha}) = \mathbf{B}\mathbf{f}$ as

$$E(\mathbf{r} - \boldsymbol{\alpha}|\mathbf{f}) = \sum_j [\mathbf{B}]_{\cdot,j} f_j$$

The expected excess return vector is the overlap of a small number of vectors (the loadings $[\mathbf{B}]_{\cdot,j}$ for a specific factor), weighted by the factor return. This makes clear that the factor component of the cross-section lives in a low-dimensional space. This is shown in Figure 3.3.

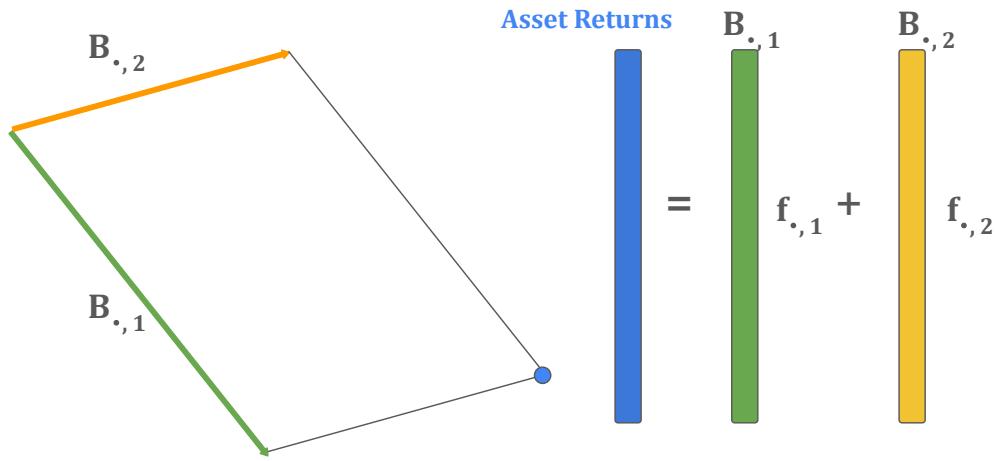


Figure 3.3: Factor models as superposition of weighted factor loadings.

3.2.3 Single-Asset Product

The last interpretation applies to single assets. The expected return of an asset given the factor returns is equal to the scalar product of the asset loadings and the vector of factor returns.

$$E(r_i - \alpha_i|\mathbf{f}) = \langle [\mathbf{B}]_{i,\cdot}, \mathbf{f} \rangle$$

While this formula is rarely used at the asset level, it does show up all the time when we apply it to portfolios. Consider a portfolio $\mathbf{w} \in \mathbb{R}^n$, where w_i is the

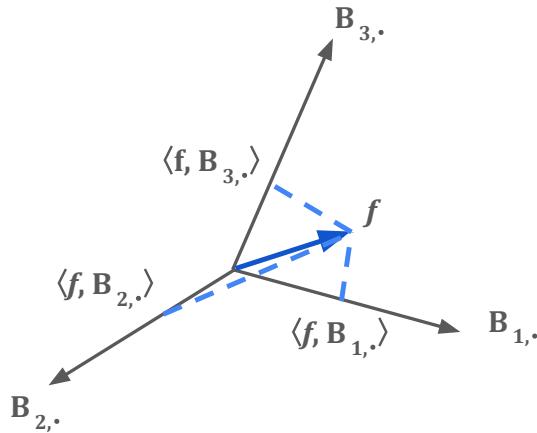


Figure 3.4: Factor models as scalar products of per-stock loadings and factor returns.

net notional value⁴ invested in asset i ; for stocks, this is the stock price times the number of shares held long or short. The expected PnL of the portfolio is

$$\begin{aligned} E(\mathbf{w}'\mathbf{r}|\mathbf{f}) &= E\left(\sum_i w_i r_i \middle| \mathbf{f}\right) \\ &= \sum_i [\alpha_i + \langle [\mathbf{B}]_{i,:}, \mathbf{f} \rangle] w_i \end{aligned}$$

We can explain the factor PnL of a portfolio in terms of a scalar product; within the scalar product identify the largest contributor, the degree of dispersion of PnL among the factors, and so on. This is the jump-off point for *performance attribution*, which we'll cover extensively later in the book.

3.3 Alpha Spanned and Alpha Orthogonal

Consider the factor equation

$$\mathbf{r}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \boldsymbol{\epsilon}_t$$

where \mathbf{f}_t are iid with finite mean and variance, and $\boldsymbol{\epsilon}_t$ are iid (independent on \mathbf{f}_t) with zero unconditional mean and finite variance.

Decompose $\boldsymbol{\alpha}$ as the sum of its projection on the column subspace of \mathbf{B} and the orthogonal complement: $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\alpha}_{\perp}$. By construction it is $\mathbf{B}'\boldsymbol{\alpha}_{\perp} = 0$. So

⁴Net Notional Value is the amount invested in the security in the reference currency (numeraire).

we have $\mathbf{r}_t = \boldsymbol{\alpha}_{\perp} + \mathbf{B}(\boldsymbol{\lambda} + \mathbf{f}_t) + \boldsymbol{\epsilon}_t$. In this relationship, you can see that there is an indeterminacy in the factor model. You can rewrite the model as

$$\mathbf{r}_t = \boldsymbol{\alpha}_{\perp} + \mathbf{B}[\boldsymbol{\lambda} + E(\mathbf{f}_t)] + \mathbf{B}[\mathbf{f}_t - E(\mathbf{f}_t)] + \boldsymbol{\epsilon}_t$$

The “alpha” spanned by the columns of \mathbf{B} is indistinguishable from the expected returns of the factors. However, the term $\boldsymbol{\alpha}_{\parallel} := \mathbf{B}[\boldsymbol{\lambda} + E(\mathbf{f}_t)]$ is independent of how you split the contribution of the two. In the following, we set $\boldsymbol{\lambda} = 0$ and set $\boldsymbol{\mu}_f := E[\mathbf{f}_t]$.

Now, for a little prestige: if you choose as a portfolio $\mathbf{w} = \boldsymbol{\alpha}_{\perp} / \|\boldsymbol{\alpha}_{\perp}\|$, its payoff is

$$\begin{aligned}\mathbf{w}'\mathbf{r}_t &= \frac{1}{\|\boldsymbol{\alpha}_{\perp}\|} \boldsymbol{\alpha}'_{\perp} \mathbf{r}_t \\ &= \|\boldsymbol{\alpha}_{\perp}\| + \frac{\boldsymbol{\alpha}'_{\perp} \boldsymbol{\epsilon}_t}{\|\boldsymbol{\alpha}_{\perp}\|}\end{aligned}$$

The expected return and variance of this portfolio are

$$\begin{aligned}E(\mathbf{w}'\mathbf{r}_t) &= \|\boldsymbol{\alpha}_{\perp}\| \\ \text{var}(\mathbf{w}'\mathbf{r}_t) &= \frac{\boldsymbol{\alpha}'_{\perp} \boldsymbol{\Omega}_{\epsilon} \boldsymbol{\alpha}_{\perp}}{\|\boldsymbol{\alpha}_{\perp}\|^2}\end{aligned}$$

There is an upper bound for the variance, given by the operator norm:

$$\frac{\boldsymbol{\alpha}'_{\perp} \boldsymbol{\Omega}_{\epsilon} \boldsymbol{\alpha}_{\perp}}{\|\boldsymbol{\alpha}_{\perp}\|^2} \leq \|\boldsymbol{\Omega}_{\epsilon}\|_2^2$$

So that

$$\text{SR} \geq \frac{\|\boldsymbol{\alpha}_{\perp}\|}{\|\boldsymbol{\Omega}_{\epsilon}\|_2}$$

In the case of a diagonal matrix, $\|\boldsymbol{\Omega}_{\epsilon}\|_2^2$ is the largest asset idiosyncratic variance. Assume that it is upper bounded for all n . This has interesting implications. Consider the case, for example, where the average absolute orthogonal return per asset is positive: $\sum_i |[\boldsymbol{\alpha}_{\perp}]_i|/n = \mu > 0$ or, equivalently, $\|\boldsymbol{\alpha}_{\perp}\|_1 \geq n\mu$. Now use the simple inequality between 1-norm and Euclidean norm: $\|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2$. Hence $\|\boldsymbol{\alpha}_{\perp}\|_2 \geq \sqrt{n}\mu$. Apply this to the Sharpe Ratio of the alpha orthogonal strategy, and we have a lower bound on the Sharpe Ratio:

$$\text{SR}(\mathbf{w}) \geq \frac{\|\boldsymbol{\alpha}_{\perp}\|}{\|\boldsymbol{\Omega}_{\epsilon}\|_2} \geq \sqrt{n} \frac{\mu}{\|\boldsymbol{\Omega}_{\epsilon}\|_2} \tag{3.3}$$

Let’s summarize the assumptions we made so far, besides the fact that the factor model is correct. We assumed that:

1. the largest idiosyncratic variance⁵ is uniformly bounded in the number of assets.
2. the average absolute value of the coordinate of α_{\perp} is bounded below by μ .

If that is the case, we have a lower bound on the Sharpe Strategy of a portfolio. And if these bounds are uniform for increasing values⁶ of n , then we have a sequence of Sharpe Ratios going to infinity! This is highly unlikely in real life, so one of the assumptions we made – factor model, bound on idio variances, bound on orthogonal expected returns – must not hold. *Under the assumption that the factor model is correct*, it appears that the assumption of finite idio variances holds in practice. This leaves us with the fact that the idio orthogonal expected returns are vanishing in n . Let us summarize this result in concrete terms:

- If a linear model is a good approximation of returns, then alpha is either “spanned” or “orthogonal”.
- “Alpha orthogonal” is extremely valuable: if you have positive expected alpha at the asset level, then your Sharpe increases at the rate \sqrt{n} , a typical rate that arises when you can diversify risk without giving up on returns.
- But Sharpe does not approach infinity in the real world, so you can expect the orthogonal alpha of an asset to be small.
- There are excess returns, but they are more likely to come from “alpha spanned”. This alpha as we will see in the next chapter comes with risk that does not diversify away with a large number of assets.

⁵Or, more generally, the largest eigenvalue of the sparse matrix Ω_r .

⁶Economists reason in terms of “increasing” economies as a function of the tradable assets. I believe there is not much to gain from this level of abstraction, so I keep the exposition and the inequalities in finite dimensions.

3.4 Transformations

3.4.1 Rotations

A factor model is not uniquely identified. Let \mathbf{C} be an $m \times m$ invertible matrix, and define

$$\tilde{\mathbf{B}} = \mathbf{B}\mathbf{C}^{-1}$$

$$\tilde{\mathbf{f}} = \mathbf{C}\mathbf{f}$$

the columns of $\tilde{\mathbf{B}}$ span the same subspace as the columns of B . The model

$$\mathbf{r} = \boldsymbol{\alpha} + \tilde{\mathbf{B}}\tilde{\mathbf{f}} + \boldsymbol{\epsilon}$$

has the same returns as the original model. This is usually termed *rotational indeterminacy* of the risk model.

The covariance matrix of the transformed factors is $\tilde{\Omega}_f = \mathbf{C}\Omega_f\mathbf{C}'$. The factor component of the asset covariance matrix is unchanged under the transformation:

$$\mathbf{B}\mathbf{C}^{-1}\mathbf{C}\Sigma\mathbf{C}'(\mathbf{C}^{-1})'\mathbf{B}' = \mathbf{B}\Omega_f\mathbf{B}'$$

There will be several applications of rotational indeterminacy in the book. Rotations enable us to provide final users with different *views* of the same model. We explore the impact on factor returns of factor transformations for three instructive examples: unit factor covariance, orthonormal loadings and z-scored loadings.

Unit Factor Covariance. Sometimes, users of a model would like to see uncorrelated factors returns with unit variances. Under this perspective, exposures of the portfolio can be interpreted directly as factor volatilities, and factor risk of a portfolio is the sum of the squared exposures, without covariance terms.

This risk model perspective can be obtained by taking the Singular Value Decomposition (SVD) of $\Omega_f = \mathbf{U}\mathbf{S}\mathbf{U}'$ and then setting $\mathbf{C} := \mathbf{S}^{-1/2}\mathbf{U}'$. Then,

$$\tilde{\Omega}_f = \mathbf{C}\Omega_f\mathbf{C}' = \mathbf{S}^{-1/2}\mathbf{U}'\mathbf{U}\mathbf{S}\mathbf{U}'\mathbf{U}\mathbf{S}^{-1/2} = \mathbf{I}$$

Orthonormal loadings. We can also choose a rotation so that the loadings are orthonormal $\tilde{\mathbf{B}}'\tilde{\mathbf{B}} = \mathbf{I}_m$. This means that each column of $\tilde{\mathbf{B}}$ has unit norm (but not unit variance, since the column may have non-zero mean), and is orthogonal to the other. In this case the transformation is $\mathbf{C}^{-1} := \mathbf{V}\mathbf{S}^{-1}$, where \mathbf{V} and \mathbf{S} come from the SVD of $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}'$. We get $\tilde{\mathbf{B}} = \mathbf{U}$.

Z-scored loadings. The z-scoring factor loadings is a common procedure. It consists of a linear rescaling of the loadings of one or more factors, so that the new loadings have zero mean and unit variance⁷. This makes the loadings easier to interpret. Is the stock more exposed than average to the factor, and by how much? What is the average portfolio exposure to the factor on a standardized basis? Is such a linear transformation resulting in an equivalent factor model? It is possible to multiplicate the loadings of factor i by constant κ_i : just consider $\mathbf{C} := \text{diag}(\kappa_1^{-1}, \dots, \kappa_m^{-1})$, which is always invertible. However, in general it is *not* possible to center the loadings (You can try to find a counterexample in Exercise 3.3). However, assume that the unit vector is in the subspace spanned by the loadings columns, i.e., there is a vector \mathbf{V} such that $\mathbf{1} = \mathbf{BV}$. In this case, the centering is possible. If we want to add constants κ_i to the loadings, then⁸ $\tilde{\mathbf{B}} = \mathbf{B} + \mathbf{1}\kappa' = \mathbf{B} + \mathbf{BV}\kappa' = \mathbf{B}(\mathbf{I} + \mathbf{V}\kappa')$, hence our transformation is

$$\mathbf{C}^{-1} = \mathbf{I} + \mathbf{V}\kappa' \quad (3.4)$$

This assumption is verified in two common cases. The first one is the use of a “market” factor, to which all assets are identically exposed. The loadings vector is then $\mathbf{1}$ by construction. The second case is when there are country or industry loadings, such that for each asset the sum of the loadings across industries is exactly one. In this case $\mathbf{1} = \mathbf{BV}$ for a \mathbf{V} with ones corresponding to industry factors, and zero otherwise.

3.4.2 Projections

On occasion, we may want to use a risk model with *fewer* factors compared to the original one. At first glance, this operation may seem unjustified. If we trust that our risk model is the most accurate, why would we want to replace it with a different one? The reasons are many. For example, it may be the case that in practice the loadings of one or more factors are changing so fast as to make portfolio management and hedging difficult. Another reason is that we are using a vendor-provided model, and that we believe that the model is not perfect, i.e., some factors do not quite belong to the model. A third reason may be that we want to give to a final user a “simplified” risk model that is as accurate as possible, while retaining the full model for other uses. For these reasons and more, we face to find a different risk model that is close, in some sense, to the original one.

⁷Like orthonormal loadings, they have unit variance. Unlike orthonormal loadings, they neither have zero mean, nor they are orthogonal.

⁸We use the notation $\mathbf{1} := (1, 1, \dots, 1)$.

We have a model $\mathbf{r} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}$ and associated covariance matrix $\boldsymbol{\Omega}_r = \mathbf{B}\boldsymbol{\Omega}_f\mathbf{B}' + \boldsymbol{\Omega}_\epsilon$, but we want to employ a different exposure matrix \mathbf{A} , in which the range of \mathbf{A} is contained in the range of \mathbf{B} . If we model returns as $\mathbf{r} = \boldsymbol{\alpha} + \mathbf{A}\mathbf{g} + \boldsymbol{\eta}$, the covariance matrix would be $\boldsymbol{\Omega}_r = \mathbf{A}\boldsymbol{\Omega}_g\mathbf{A}' + \boldsymbol{\Omega}_\eta$. What is the value $\boldsymbol{\Omega}_g$ resulting in the best approximation to our original model? Let the distance between the original factor returns \mathbf{f} and the approximate factor returns \mathbf{g} be $\|\mathbf{B}\mathbf{f} - \mathbf{A}\mathbf{g}\|^2$. The distance-minimizing approximate factor returns are $\mathbf{g} = \mathbf{H}\mathbf{f}$, where $\mathbf{H} := (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}$. The corresponding value of $\boldsymbol{\Omega}_g$ is

$$\boldsymbol{\Omega}_g := \mathbf{H}\boldsymbol{\Omega}_f\mathbf{H}'$$

The factor component of asset returns is Bf . We solve

$$\min_{\mathbf{g}} \|\mathbf{B}\mathbf{f} - \mathbf{A}\mathbf{g}\|^2$$

from which $\mathbf{g} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}\mathbf{f} = \mathbf{H}\mathbf{f}$. The corresponding covariance matrix is $\boldsymbol{\Omega}_g = \mathbf{H}\boldsymbol{\Omega}_f\mathbf{H}'$.

We call these transformations projections because, like projections, they are *idempotent*. An idempotent linear operator $\mathbf{\Pi}$ is such that $\mathbf{\Pi}^2\mathbf{x} = \mathbf{\Pi}\mathbf{x}$. The geometric intuition is that, once you have projected a vector on a plane, projecting the resulting vector again on the same plane does not result in another vector, since the input vector is already on the plane. Similarly, consider the “projection” of the model using the matrix \mathbf{A} . The new loadings matrix is now \mathbf{A} . If you project once again using the same matrix \mathbf{A} , the transformation H becomes $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{A}$, i.e., the identity. The transformed covariance matrix is unchanged.

3.4.3 Push-Outs

In the previous two sections we introduced a transformation that preserves the number of factors and a transformation that reduces it. The last section focuses on a transformation that *increases* the number of factors. We therefore lift the loadings matrix into a new one, whose column space contains the column space of the old one. Why could this be of interest? A possible scenario that occurs in practice is that that our factor model may have been developed on historical data that are not representative of the current regime. As a result, the idiosyncratic returns show some structure, in the sense that they themselves are amenable to be formulated as a different factor model:

$$\boldsymbol{\epsilon} = \mathbf{A}\mathbf{g} + \boldsymbol{\eta}$$

with $\mathbf{A} \in \mathbb{R}^{n \times p}$, \mathbf{g} a rv taking values in \mathbb{R}^p and $\boldsymbol{\eta}$ a rv taking values in \mathbb{R}^n . The new model becomes

$$\mathbf{r} = \boldsymbol{\alpha}_{\perp} + \mathbf{B}\mathbf{f} + \mathbf{A}\mathbf{g} + \boldsymbol{\eta} \quad (3.5)$$

With $\boldsymbol{\eta}$ uncorrelated from \mathbf{f}, \mathbf{g} . In the specification of the model (3.5), we require that $\mathbf{A}'\mathbf{B} = 0$. If not, the factor returns of the original model would have to be modified. Assume that $\mathbf{B}'\mathbf{A} \neq 0$. Then we can decompose the columns of \mathbf{A} into the sum of parallel and orthogonal components. In matrix terms, $\mathbf{A} = \mathbf{BC} + \mathbf{H}$, for some $\mathbf{C} \in \mathbb{R}^{m \times p}$. It follows that the model $\mathbf{r} = \mathbf{B}\mathbf{f} + \mathbf{A}\mathbf{g} + \boldsymbol{\eta}$ can be written as $\mathbf{r} = \mathbf{B}(\mathbf{f} + \mathbf{C}\mathbf{g}) + \mathbf{H}\mathbf{g} + \boldsymbol{\eta}$.

$$\begin{aligned} \boldsymbol{\epsilon}'\mathbf{B}\mathbf{f} &= 0 && \text{(original model orthogonality condition)} \\ \boldsymbol{\eta}'\mathbf{A}\mathbf{g} &= 0 && \text{(residual model orthogonality condition)} \\ \boldsymbol{\eta}'(\mathbf{B}\mathbf{f} + \mathbf{A}\mathbf{g}) &= 0 && \text{(final model orthogonality condition)} \end{aligned}$$

From the second and the third equality it follows that $\boldsymbol{\eta}'\mathbf{B}\mathbf{f} = 0$; the first equality can be rewritten as $0 = (\mathbf{g}'\mathbf{A}' + \boldsymbol{\eta}')\mathbf{B}\mathbf{f} = \mathbf{g}'\mathbf{A}'\mathbf{B}\mathbf{f}$ for all realization of \mathbf{f}, \mathbf{g} , hence $\mathbf{A}'\mathbf{B} = 0$. In the estimation chapter we will see how to augment a risk model in a characteristic-model framework.

3.5 Applications

3.5.1 Performance Attribution

What is the PnL of a portfolio at time t ?

$$\begin{aligned} (\text{portfolio PnL}_t) &= \mathbf{w}'_t \mathbf{r}_t \\ &= \mathbf{w}'_t \mathbf{B}\mathbf{f}_t + \mathbf{w}'_t (\boldsymbol{\alpha}_{\perp} + \boldsymbol{\epsilon}_t) \\ &= \mathbf{b}'_t \mathbf{f}_t + \mathbf{w}'_t (\boldsymbol{\alpha}_{\perp} + \boldsymbol{\epsilon}_t) \quad (\mathbf{b}_t := \mathbf{B}'\mathbf{w}_t) \end{aligned}$$

The vector $\mathbf{b}_t \in \mathbb{R}^m$ are the *factor exposures* of the portfolio at time t . The term $[\mathbf{b}_t]_i$ is the sum of the characteristics of factor i of each stock, weighted by the portfolio holdings; keep in mind that the characteristics and the weights can both be negative. The term $\mathbf{b}'_t \mathbf{f}_t$ is the factor PnL in time interval t , while the term is the idiosyncratic PnL. Summing up over a time interval $[1, \dots, T]$, the PnL of a strategy is

$$\text{PnL} = \text{Factor PnL} + \text{Residual PnL}$$

We can also distribute the sum differently:

$$\begin{aligned}
 \text{PnL} &= \sum_{t=1}^T (\text{Factor PnL}_t) + (\text{Residual PnL}_t) \\
 &= \sum_{t=1}^T \sum_{j=1}^m [\mathbf{b}_t]_j [\mathbf{f}_t]_j + \sum_{t=1}^T \sum_{i=1}^n [\mathbf{w}_t]_i (\alpha_{\perp,i} + [\boldsymbol{\epsilon}_t]_i) \\
 &= \sum_{j=1}^m (\text{Factor } j \text{ PnL}) + \sum_{i=1}^n (\text{Stock } i \text{ Residual PnL})
 \end{aligned}$$

And then, of course, one can partition factors and stocks in groups, to highlight, for example the performance arising from style factors, from industry factors, or from a specific group of stocks.

3.5.2 Risk Management: Forecast and Decomposition

If we have a covariance matrix (not specifically from a factor model), the variance of a portfolio \mathbf{w} is easy to compute: $\text{var}(\mathbf{r}'\mathbf{w}) = \sum_{i,j} \text{var}(r_i w_i, r_j w_j) = \sum_{i,j} w_i \text{var}(r_i, r_j) w_j = \mathbf{w}' \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w}$. We can apply this to formula to a covariance matrix associated to a factor model:

$$\begin{aligned}
 \text{var}(\mathbf{r}'\mathbf{w}) &= \mathbf{w}' (\mathbf{B} \boldsymbol{\Omega}_{\mathbf{f}} \mathbf{B}' + \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}) \mathbf{w} \\
 &= \mathbf{b}' \boldsymbol{\Omega}_{\mathbf{f}} \mathbf{b} + \mathbf{w}' \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \mathbf{w}
 \end{aligned}$$

This has two application. The first one is an estimate of a portfolio's *ex ante* volatility at any point in time. This is an essential data for risk managers, since they monitor volatility and allocate risk based on this measure. The second application is the decomposition of variance in factor and idiosyncratic components. Like in the attribution case, the formula is a jumping-off point. For example, a commonly quoted statistic for a strategy is the *percentage of idio variance*, defined as $100 * (\text{dollar idio variance}) / (\text{total variance})$; the percentage of idio variance and factor variance sum to 100, of course. The factor variance can be decomposed further by making factor partitions. The most detailed one has each factor being a singleton, but very common choices are (style group)/(industry group)/(country group), or (subsectors group)/(style group)/(country group). This measure, and the associated sensitivities, are commonly used to monitor strategies. Every partition, either of factors or of assets, induces a covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ where $[\boldsymbol{\Omega}]_{i,j}$ is the covariance between partition group i and j . For example, say that a portfolio has factor exposure \mathbf{b} , and we partition the

factors into groups $1, \dots, p$, with group i containing the subset of elements $\mathcal{S}(i)$. Define $\mathbf{b}_{\mathcal{S}(i)}$ as a vector of factor exposures where all the terms not in $\mathcal{S}(i)$ are set equal to zero. Define $\boldsymbol{\Omega}$ as the covariance matrix of the sets' returns:

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{b}'_{\mathcal{S}(1)} \boldsymbol{\Omega}_f \mathbf{b}_{\mathcal{S}(1)} & \dots & \mathbf{b}'_{\mathcal{S}(1)} \boldsymbol{\Omega}_f \mathbf{b}_{\mathcal{S}(p)} \\ \dots & \dots & \dots \\ \mathbf{b}'_{\mathcal{S}(p)} \boldsymbol{\Omega}_f \mathbf{b}_{\mathcal{S}(1)} & \dots & \mathbf{b}'_{\mathcal{S}(p)} \boldsymbol{\Omega}_f \mathbf{b}_{\mathcal{S}(p)} \end{pmatrix}$$

Then the total factor variance is $v_{TOT}^2 := \mathbf{e}' \boldsymbol{\Omega} \mathbf{e}$, where $\mathbf{e} := (1 \ \dots \ 1)'$. The group's i variance is $v_i^2 := \mathbf{b}'_{\mathcal{S}(i)} \boldsymbol{\Omega}_f \mathbf{b}_{\mathcal{S}(i)}$.

- *Fraction of total variance* for group i is

$$\begin{aligned} p_i &:= \frac{(variance \ of \ group \ i \ + \ half \ of \ covariance \ contributions)}{(portfolio \ variance)} \\ &= \frac{\sum_j [\boldsymbol{\Omega}]_{i,j}}{v_{TOT}^2} \\ &= \frac{\text{cov}(\mathbf{r}_i, \mathbf{r}_{TOT})}{v_{TOT}^2} \\ &= \beta_i \end{aligned}$$

So that $\sum_i p_i = 1$. The percentage of variance of a group $\mathcal{S}(i)$ (again, this includes single factors and single assets—perhaps the most commonly used partition!) is simply the beta of returns of the group to the overall portfolio.

- *Marginal contribution to risk* (MCR) of a group $\mathcal{S}(i)$ is defined as

$$\begin{aligned} m_i &:= \frac{(portfolio \ $vol \ change \ when \ we \ buy \ $1M \ vol \ of \ set \ \mathcal{S}(i))}{\$1M} \\ &= \frac{\partial}{\partial(x_i v_i)} \sqrt{\mathbf{x}' \boldsymbol{\Omega} \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{e}} \\ &= \frac{1}{v_i v_{TOT}} \sum_j [\boldsymbol{\Omega}]_{i,j} \\ &= \rho_i \\ &= \frac{v_{TOT}}{v_i} p_i \end{aligned}$$

- *Sharpe Ratio sensitivity*. It is also useful to compute the sensitivity of the Sharpe Ratio to changes in volatility of a group. The Total Portfolio's

Sharpe Ratio Sensitivity with respect to volatility increase of group i is given by

$$\begin{aligned}\frac{\partial}{\partial x_i} \frac{E(PnL)}{\text{vol}(PnL)} &= \frac{\text{vol}(PnL) \times \partial_i E(PnL) - \partial_i \text{vol}(PnL) \times E(PnL)}{\text{var}(PnL)} \\ &= \frac{\text{vol}(PnL) \times \partial_i E(PnL) - m_i \times \text{SR}_{\text{TOT}} \times \text{vol}(PnL)}{\text{var}(PnL)} \\ &= \frac{\text{SR}_{\text{TOT}}}{v_{\text{TOT}}} \left(\frac{\text{SR}_i}{\text{SR}_{\text{TOT}}} - m_i \right)\end{aligned}$$

The contribution to total Sharpe Ratio is positive if the Sharpe Ratio of a group exceeds a threshold, which is its marginal contribution to risk.

3.5.3 Portfolio Management

Factor models are useful for portfolio management in more than one way. The first one is adjacent to risk management: volatility is the common language spoken by risk managers and portfolio managers, and it is oftentimes generated by a factor model. The second one is the inverse of the covariance matrix also known as *precision matrix*. This matrix plays a central role in portfolio optimization. As discussed in FAQ 3.2, the empirical covariance matrix is usually not invertible. The factor structure makes possible to estimate both Ω_r and Ω_r^{-1} . The third is the model of the expected asset returns as the sum of two terms: α and $\mathbf{B}'E(\mathbf{f})$. These two terms give rise to two qualitatively different classes of expected returns. This makes sense intuitively, since the factor-based returns come with some variability and risk, which is itself captured by the factor covariance matrix Ω_f . The alpha term, instead, comes with apparently no risk. How to manage these sources of returns is the concern of Portfolio Management. Lastly, a factor model is *legible*: when applied to a portfolio, it produces factor exposures, risk and performance decompositions, as discussed above. This makes the job of the portfolio managers easier, since it enables them to plan (before the trade), monitor (during the trade) and understand (after the trade) their strategies.

3.5.4 Alpha Research

As volatility is the *lingua franca* spoken by risk managers and portfolio managers, so alpha is what a signal researcher and a portfolio manager both understand.

FAQ 3.2: Why not the Empirical Covariance Matrix?

Before delving in the details of factor model estimation, we address a preliminary question. Given a time series of returns \mathbf{r}_t with population covariance matrix $\Omega_{\mathbf{r}}$, its simplest estimator is the empirical covariance

$$\hat{\Omega}_{\mathbf{r}} = \frac{1}{T} \sum_{t=1}^T \mathbf{r}_t \mathbf{r}'_t$$

or, if we denote $\mathbf{R} \in \mathbb{R}^{n \times T}$ the matrix of returns where $\mathbf{r}_t = \mathbf{R}_{\cdot,t}$, we can write $\hat{\Omega}_{\mathbf{r}} = T^{-1} \mathbf{R} \mathbf{R}'$. It is well known (and easy to establish) that the estimate maximizes log likelihood (for a normal multivariate distribution), is asymptotically consistent, and a Central Limit Theorem is available for it (Anderson, 1963). Why not use this as our estimate for our covariance matrix? The reason is that, when $T \ll n$, the estimator is very inadequate for volatility estimation and portfolio optimization purposes. The covariance matrix has at most rank T . let $\mathbf{w}_i, i = 1, \dots, T - n$ a basis for the null space of \mathbf{R}' , i.e., $\mathbf{R}' \mathbf{w}_i = 0$. We can interpret these vectors as portfolios. The volatility of portfolio i is $\mathbf{w}'_i \hat{\Omega}_{\mathbf{r}} \mathbf{w}_i = T^{-1} \|\mathbf{R}' \mathbf{w}_i\| = 0$. So, a majority of independent portfolios has zero volatility. The situation is even worse in portfolio optimization. The solution of the mean-variance problem $\max_{\mathbf{w}} [\boldsymbol{\alpha}' - (2\lambda)^{-1} \mathbf{w}' \hat{\Omega}_{\mathbf{r}} \mathbf{w}]$ is $\mathbf{w} = \lambda \hat{\Omega}_{\mathbf{r}}^{-1} \boldsymbol{\alpha}$. In this case, if $\boldsymbol{\alpha}$ is in the null space, the portfolio is undefined. Choosing an alpha close to the null space yields an arbitrarily large portfolio, and an arbitrarily large Sharpe Ratio.

At the cost of excessive generalization, one could say that the signal researcher cares about $\boldsymbol{\alpha}$, the risk manager cares about \mathbf{Bf} and the portfolio manager adds trading costs and tries to combine all these terms into a profitable strategy. In reality, there is “risky” alpha worth trading in \mathbf{Bf} –again, I use the language loosely for the time being, with the goal of tightening it in coming chapters. Furthermore, sometimes these people are one and the same, although the roles are increasingly separated in sufficiently large and complex strategies. Alpha research is improved by factor models in two ways. First, \mathbf{Bf} is important, and for a certain class of investors it is the only thing that matters⁹. For researchers focusing on $\boldsymbol{\alpha}$, the factor-based approach helps separate the two

⁹A marketing term used for this investment style is *smart beta*

source of expected returns: priced factors and unpriced factors.

3.6 Factor Models Types

The fundamental model we use from here on is the factor model of Equation (3.1). We have taken the model for granted. But where the data and parameters of the model come from? In the case of factor models, the answer is especially important, because the meaning attached to the various symbols matters. Practitioners use three broad approaches to identify all the parameters in the equation:

- *Characteristic Model*: this is the most common approach. The input data to the model are the time series \mathbf{r}_t and \mathbf{B}_t . Factor and idiosyncratic returns are estimated from these data. \mathbf{B}_t is a matrix of *asset characteristics*. The intuition is that these characteristics are partially responsible for the stock return. I cover this in Chapter 8.
- *Statistical Model*: in this model, the only primitive is \mathbf{r}_t , and \mathbf{B}_t , \mathbf{f}_t and $\boldsymbol{\epsilon}_t$ are all estimated. I cover this in Chapter 9.
- *Macroeconomic Model*: in this approach, the primitives are \mathbf{r}_t and \mathbf{f}_t , and \mathbf{B}_t and $\boldsymbol{\epsilon}_t$ are estimated. \mathbf{f}_t usually represents a vector of macroeconomic time series.

The relevant methodological issues the modeler must address are:

1. What are the best loss functions to rank a model?
2. Once we have estimates (or primitive data) about factor and idiosyncratic returns, how do we estimate the covariance matrices from cross-sectional estimates?
3. What is the best approach within each framework?
4. How do we select the best model, without optimizing for noise, or data mining?

3.7 Further Reading

Factor models go back to the birth of psychometrics at the turn of the XIX Century. Recent textbooks treatment on the subject are Rencher and Christensen

(2012); Johnson and Wichern (2007). In finance, factor models were first introduced by Sharpe (1964), Sharpe (1965), and Sharpe (1966) for the one-factor case, which was extended to multiple factors by Ross (Ross, 1976). Good introductions to factor models in finance are the survey papers by Connor and Korajczyk (2010), Fan et al. (2016) and the books Connor et al. (2010) and Connor and Korajczyk (2010), MacKinlay (1995).

Graphical models are covered in monographies (Lauritzen, 1996), books on Machine Learning (Bishop, 2006; Murphy, 2012), and survey papers (Models, 2004).

3.8 ★Appendix

3.8.1 Linear Regression

Linear models are by far the most important class of models in Statistics. There are more books on the subject than citizens of the sovereign state of the Vatican¹⁰. In fact, one could argue there is so much material on linear models, that two humans on planet Earth may have completely interpretations of them. In order to have some common ground, I will describe the salient aspects some less-well known aspects which will be needed later. Our setting is as follows. We are given a pair (y, \mathbf{x}) , where y is a random variable taking values in \mathbb{R} and \mathbf{x} is a random vector taking values in \mathbb{R}^m . y and \mathbf{x} ; y and \mathbf{x} are in general dependent random variables: knowing the value of a realization of \mathbf{x} tells us something about the values of y and this makes the problem infinitely interesting. Say that we want to provide a forecast of y , which we denote $\hat{y}(\mathbf{x})$. One way to select such forecast is to try to minimize a loss function; we should pay a price for being wrong. One natural choice of loss is the quadratic loss: it is nonnegative; it is symmetric; it is differentiable; and it penalizes more for large errors. The problem we face is

$$\min_{\hat{y}} E[(\hat{y}(\mathbf{x}) - y)^2 | \mathbf{x}] \quad (3.6)$$

One basic result in statistics¹¹ and in control theory is that, if $E(y^2) < \infty$, the function that minimizes this expectation is the conditional expectation of y

¹⁰Not a joke: as of October 2017, the Vatican has 842 citizens; Amazon lists 1,392 books in the “Probability and Statistics” section with “regression” in their title or subject, the vast majority of them covering linear models.

¹¹Linear regression is an inexhaustible topic. Some useful references are, in order of increasing detail, Wasserman (2004); Hastie et al. (2008); Johnson and Wichern (2007); Harrell (2015); Gelman et al. (2022); Hansen (2022).

given \mathbf{x} . We introduce a new variable ϵ :

$$y = E(y|\mathbf{x}) + \epsilon \quad (3.7)$$

It follows that $E(\epsilon) = E(y) - E(E(y|\mathbf{x})) = E(y) - E(y) = 0$. Then use the chain the following chain:

$$\begin{aligned} E[(\hat{y}(\mathbf{x}) - y)^2|\mathbf{x}] &= E[(\hat{y}(\mathbf{x}) - E(y|\mathbf{x}) + E(y|\mathbf{x}) - y)^2|\mathbf{x}] \\ &= E[\epsilon^2|\mathbf{x}] + E[(\hat{y} - E(y|\mathbf{x}))^2|\mathbf{x}] - 2E[\epsilon|\mathbf{x}](y - E(y|\mathbf{x})) \\ &= E[\epsilon^2|\mathbf{x}] + E[(\hat{y} - E(y|\mathbf{x}))^2|\mathbf{x}] \\ &\geq E[\epsilon^2|\mathbf{x}] \end{aligned}$$

The equality holds only if $\hat{y} = E(y|\mathbf{x})$. The term $E[\epsilon^2|\mathbf{x}]$ is finite, because

$$\begin{aligned} E[\epsilon^2|\mathbf{x}] &\leq 2E(y^2) + 2E[E(y|\mathbf{x})^2] \\ &\leq 2E(y^2) + 2E[E(y^2|\mathbf{x})] \quad (\text{Jensen}) \\ &= 4E(y^2) \quad (\text{Iterated Expectation}) \\ &< \infty \end{aligned}$$

In applications, we have n samples (y_i, \mathbf{x}_i) and we choose a functional form for $\hat{y} = g(\mathbf{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a finite- or infinite-dimensional vector. We then minimize the *empirical* squared loss $n^{-1} \sum_i (y_i - g(\mathbf{x}, \boldsymbol{\theta}))^2$. The simplest form of g is linear: $g(\mathbf{x}, \boldsymbol{\beta}) = \sum_i \beta_i x_i$. In matrix form, Equation (3.7) becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3.8)$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times m}$, $\boldsymbol{\beta} \in \mathbb{R}^m$. n are the observations, and m are the “features”. we want to estimate the parameters $\boldsymbol{\beta}$, and estimates for $\mathbf{X}\boldsymbol{\beta}$. We then minimize the empirical loss

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (3.9)$$

which is equal to the unweighted sum of squared errors (*Ordinary Least Squares*, or OLS) $(y_i - \sum_j [\mathbf{X}]_{i,j} \beta_j)^2$. A different way to arrive at the same problem is to posit that the true model is Equation (3.8), and to further assume that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$. If we fix $\boldsymbol{\beta}$, we have $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$; and since we know the distribution of $\boldsymbol{\epsilon}$, we can associate to a choice of $\boldsymbol{\beta}$ a *likelihood* $f(\boldsymbol{\epsilon}|\boldsymbol{\beta})$. If we choose the parameter $\boldsymbol{\beta}$ to maximize the likelihood, we end up solving the same problem as Equation (3.9). The choice of maximizing the likelihood is called *the Maximum Likelihood Principle*¹².

¹²For a detailed discussion of the MLP, see Robert (2007).

Finally there is a geometrical interpretation for the regression problem. You can interpret the set $S := \{\mathbf{X}\boldsymbol{\beta} | \boldsymbol{\beta} \in \mathbb{R}^m\}$ as a subspace of \mathbb{R}^n . The columns of \mathbf{X} are a (generally non-orthonormal) basis of the subspace. We are then given a point $\mathbf{y} \in \mathbb{R}^n$ and find the point $\hat{\mathbf{y}} \in S$ that is closest to \mathbf{y} . This is the definition of a projection of \mathbf{y} on S . The projection is a linear operator. The minimum¹³ is attained at

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3.10)$$

and the estimates $\hat{\mathbf{y}} := E(\mathbf{y}|\boldsymbol{\beta})$ are

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \quad (3.11)$$

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the *hat matrix* or *projection matrix*. The estimated residuals are

$$\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Intuitively, the optimal estimates should not change if we change the base of the subspace. To see this rigorously, transform \mathbf{X} into \mathbf{XQ} , where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is non-singular. The transformed set of predictors spans the same subspace as \mathbf{X} . Then

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{XQ}((\mathbf{XQ}')\mathbf{XQ})^{-1}(\mathbf{XQ}')\mathbf{y} \\ &= \mathbf{XQ}(\mathbf{Q}'\mathbf{X}'\mathbf{xQ})^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{XQQ}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{Q}')^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \quad (3.12)$$

hence \mathbf{y} is independent of base representation.

Another property of the estimate $\hat{\mathbf{y}}$ is that, if we iterate the estimation process on the estimate $\hat{\mathbf{y}}$, we obtain again $\hat{\mathbf{y}}$. This also has geometric interpretation. Once a point has been projected on a hyperplane, the projection of the projection is unchanged. In algebraic terms, $\mathbf{H}\hat{\mathbf{y}} = \mathbf{H}^2\mathbf{y} = \mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$.

Here is another facet of linear regression tying geometric and algebraic interpretations of linear regression. Decompose \mathbf{x} using the SVD: $\mathbf{x} = \mathbf{U}\Lambda\mathbf{V}'$.

¹³The minimum is unique if the rank of \mathbf{X} is m , i.e., if all the columns of \mathbf{X} are linearly independent. In Chapter 8 we will encounter cases of rank-deficient matrices.

\mathbf{U} is an orthonormal basis for the column subspace of \mathbf{x} . Then

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{U}\Lambda\mathbf{V}'(\mathbf{V}\Lambda\mathbf{U}'\mathbf{U}\Lambda\mathbf{V}')^{-1}\mathbf{V}\Lambda\mathbf{U}'\mathbf{y} \\ &= \mathbf{U}\mathbf{U}'\mathbf{y}\end{aligned}$$

So \mathbf{y} is projected on the column space of \mathbf{U} .

Replace Equation (3.8) in beta estimation formula (3.10) to obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.\end{aligned}$$

The estimate of beta is unbiased, because $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] = 0$; and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is The standard deviations of the estimates are

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (3.13)$$

Similarly,

$$\text{var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

We can write these formulas using the SVD:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{V}\Lambda^{-2}\mathbf{V}' \quad (3.14)$$

$$\text{var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{U}\mathbf{U}' \quad (3.15)$$

The variance of the estimates $\text{var}(\hat{\boldsymbol{\beta}})$ becomes larger as the columns of X become more collinear. In our interpretation of the matrix X , this occurs when we include factors that overlap heavily with pre-existing factors.

The estimation formulas extend directly to the case of *heteroskedastic noise*. In this case we assume that $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Omega}_\epsilon)$, where $\boldsymbol{\Omega}_\epsilon$ is a positive definite symmetric matrix. The estimates for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{y}}$ and $\text{var}(\hat{\boldsymbol{\beta}})$ can be derived directly from the previous formulas, by left-multiplying by $\boldsymbol{\Omega}_\epsilon^{-1/2}$ both sides of Equation (3.8):

$$\boldsymbol{\Omega}_\epsilon^{-1/2}\mathbf{y} = \boldsymbol{\Omega}_\epsilon^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}_\epsilon^{-1/2}\boldsymbol{\epsilon}$$

Notice that the $\boldsymbol{\Omega}_\epsilon^{-1/2}\boldsymbol{\epsilon}$ is distributed according to a standard normal (exercise), so that the noise is homoskedastic; and we apply the OLS results to obtain *Weighted Least Squares* (WLS) formulas:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{y} \quad (3.16)$$

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{X})^{-1} \quad (3.17)$$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{y} \quad (3.18)$$

3.8.2 Linear Regression Decomposition

Split Equation (3.8) into two parts:

$$\mathbf{y} = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \quad (3.19)$$

where we have partitioned the predictors $\mathbf{x} = (\mathbf{x}_1|\mathbf{x}_2)$. Equation (3.10) can be rewritten by using block submatrices for $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$, and the formula for the inverse of submatrices, in order to obtain $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$. It can be shown that the coefficient $\hat{\boldsymbol{\beta}}_2$ can be estimated by a two-step process. First, regress the columns of \mathbf{y}_2 on \mathbf{x}_2 : $\mathbf{x}_2 = \mathbf{x}_1\boldsymbol{\gamma} + \mathbf{u}$, where $\mathbf{x}_2 - \mathbf{x}_1\boldsymbol{\gamma} \perp \mathbf{u}$. Second, regress y on \mathbf{u} : $y = \mathbf{u}\boldsymbol{\beta}_3 + \mathbf{v}$. The least-squared coefficient of this regression is the same as $\hat{\boldsymbol{\beta}}_2$, i.e. $\hat{\boldsymbol{\beta}}_3 = \hat{\boldsymbol{\beta}}_2$. The proof can be found in Hansen (2022).

Exercise 3.1. If a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ has near collinear columns, then there is a vector u such that $\|Xu\|^2 = h$ for some small positive h .

1. Show that $u'(X'X)u = h$.
2. Show that $\min_i \lambda_i^2 \leq h$.
3. From this, show that $\|var(\hat{\boldsymbol{\beta}})\| = \max_i \lambda_i^{-2} \geq 1/h$.

3.8.3 The Frisch-Waugh-Lovell Theorem

Suppose that you estimate a model in which we have two separate groups of independent variables. In matrix form, the model is

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{Z}] \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix} + \boldsymbol{\eta} \quad (3.20)$$

We could use the Equations (3.10) and (3.12) to estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}$.

However, you could follow a different approach. The idea is to remove from the columns of \mathbf{Z} the terms collinear to the columns of \mathbf{x} . We decompose $\mathbf{Z} = \mathbf{X}\mathbf{a} + \tilde{\mathbf{Z}}$; the columns of $\tilde{\mathbf{Z}}$ are regression residuals and therefore orthogonal to $\mathbf{x}'\tilde{\mathbf{Z}} = \mathbf{0}$. The subspace spanned by $(\mathbf{X}|\tilde{\mathbf{Z}})$ is the same as the subspace spanned by $(\mathbf{X}|\tilde{\mathbf{Z}})$ (if you do not see it, prove it). Therefore, the estimates $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$ are unchanged. However, this transformation enables us to perform regressions in stages, without having to reestimate the entire model. This is formalized in the Frish-Waugh-Lovell Theorem. It states that the $\hat{\mathbf{y}}$ forecast from Equation (3.20) is the same as $\mathbf{X}\hat{\boldsymbol{\beta}} + \tilde{\mathbf{Z}}\hat{\boldsymbol{\delta}}$ from Procedure 3.1. Correspondingly,

Procedure 3.1: *Stagewise Linear Regression*

1. Estimate the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \quad (3.21)$$

to obtain estimates $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\epsilon}}$.

2. Regress the columns of \mathbf{Z} on \mathbf{X} and take the residuals of each regression. Define $\tilde{\mathbf{Z}}$ as a matrix whose i th column is the residual vector of $[\mathbf{Z}]_i$ on \mathbf{X} .
3. Estimate the model $\hat{\boldsymbol{\epsilon}} = \tilde{\mathbf{Z}}\boldsymbol{\beta}_2 + \boldsymbol{\xi}$

the estimated residuals in the two procedures, $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\xi}}$, are identical. In addition if we regress the model

$$\mathbf{y} = [\mathbf{X}|\tilde{\mathbf{Z}}] \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{bmatrix} + \tilde{\boldsymbol{\eta}} \quad (3.22)$$

then $\hat{\boldsymbol{\delta}}_i = \hat{\boldsymbol{\beta}}_i$, with $i = 1, 2$.

To prove the statement, first write the regression residuals:

$$\tilde{\mathbf{Z}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Z} = (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}})\mathbf{Z} . \quad (3.23)$$

The operator $\mathbf{I}_n - \mathbf{H}_{\mathbf{X}}$ is termed the *annihilator* of \mathbf{X} . Because $\mathbf{H}_{\mathbf{X}}$ is a projection, so is $\mathbf{I}_n - \mathbf{H}_{\mathbf{X}}$: $(\mathbf{I}_n - \mathbf{H}_{\mathbf{X}})^2 = \mathbf{I}_n - 2\mathbf{H}_{\mathbf{X}} + \mathbf{H}_{\mathbf{X}}^2 = \mathbf{I}_n + \mathbf{H}_{\mathbf{X}}$. Therefore, it is $\tilde{\mathbf{Z}} = (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}})^2\mathbf{Z}$. Because the column vectors of $\tilde{\mathbf{Z}}$ are orthogonal to those of \mathbf{X} , it follows that $\mathbf{X}'\tilde{\mathbf{Z}} = \mathbf{0}$. Finally, $\hat{\boldsymbol{\epsilon}} = (\mathbf{I}_n - \mathbf{H}_{\mathbf{X}})\mathbf{y}$. The matrices $(\mathbf{X}|\tilde{\mathbf{Z}}) = (\mathbf{x}|\mathbf{Z})$ are in the following relationship:

$$(\mathbf{x}|\tilde{\mathbf{Z}}) = (\mathbf{x}|\mathbf{Z}) \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}} \end{pmatrix} \quad (3.24)$$

Now, we estimate the parameters of model (3.22):

$$\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix} = (\mathbf{x} | \tilde{\mathbf{Z}}) \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\tilde{\mathbf{Z}} \\ \tilde{\mathbf{Z}}'\mathbf{X} & \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}} \end{pmatrix}^{-1} \begin{bmatrix} \mathbf{X}' \\ \tilde{\mathbf{Z}}' \end{bmatrix} \mathbf{y} \quad (3.25)$$

$$= (\mathbf{X} | \tilde{\mathbf{Z}}) \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{y} \end{bmatrix} \quad (3.26)$$

$$\hat{\delta}_1 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3.27)$$

$$\hat{\delta}_2 = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\hat{\epsilon} \quad (3.28)$$

So that $\hat{\delta}_i = \hat{\beta}_i$, with $i = 1, 2$. Figure 3.5 illustrates the steps of the theorem.

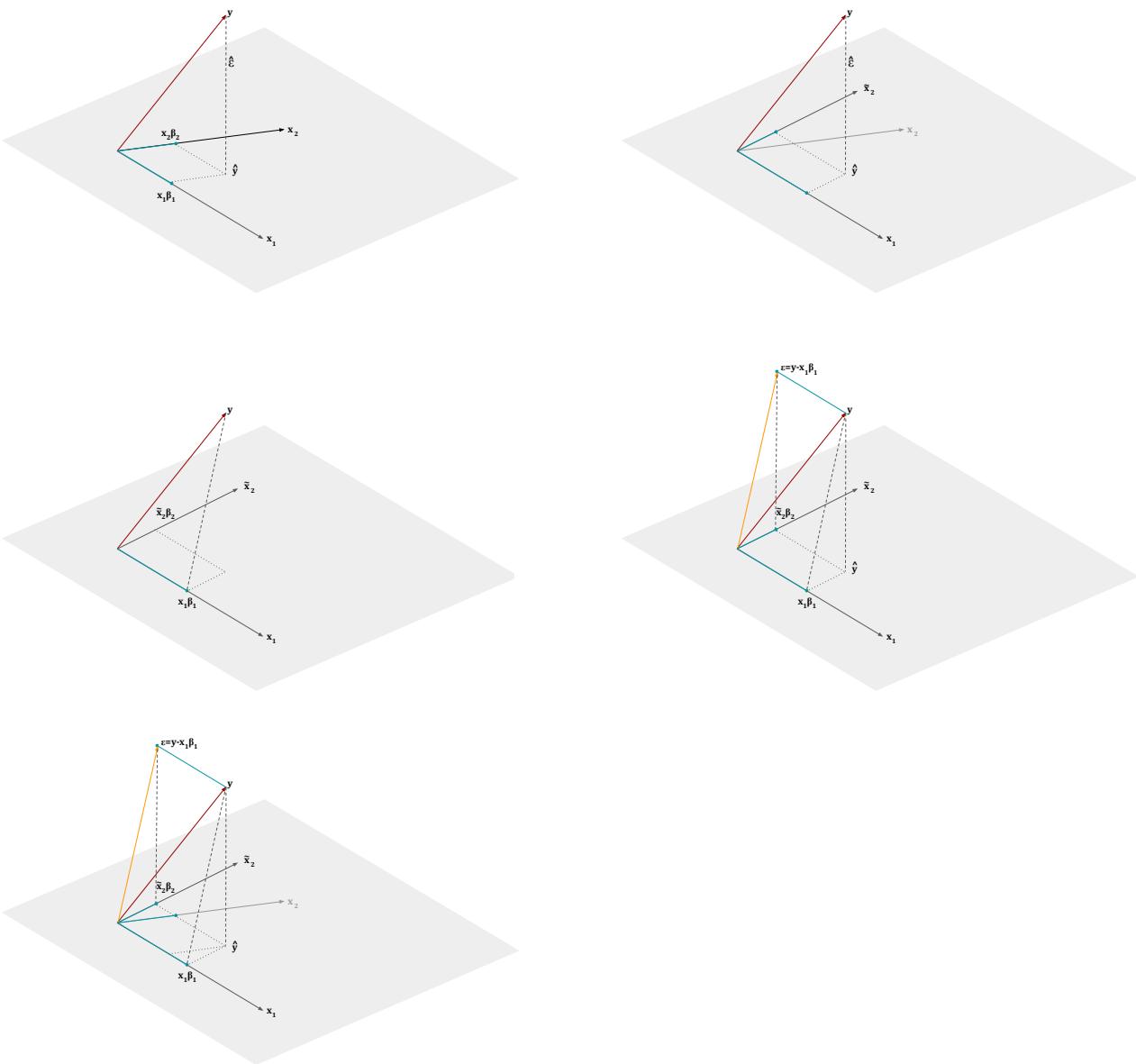


Figure 3.5: Frisch-Vaugh-Lovell theorem procedure. From top left: Original regression ($\mathbf{X}_1, \mathbf{X}_2$). Regression performed in a orthonormalized basis ($\mathbf{X}_1, \tilde{\mathbf{X}}_2$). Univariate regression on \mathbf{X}_1 . Regression of the residual from previous regression, performed on $\tilde{\mathbf{X}}_2$). Lastly the combined univariate regressions yield the same result as the two bivariate regressions above.

3.8.4 The Singular Value Decomposition

The Singular Value Decomposition (SVD) is a fundamental factorization in numerical linear algebra. It powers many numerical computations, as [Golub and Van Loan \(2012\)](#) beautifully explains. In addition, it is extremely insightful in theoretical analysis. Much of this book relies on it. Since it is not always

covered in linear algebra courses, this appendix provides a crash course on the subject. For gentler introductions, see [Trefethen and Bau \(1997\)](#), [Horn and Johnson \(2012\)](#), [Strang \(2019\)](#), and the aforementioned classic book by Golub and Van Loan.

We start by recalling a basic fact of algebra. We are given a square matrix \mathbf{A} that is symmetric and positive semidefinite ($\mathbf{x}\mathbf{Ax}' \geq 0$ for all \mathbf{x}). Let λ_i, \mathbf{v}_i be the i th eigenvalue and eigenvector of \mathbf{A} , i.e., $\mathbf{Av}_i = \lambda_i \mathbf{v}_i$, where \mathbf{v}_i are unit-norm vectors. Then, the eigenvalues are real, positive, and the eigenvectors are orthonormal $\mathbf{v}_i' \mathbf{v}_j = \delta_{i,j}$. What can be said about generic rectangular matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$? A possible generalization is to relax the condition that \mathbf{v}_i appear on both sides of the eigenvalue equation. We posit an equation of the form: $\mathbf{Av}_i = s_i \mathbf{U}_i$, with $\mathbf{v}_i \in \mathbb{R}^n$ and $\mathbf{U}_i \in \mathbb{R}^m$. However, we keep the requirement that \mathbf{v}_i be orthonormal and similarly for \mathbf{U}_i . Let $r \leq \min\{m, n\}$ be the norm of \mathbf{A} . The image subspace of \mathbf{A} has dimension r : there are r independent vectors \mathbf{x}_i such $\mathbf{Ax}_i \neq 0$. The kernel subspace has dimension $n - r$: there are $m - r$ independent vectors \mathbf{y}_i such that $\mathbf{Ay}_i = 0$. We partition \mathbf{v}_i in image and kernel vectors:

$$\mathbf{Av}_i = s_i \mathbf{U}_i \quad 1 \leq i \leq r \quad (3.29)$$

$$\mathbf{Av}_i = 0 \quad r < i \leq m \quad (3.30)$$

We can write these equations in matrix form:

$$\mathbf{A} (\mathbf{v}_1 | \dots | \mathbf{v}_n) = (\mathbf{U}_1 | \dots | \mathbf{U}_m) \begin{pmatrix} s_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & s_2 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & s_r & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \end{pmatrix} \quad (3.31)$$

Here, in addition to the vectors $\mathbf{U}_1, \dots, \mathbf{U}_r$, we have completed this orthonormal basis with $\mathbf{U}_{r+1}, \dots, \mathbf{U}_m$ so that it spans \mathbb{R}^m . In compact form, Equation (3.31) can be written as $\mathbf{AV} = \mathbf{US}$, where \mathbf{U}, \mathbf{V} are orthonormal matrices, i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}_m$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_n$; and $\mathbf{S} \in \mathbb{R}^{m \times n}$ may have non-zero elements only on the main diagonal. Finally, we rewrite the equation after right-multiplying by \mathbf{V}' as

$$\mathbf{A} = \mathbf{USV}' \quad (3.32)$$

We show the decomposition visually in Figure 3.6.

$$\begin{array}{c} \boxed{} \\ \mathbf{A} \end{array} = \begin{array}{c} \boxed{} \\ \mathbf{U} \end{array} \times \begin{array}{c} \boxed{} \\ \mathbf{S} \end{array} \times \begin{array}{c} \boxed{} \\ \mathbf{V}' \end{array}$$

\times

Figure 3.6: Singular Value Decomposition, full form.

We prove Equations (3.29), (3.30) by noting that $\mathbf{A}'\mathbf{A}$ is a symmetric positive semidefinite matrix of rank r , so that there are r pairs $(\mathbf{v}_i, \lambda_i)$ satisfying $(\mathbf{A}'\mathbf{A})\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Define $s_i := \sqrt{\lambda_i}$ and $\mathbf{U}_i := (\mathbf{A}\mathbf{v}_i)/s_i$. These satisfy Equation (3.29). We prove that the \mathbf{U}_i are orthonormal:

$$\mathbf{U}'_i \mathbf{U}_j = \frac{\mathbf{v}'_i (\mathbf{A}' \mathbf{A} \mathbf{v}_j)}{s_i s_j} = \frac{s_j}{s_i} \mathbf{v}_i \mathbf{v}_j = \delta_{i,j} \quad (3.33)$$

because $\mathbf{v}_1, \dots, \mathbf{v}_r$ are orthonormal. Now we complete the basis in \mathbb{R}^n by adding orthonormal vectors $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$, where $\mathbf{A}\mathbf{v}_i = 0$. These make a basis for the nullspace of \mathbf{A} . Correspondingly, we add orthonormal vectors $\mathbf{U}_{r+1}, \dots, \mathbf{U}_m$ to complete the basis in \mathbb{R}^m . A few observations (among many) on the SVD:

1. If all the singular values are distinct, the first r columns of \mathbf{U} and \mathbf{V} are uniquely determined. However, they are not in the case of identical singular values.
2. Equation (3.32) can be rewritten as

$$\mathbf{A} = \sum_{i=1}^r s_i \mathbf{U}_i \mathbf{v}'_i \quad (3.34)$$

The SVD decomposes a matrix into a sum of rank-one matrices.

3. For all $i \leq r$,

$$\mathbf{A}' \mathbf{A} \mathbf{v}_i = s_i^2 \mathbf{v}_i \quad (3.35)$$

$$\mathbf{A} \mathbf{A}' \mathbf{U}_i = \mathbf{A} (\mathbf{A}' \mathbf{A} \mathbf{v}_i)/s_i = s_i \mathbf{A} \mathbf{v}_i = s_i^2 \mathbf{U}_i \quad (3.36)$$

In other terms, $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ have the same eigenvalues.

4. The SVD decomposes the operations on an element in \mathbb{R}^n into a rotation, a rescaling of the axes turning a ball into an ellipsoid, followed by another rotation. The net result is that any operator \mathbf{A} maps a point on a ball into a point on a rotated ellipsoid. Figure 3.7 illustrates the steps of the SVD.

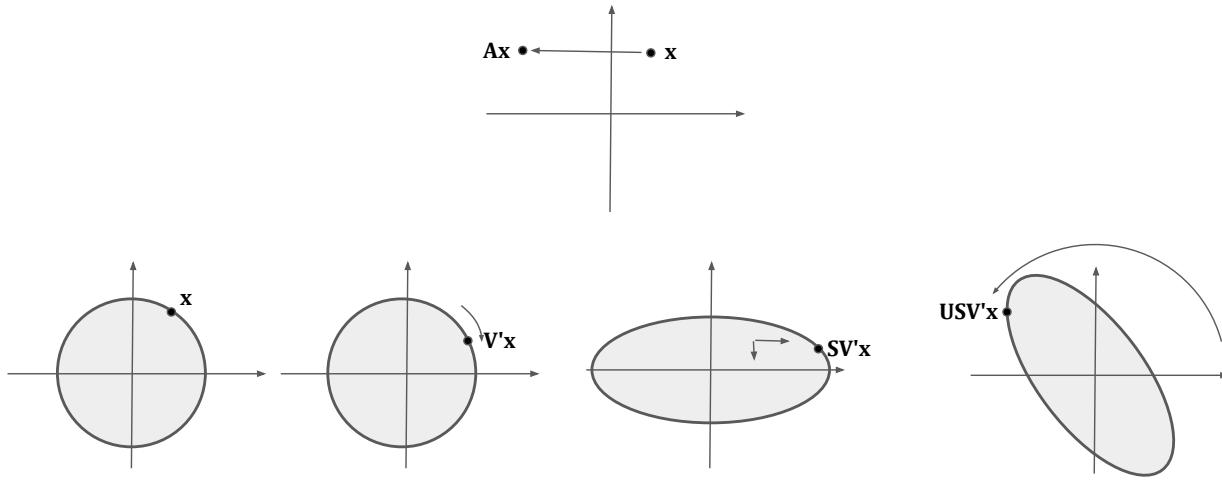


Figure 3.7: Singular Value Decomposition as a sequence of steps: rotation, scaling, rotation.

3.9 Exercises

Exercise 3.2 (Portfolio Covariances).

1. (5) Prove that, if $R - e$ is Gaussian with covariance matrix Ω_r , then the net return portfolio w has variance $\mathbf{w}'\Omega_r\mathbf{w}$.
2. (10) Generalize this result. Let x be a random vector taking values in \mathbb{R}^n with covariance matrix Ω . Let A be an $m \times n$ matrix. Prove that the covariance matrix of the random vector \mathbf{Ax} is $\mathbf{A}\Omega\mathbf{A}'$.
3. (10) Say that a random vector \mathbf{x} follows a multivariate normal distribution with covariance matrix Ω . Let the Singular Value Decomposition of Ω be $\mathbf{U}\Lambda\mathbf{U}'$, and define

$$\Omega^{1/2} = \mathbf{U} \begin{pmatrix} \lambda_1^{1/2} & 0 & \dots & 0 \\ 0 & \lambda_2^{1/2} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & \lambda_n^{1/2} \end{pmatrix} \mathbf{U}'$$

Let ξ a gaussian distribution with unit covariance matrix. Prove that $\Omega^{1/2}\xi$ has covariance Ω .

Exercise 3.3. (27) Provide a counterexample in which it is not possible to center the loadings with a rotation. Hint: use a one-factor model.

Exercise 3.4. (30) Find conditions under which matrix (3.4) is invertible.

Exercise 3.5 (Factor Model of log returns). In the original formulation of our factor model we have considered a linear model of returns. Assume instead that we have a linear model for the vector of log returns. Denote this vector \tilde{r} . The net return of asset i is $r_{i,t} = \exp(\tilde{r}_{i,t}) - 1$. Denote the covariance matrices of log returns and of returns $\tilde{\Omega}_r$, Ω_r respectively.

1. (10) Prove that the covariance matrix of returns is such that

$$[\Omega_r]_{i,j} = \exp \left[\frac{1}{2} [\tilde{\Omega}_r]_{i,i} + \frac{1}{2} [\tilde{\Omega}_r]_{j,j} \right] \exp \left[[\tilde{\Omega}_r]_{i,j} - 1 \right] \quad (3.37)$$

2. (5) Show that the first-order approximation for the exact equation above is

$$[\Omega_r]_{i,j} \simeq [\tilde{\Omega}_r]_{i,j} \left[1 + \frac{1}{2} \left([\tilde{\Omega}_r]_{i,i} + [\tilde{\Omega}_r]_{j,j} + [\tilde{\Omega}_r]_{i,j} \right) \right] \quad (3.38)$$

Prove that $[\Omega_r]_{i,j} \geq [\tilde{\Omega}_r]_{i,j}$.

Exercise 3.6 (Excess Returns). (25) It is commonly assumed that among the investable assets there is an asset that has return r_{free} with probability 1. r_{free} changes over time, but is known in each period. In the academic literature the standard factor model (3.1) models the excess returns, defined as $(r_t)_i - r_{free}$, as per Section 2.1.2 On the other side, practitioners think in terms of returns, not excess returns.

- When in portfolio management is incorrect to reason in term of excess returns? When it is not?
- Show that a model of excess returns could be recasts as a model of returns by adding a factor.
- Can you extend the modeling to incorporate sensitivities to interest rates?

The Takeaways

1. Factor Models are intuitive and tractable.
 - a) Intuitive, because they model additive sources of returns, each one with interpretable characteristics;
 - b) Tractable, because linear models can scale in complexity and breadth of covered securities.
2. They can be used for risk estimation, portfolio construction, signal research, and performance attribution. You can:
 - a) Separate risk into factor and idiosyncratic, and manage them according to your investment philosophy;
 - b) Optimize a portfolio having full visibility into the type of risk you want to tolerate;
 - c) Understand profit and loss in terms of factor and idiosyncratic terms.
3. They are flexible, i.e., they can be modified, simplified, or extended to suit a specific need.
 - a) Simplify a model, by projecting it on a lower-dimensional space;
 - b) Transform a model, by changing its factor representation;
 - c) Extend a model, by adding factors.

Chapter 4

Portfolio Management: The Basics

The Questions

1. What are the basic single-period portfolio optimization formulations?
How do we interpret the results?
2. What can go wrong?
3. How do we incorporate prior information in portfolio optimization?

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

This chapter is devoted to the very basics of portfolio construction. The common theme throughout the chapter is that we limit ourselves to a single-period optimization setting. This a chapter for hedgehogs, not foxes: we set a narrow playing field, but dig a deep hole. The chapter requires some knowledge of basic results optimization Please consult the Appendix, Section 4.6.1, or re-open any optimization textbook you enjoyed reading as an undergraduate or MS student. I give references on this subject, and other standard topics like utility theory, in the “Further Reading” Section at the end of the chapter.

4.1 Why Mean-Variance Optimization?

Investors have objectives, information, and constraints. Besides this generic statement, there is not much in common among them. A large fraction of investment professionals cannot – and would not – articulate a clear objective function; their constraints are sometimes *ad hoc*, vague, or inconsistently enforced. Neither George Soros nor Warren Buffett, nor others among the most successful investors in history, have ever known what the volatility of their portfolios was at any point in time. At the other extreme, academics have developed several normative theories for portfolio construction. In this book I am trying to use relevance to applications as a guiding principle. In the vast majority of applications, the optimization formulations are single-period. This is explainable by a combination of the following¹:

- *Interpretability.* Multi-period optimization problems are vastly more complex to formulate and, once solved, their solutions are also harder to interpret.
- *Computational tractability.* Single-period optimization problems are solvable by commercial solvers in a matter of seconds.
- *Short-term investment horizon.* Investors think only about the short term, partly because they heavily discount the future, partly because they do not know how to quantify information uncertainty and rate of change.
- *Lack of standard theoretical results.* Few strong results are available characterizing the formulations and the improvement in performance.

The objective function V is a function of the portfolio weights \mathbf{w} and of the returns \mathbf{r} . Economic theory interprets V as a utility function, taking different values under different realizations of the future. The expected value of the utility function gives the investor the *ex ante* value of the bet she would be taking by investing in a portfolio. We assume that the investor has initial wealth W_0 , that she knows the distribution of the random vector \mathbf{r} , and that she solves the

¹On justifications of the mean-variance approach to portfolio optimization, see [Cochrane \(2005\)](#); [Huang and Litzenberger \(1988\)](#). Both cover the standard cases of exponential and quadratic utilities. A number of textbooks exist covering portfolio construction. A classic is [Grinold and Kahn \(1999\)](#); see also [Chincarini and Kim \(2022\)](#); [Qian et al. \(2007\)](#); [Isichenko \(2021\)](#). On the statistics of the Sharpe Ratio, see [Lo \(2002\)](#) and, for a comprehensive and definitive reference, [Pav \(2023\)](#) and references therein.

problem

$$\max E [V(W_0 + \mathbf{w}' \mathbf{r})] \quad (4.1)$$

The choice of V is not obvious. Common properties of V are that it is be monotonically increasing (more wealth is better than less) and concave (corresponding to risk aversion). One approach, followed by [Markowitz \(1959\)](#), is to consider a polynomial local approximation of the objective function: $V(W_0 + \mathbf{w}' \mathbf{r}) \simeq V(W_0) + V'(W_0)\mathbf{w}' \mathbf{r} + V''(W_0)(\mathbf{w}' \mathbf{r})^2/2$. Taking expectations, we obtain

$$\begin{aligned} E [V(W_0 + \mathbf{w}' \mathbf{r})] &\simeq V(W_0) + V'(W_0)\mathbf{w}' \boldsymbol{\alpha} + \\ &\quad \frac{V''(W_0)}{2} (\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} + (\mathbf{w}' \boldsymbol{\alpha})^2) \\ &\simeq V(W_0) + V'(W_0)\mathbf{w}' \boldsymbol{\alpha} + \\ &\quad \frac{V''(W_0)}{2} \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \end{aligned}$$

We maximize a quadratic objective function which is the weighted sum of expected return and variance; hence the name *mean-variance portfolio optimization* ([De Finetti, 1940](#); [Markowitz, 1952](#)):

$$\begin{aligned} \frac{E [V(W_0 + \mathbf{w}' \mathbf{r})] - V(W_0)}{V'(W_0)} &\simeq \mathbf{w}' \boldsymbol{\alpha} - \frac{\rho}{2} \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \\ \rho &:= -\frac{V''(W_0)}{V'(W_0)} \end{aligned}$$

$\rho > 0$ is called the *coefficient of absolute risk aversion* (CARA). The higher ρ , the more risk-averse the investor is.

As examples, consider an objective function of the form $V(x) = -\exp(-ax)$. The CARA for this function is constant $\rho = a$: it is independent of the wealth W_0 of the investor, and so are her allocation decisions. The optimization problem is

$$\max_{\mathbf{w}} \mathbf{w}' \boldsymbol{\alpha} - \frac{a}{2} \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w}$$

Alternatively, consider the objective function $V(x) = \log(x)$. This function is associated to the *Kelly criterion* for investing. It has unique properties which warrant a dedicated section to it in this chapter. Here, let us consider its

implications for approximate portfolio optimization. The CARA is $\rho = 1/W_0$, so that we solve

$$\max_{\mathbf{w}} \mathbf{w}'\boldsymbol{\alpha} - \frac{1}{2W_0}\mathbf{w}'\boldsymbol{\Omega}_r\mathbf{w}$$

The wealthier the investor is, the more risk-seeking she becomes.

We have shown that a quadratic utility function implies a mean-variance optimization problem for the investor. This result is standard. Less known is the converse: if an investor selects an investment on the basis of mean and variance only, her utility function is necessarily quadratic (Baron, 1977; Johnstone and Lindley, 2011). Viewed in the context of axiomatic decision theory, portfolio mean-variance optimization is not satisfactory, because a quadratic utility implies that investors are satiated, and have even a dislike of wealth beyond a certain threshold. As a local approximation, however, mean-variance optimization is appropriate. Moments of returns (and portfolios) beyond the second one are beyond the realm of what's possible, as seen in Chapter 2. A portfolio manager settled a long discussion on the topic with the laconic statement that “the first two moments should be enough for everybody”.

4.2 Mean-Variance Optimal Portfolios

A factor model gives us an asset-asset covariance matrix $\boldsymbol{\Omega}_r \in \mathbb{R}^{n \times n}$. Given this information, it is straightforward to compute the variance of a portfolio, as we saw in Section 3.5.2 on risk decomposition. The other essential input to the optimization problem is a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ of expected returns, over the same interval at which we have a volatility forecast. The simplest optimization problem is to maximize expected PnL, subject to a constraint on the maximum tolerable volatility, denoted by $\sigma > 0$. The problem can be stated as

$$\begin{aligned} \max & \boldsymbol{\alpha}'\mathbf{w} \\ \text{s.t. } & \mathbf{w}'\boldsymbol{\Omega}_r\mathbf{w} \leq \sigma^2 \end{aligned} \tag{4.2}$$

This is not only the simplest optimization problem. One of the most important metrics used for the evaluation of strategies is the *Sharpe Ratio*. Most readers have at least a superficial acquaintance with it; and it will be covered extensively in Chapter 7. For now, it suffices to say that the Sharpe Ratio of a portfolio is the ratio of its expected return to its predicted volatility at a certain time horizon, and therefore serves as a risk-adjusted measure of performance. If we

have covariance matrix and expected returns, we can formulate the Sharpe Ratio optimization thus:

$$\max_{\mathbf{w}} \frac{\boldsymbol{\alpha}' \mathbf{w}}{\sqrt{\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w}}}$$

This optimization, however, is indefinite because the objective function $\text{SR}(\mathbf{w})$ is independent of the portfolio size, i.e., homogeneous of degree 0: $\text{SR}(t\mathbf{w}) = \text{SR}(\mathbf{w})$ for all $t > 0$. We can address this issue by bounding the denominator. This means fixing the portfolio size. The upper bound constraint on the denominator is always binding:

$$\begin{aligned} & \max_{\mathbf{w}} \frac{\boldsymbol{\alpha}' \mathbf{w}}{\sqrt{\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w}}} \\ & \text{s.t. } \sqrt{\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w}} \leq \sigma \\ & \text{equivalent to } \max_{\mathbf{w}} \frac{\boldsymbol{\alpha}' \mathbf{w}}{\sigma} \\ & \text{s.t. } \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \leq \sigma^2 \\ & \text{equivalent to } \max_{\mathbf{w}} \boldsymbol{\alpha}' \mathbf{w} \\ & \text{s.t. } \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \leq \sigma^2 \end{aligned}$$

Which is optimization problem (4.2). The First-Order Necessary Conditions (FONC) for this problem are:

$$\begin{aligned} \nabla_{\mathbf{w}} (\boldsymbol{\alpha}' \mathbf{w} - \lambda \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w}) &= \boldsymbol{\alpha} + 2\lambda \boldsymbol{\Omega}_r \mathbf{w} \\ &= 0 \\ \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} &\leq \sigma^2 \\ \lambda &\geq 0 \\ \lambda (\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} - \sigma^2) &= 0 \end{aligned}$$

The solution to these equations is

$$\mathbf{w}^* = \frac{\sigma}{\sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}}} \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha} \quad (4.3)$$

$$\lambda^* = \frac{\sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}}}{2\sigma} \quad (4.4)$$

The expected return and the Sharpe Ratio of the portfolio are

$$\begin{aligned} E(\mathbf{r}'\mathbf{w}^*) &= \sigma \sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}} \\ SR^* &= \sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}} \end{aligned} \quad (4.5)$$

A way to interpret (and derive quickly) the solution is to recall that the optimal portfolio is proportional to $\boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}$, and then to find the proportionality factor so that the variance constraint is met. The optimal portfolio is proportional to the volatility budget: the higher the budget, the bigger the portfolio. However, the portfolio is independent of the magnitude of the alpha vector (it is homogeneous of degree zero in alpha): replacing $\boldsymbol{\alpha}$ with $\kappa \boldsymbol{\alpha}$ gives the same solution. This is interesting. The parameter λ^* also merits special consideration. It is

Insight 4.1: Miscalibration of alpha size is not catastrophic

If you have a volatility constraint, a good volatility model, and your *relative* alphas are accurate then the error in the *absolute* size of the alphas does not matter.

the *shadow price* (or Lagrange multiplier) of the volatility constraint. If we increase the variance budget by one unit, the expected return increases by λ^* . In other terms, the shadow price of the variance constraint is the derivative of the objective function with respect to the variance. While this relationship is not very useful in this specific case, it will come in handy for other constraints.

In its simplicity, the solution contains the essential data of the problem: the inverse of the covariance matrix (also called the *precision matrix*) and the vector of expected return. In the next few pages I would like to interpret, eviscerate, extend this simple formula; and finally, as you start believing it is useful, to caution you against its use. Like all the good things in life, mean-variance optimization is at its most pleasant when it is accompanied by precautionary measures. First of all, we can derive the same solution when we solve an unconstrained problem:

$$\max \boldsymbol{\alpha}' \mathbf{w} - \lambda \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \quad (4.6)$$

We have added the constraint to the objective function in the form of a penalty term; the informal term for this operation is *pricing out* the constraint. The

objective function is convex, and the solution is given by

$$\mathbf{w}^* = \frac{1}{2\lambda} \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}$$

which gives the same solution as the vol-constrained problem when

$$\lambda = \frac{\sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}}}{2\sigma}$$

The larger the volatility budget, the smaller the penalty coefficient.

Notice that this penalty value is the same as the shadow price in the previous formulation. This is not a coincidence. We obtain the same solution when we price out the constraint and we give the a unit price equal to the shadow price of that constraint.

A third equivalent formulation is the one where we minimize volatility, subject to a return constraint:

$$\min \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \tag{4.7}$$

$$\text{s.t. } \boldsymbol{\alpha}' \mathbf{w} \geq \mu \tag{4.8}$$

The solution is

$$\mathbf{w}^* = \frac{\mu}{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}} \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}$$

There is yet another formulation that is equivalent to the previous ones. Oftentimes, we think of portfolio positions not in terms on net market value, but of volatility. We do not invest \$10M in AAPL. The annualized volatility of AAPL is 20%, and therefore we have a \$2M volatility position in the stock. This conveys the position size in terms of its range of dollar movement over the course of a year. Now, we can express the Sharpe-Optimal portfolio in terms of volatility in the following way. Let the stock volatilities be $\sigma_1, \dots, \sigma_n$, and define \mathbf{V} a diagonal matrix with these volatilities on the main diagonal. Denote the asset correlation matrix with \mathbf{C} . Then, the covariance matrix can be written $\boldsymbol{\Omega}_r = \mathbf{VCV}$. Now let's rewrite the solution to the MVO problem:

$$\begin{aligned} \mathbf{w}^* &= \frac{1}{2\lambda} (\mathbf{VCV})^{-1} \boldsymbol{\alpha} \\ \mathbf{V}\mathbf{w}^* &= \frac{1}{2\lambda} \mathbf{C}^{-1} (\mathbf{V}^{-1} \boldsymbol{\alpha}) \end{aligned} \tag{4.9}$$

$$\begin{aligned} \mathbf{v}^* &= \frac{1}{2\lambda} \mathbf{C}^{-1} \mathbf{s} \\ \text{SR}^* &= \sqrt{\mathbf{s}' \mathbf{C}^{-1} \mathbf{s}} \end{aligned} \tag{4.10}$$

In the formula above, $\mathbf{v}^* := \mathbf{V}\mathbf{w}^*$ is the vector of dollar volatilities, and \mathbf{s} is the vector of asset-level Sharpe Ratios. Therefore, the optimal dollar volatilities are proportional to the Sharpe Ratios, multiplied by the inverse of the correlation matrix. This is interesting, because dollar vols, correlations, and asset Sharpe Ratios are more intuitive quantities than covariances and returns. A direct implication of this result is that, when assets are uncorrelated, the optimal dollar vol allocation is proportional to the asset Sharpe Ratios.

Insight 4.2: Reading the entries of the precision matrix \star

Is there a way to interpret further the relationship $\mathbf{w}^* \propto \Omega_{\mathbf{r}}^{-1} \boldsymbol{\alpha}$? The optimal position of asset i is a weighted sum of alphas. The $\Omega_{\mathbf{r}}^{-1}]_{i,j}$ are proportional to minus the *partial correlations* of the returns of i and j after controlling for the other asset returns. The interpretation of partial correlation is that it captures collinearity between two random variables, after removing the collinearity of these variables with a set of controlling variables. In practice, one follows this procedure: 1. regress the returns of asset i and j on the returns of the other assets; 2. compute the correlation between the residuals from the two regressions, which we denote $\rho_{i,j}$. The formula for the optimal portfolio is

$$w_i \propto [\Omega_{\mathbf{r}}^{-1}]_{i,i} \left(\alpha_i - \sum_{j \neq i} \rho_{i,j} [\Omega_{\mathbf{r}}^{-1}]_{j,j} \alpha_j \right)$$

The diagonal terms of the precision matrix are always positive. The interpretation of this rather convoluted formula is that, whenever the returns of two assets are positively correlated after removing the joint effect of correlations with other variables, the size of the portfolio is reduced, because the collinearity makes the alpha common to both asset i and j .

4.3 Trading in Factor Space

We have a factor model, and we estimate the expected factor returns $\boldsymbol{\lambda}$. Say that we want to generate a portfolio which has the closest possible return to one of the factors. For example, we want to generate a “momentum factor” portfolio. What would it be? The returns of the *factor-mimicking portfolio* (FMP) should be as close as possible to those of the portfolio: the variance of the difference of the two returns should be minimized. A portfolio \mathbf{w} has an associated factor exposure \mathbf{b} . Its returns are $\mathbf{r}'\mathbf{w} = \mathbf{b}'\mathbf{f} + \mathbf{w}'\boldsymbol{\epsilon}$. The tracking variance² between f_i and $\mathbf{r}'\mathbf{w}$ is $E[((b_i - 1)f_i + \sum_{j \neq i} b_j f_j + \mathbf{w}'\boldsymbol{\epsilon})^2]$. This is minimized when $b_i - 1 = 0$, $b_j = 0$ for $j \neq i$, and the portfolio’s idio variance is minimized. The optimization formulation is

$$\begin{aligned} & \min \mathbf{w}'\boldsymbol{\Omega}_\epsilon\mathbf{w} \\ & \text{s.t. } \mathbf{B}'\mathbf{w} = \mathbf{e}_i \end{aligned}$$

The solution is $\mathbf{v}_i = \boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B}(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{e}_i$. The matrix whose column vectors are the factor-mimicking portfolios is (\mathbf{P} is for “portfolios”)

$$\begin{aligned} \mathbf{P} &:= (\mathbf{v}_1 | \dots | \mathbf{v}_m) \\ &= \boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B}(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}(\mathbf{e}_1 | \dots | \mathbf{e}_m) \\ &= \boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B}(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1} \end{aligned} \tag{4.11}$$

We now have factor portfolios as tradable instruments. The expected return of a factor portfolio is $(\boldsymbol{\alpha}'_\perp + \boldsymbol{\lambda}'\mathbf{B}')\mathbf{v}_i = \lambda_i$. If the factor portfolios are sufficiently diversified, they have a low idio variance compared to their factor variance. As we will see in Chapter 8, this has an interpretation in the estimation of a model. If we can ignore the idio variance of the portfolios, then investing in factor portfolios becomes a simpler problem, where the covariance matrix of these m synthetic assets is $\boldsymbol{\Omega}_f$, their expected return is $\boldsymbol{\lambda}$, and the MVO allocation is proportional to $\boldsymbol{\Omega}_f^{-1}\boldsymbol{\lambda}$, while the Sharpe Ratio is $\sqrt{\boldsymbol{\lambda}'\boldsymbol{\Omega}_f^{-1}\boldsymbol{\lambda}}$.

4.4 Trading in Idio Space

In Section 3.3 we introduced the concepts of alpha spanned and alpha orthogonal. Alpha spanned are asset expected excess returns attributable to non-zero factor expected returns; alpha orthogonal are non explainable by factor returns.

²We ignore the term $\boldsymbol{\alpha}'_\perp$, both out of simplicity and because it is very small.

Because of Equation (3.3), Sharpe Ratio scales at least like \sqrt{n} . Because of this, alpha orthogonal is the golden currency in investing. How does one build a portfolio that exploits only this alpha? By requiring that the optimal portfolio have no exposures, and therefore no returns, to factors. There are many ways to achieve this, and they are encountered in practice. The first one is to build a portfolio that has a maximum volatility, maximizes expected returns, and has no factor exposures.

Exercise 4.1 (MVO idio portfolio).

1. (10) Solve a MVO problem where the exposures are set to zero.

$$\begin{aligned} & \max \boldsymbol{\alpha}'_{\perp} \mathbf{w} \\ \text{s.t. } & \mathbf{B}' \mathbf{w} = 0 \\ & \mathbf{w}' \boldsymbol{\Omega}_{\epsilon} \mathbf{w} \leq \sigma^2 \end{aligned} \tag{4.12}$$

2. (5) Compare this solution to the “heuristic” solution proportional to alpha orthogonal: $\mathbf{w}^* = \sigma^2 \boldsymbol{\alpha}_{\perp} / \sqrt{\boldsymbol{\alpha}'_{\perp} \boldsymbol{\Omega}_{\epsilon} \boldsymbol{\alpha}_{\perp}}$. How they compare? When are they the same?

The solution of which is left as an exercise. We could also choose a simpler portfolio, proportional to $\boldsymbol{\alpha}_{\perp}$.

4.4.1 Drivers of Information Ratio: Information Coefficient and Diversification

What makes a good strategy? Before we are trading (i.e., *ex ante*), we are living the dream, i.e. Equation (4.5):

$$\text{SR}^* = \sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}}$$

A substantial part of this and of the next chapter is dedicated to the notion that we do not exist in the Dreamtime. Our forecasted returns and risk models are incorrect, and we should take this knowledge in the investment process. A first step is to establish some *ex post* relationship for the Sharpe Ratio. Start with the solution to the MVO problem, Equation (4.3):

$$\mathbf{w}^* = \frac{\sigma}{\sqrt{\boldsymbol{\alpha}' \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}}} \boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}$$

And assume that the covariance matrix Ω_r is accurate; admittedly a strong assumption. The expected *realized* return is

$$\begin{aligned} E(\mathbf{r}' \mathbf{w}^*) &= \frac{\sigma}{\sqrt{\boldsymbol{\alpha}' \Omega_r^{-1} \boldsymbol{\alpha}}} E(\mathbf{r}' \Omega_r^{-1} \boldsymbol{\alpha}) \\ &= \sigma \frac{E(\mathbf{r}' \Omega_r^{-1} \boldsymbol{\alpha})}{\sqrt{\boldsymbol{\alpha}' \Omega_r^{-1} \boldsymbol{\alpha}} \sqrt{E(\mathbf{r}' \Omega_r^{-1} \mathbf{r})}} \sqrt{\frac{E(\mathbf{r}' \Omega_r^{-1} \mathbf{r})}{n}} \sqrt{n} \end{aligned}$$

Define the *Information Coefficient*:

$$\text{IC} := \frac{E(\mathbf{r}' \Omega_r^{-1} \boldsymbol{\alpha})}{\sqrt{\boldsymbol{\alpha}' \Omega_r^{-1} \boldsymbol{\alpha}} \sqrt{E(\mathbf{r}' \Omega_r^{-1} \mathbf{r})}}$$

The important thing to know is that the Information Coefficient is a correlation. To see why, we need to transform variables:

$$\begin{aligned} \tilde{\mathbf{r}} &= \Omega_r^{-1/2} \mathbf{r} \\ \hat{\boldsymbol{\alpha}} &= \Omega_r^{-1/2} \boldsymbol{\alpha} \end{aligned} \tag{4.13}$$

So that the Information Coefficient can be rewritten in a more succinct form

$$\text{IC}(\hat{\boldsymbol{\alpha}}, \tilde{\mathbf{r}}) := \frac{E(\tilde{\mathbf{r}}' \hat{\boldsymbol{\alpha}})}{\sqrt{\hat{\boldsymbol{\alpha}}' \hat{\boldsymbol{\alpha}}} \sqrt{E(\tilde{\mathbf{r}}' \tilde{\mathbf{r}})}}$$

which can be interpreted as a cross-sectional uncentered correlation.

We can simplify things further by proving that $E(\mathbf{r}' \Omega_r^{-1} \mathbf{r}) = n$. The random vector \mathbf{r} has the same covariance matrix³ as $\Omega_r^{1/2} \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a standard multivariate normal.

$$\begin{aligned} E(\mathbf{r}' \Omega_r^{-1} \mathbf{r}) &= E(\boldsymbol{\xi}' \Omega_r^{1/2} \Omega_r^{-1} \Omega_r^{1/2} \boldsymbol{\xi}) \\ &= \sum_{i=1}^n E(\xi_i^2) \\ &= n \end{aligned}$$

Putting everything together, the $E(\mathbf{r}' \mathbf{w}^*)/\sigma$ is

$$\text{SR} = \text{IC}(\hat{\boldsymbol{\alpha}}, \tilde{\mathbf{r}}, \Omega_r) \sqrt{n}$$

Insight 4.3: Information Coefficient and Predictive Regression

Being a correlation, the IC is also naturally related to the predictive strength of our alphas, as measured by a cross-sectional regression. As an important step in exploring alphas, we perform a cross-sectional weighted-least squares regression of residual returns against alpha. We estimate a coefficient x that solves the following minimization problem:

$$\min_b \sum_i \frac{(r_i - \alpha_i b)^2}{\sigma_{\epsilon,i}^2} = \min_b \|\tilde{\mathbf{r}} - \hat{\boldsymbol{\alpha}}b\|^2$$

The solution is given by $b^* = \tilde{\mathbf{r}}' \hat{\boldsymbol{\alpha}} / \|\hat{\boldsymbol{\alpha}}\|^2$ and the residual sum of squares is $\|\tilde{\mathbf{r}}\|^2 - (\tilde{\mathbf{r}}' \hat{\boldsymbol{\alpha}})^2 / \|\hat{\boldsymbol{\alpha}}\|^2$, while the total sum of squares is $\|\tilde{\mathbf{r}}\|^2$. The Coefficient of Determination (“R squared”) is, in expectation, equal to:

$$R^2 = \frac{(\tilde{\mathbf{r}}' \hat{\boldsymbol{\alpha}})^2}{\|\hat{\boldsymbol{\alpha}}\|^2 \|\tilde{\mathbf{r}}\|^2} = (\text{IC})^2$$

And we can link the coefficient of determination in predictive regressions to the Information Ratio:

$$\text{IR} = \sqrt{R^2 n}$$

If there are T investment periods in a year, the annualized Information has a convenient form as a function of per-period cross-sectional R squared

$$\text{IR} = \sqrt{R^2 n T} = \text{IC} \sqrt{n T}$$

Otherwise stated, the annualized Information Ratio is equal to the Information Coefficient times the *independent number of return forecasts in a year*.

This relationship goes back to Grinold (1989) and Grinold and Kahn (1999), who named it *The Fundamental Law of Active Management*⁴. It is often invoked by practitioners. In practice, users of the formula do not whiten returns and alphas in Equation (4.13). Instead, they apply the law to the active component of a portfolio, so that Ω_r is replaced by the diagonal Ω_ϵ and α by α_\perp . Then, IC becomes the cross-sectional correlation between the standardized alphas (expected excess returns in units of volatility) and the standardized idio returns (so that they have unit variance).

The Fundamental Law has several important implications. The first, and most obvious one, is that performance is driven by two factors. The first one is a measure of skill: the Information Coefficient. It is a strength of the predictive strength of the signal. If we have the ability to extend our forecast to a larger panel of stocks without degrading its predictive power, we should do so! This is never the case in real life. Many investors also have a notion of “idea velocity”, expressed as the number of forecasts T per year. A higher idea velocity increases, in principle, the Information Ratio. It is really difficult, however, to increase effectively the frequency of *independent* forecasts.

The Fundamental Law also connects Information Ratio to an *ex post* measure (the IC) that is interpretable as a correlation, which can be related to a special kind of regression (as per Insight 4.3).

4.5 Investment Performance Metrics

A discretionary portfolio manager satisfies, a quantitative portfolio manager optimizes. Maybe this is not entirely fair to discretionary portfolio managers: they optimize too, the way tiger sharks optimize for food intake or migrating warblers minimize traveling distance to Cuba. On the other side, this describes accurately what a quant does. A quantitative manager has an objective, constraints, and information. Investment objectives are not the only ones entering the manager’s process. For example, the estimation process of a factor model requires a loss function to minimize. This chapter introduces and justifies the investment metrics that will appear in later chapters. In some cases, the metrics enter the optimization problem as objectives; in others, as constraints. The role played by objectives and constraints is, to some extent, interchangeable.

³Prove this step-by-step in Exercise 3.2.

⁴Quite an important name for a “law”! And why not? Nobody had thought of using this title at the time. I can imagine Grinold playing air guitar and singing to an Iron Maiden tune as he was drafting the original 1989 paper.

The performance metrics of a portfolio manager are, by and large, three:

- The expected return of the strategy;
- The volatility of the strategy;
- The Sharpe Ratio;
- The Information Ratio;
- Capacity.

4.5.1 Expected Return

The *expected return* of the strategy is defined as the ratio of profit and loss (PnL) to Assets Under Management. With the possible exception of Mother Teresa, investors prefer more money to less, and returns are an adequate way to describe this. Returns are preferred to actual money because the normalization makes the measure *stationary*, i.e., having (approximately) the same distribution across different investment periods, and *intensive* (as opposed to extensive), i.e., independent of the amount invested. This allows for better comparison across periods and across funds. Returns can be optimized either over the course of an investment period or over the lifetime of the strategy. In practice, the two problems are separable: we solve a sequence of single-period problems, which we embed in a larger multi-period problem.

4.5.2 Volatility

We introduced the *volatility* of returns in Chapter 2. The volatility of a strategy is also defined as the standard deviation of its returns. The use of volatility in investing is ubiquitous. From a descriptive point of view, one may stop here. From a normative point of view, volatility may be justified by making additional assumptions. Say that the investor solves a one-period utility optimization problem of the form $\max_{\mathbf{w}} E[u(\mathbf{w}' \mathbf{R})]$; the utility function describes the preference of the investor and is increasing and concave. Then, one can justify the use of volatility using three assumptions. The first, is by assuming that the utility function is well approximated by a second-order Taylor expansion centered at the expected payoff, so that $u \simeq aE(\mathbf{w}' \mathbf{R}) - bE[(\mathbf{w}' \mathbf{R} - E(\mathbf{w}' \mathbf{R}))^2]$. The second justification comes from assuming a quadratic utility⁵ $u = ax - bx^2$ and

⁵This is not an increasing utility, and this assumption is unrealistic anyway.

arbitrarily distributed returns; then the portfolio has also normally distributed payoff, and taking expectations we can express utility as a function of mean and variance. Lastly, one can assume arbitrary utility and normally distributed returns. Since mean and variance identify a normal distribution, we can express the expected utility as a function of these two parameters, albeit not necessarily additive. All the approaches above require that the second moment of returns be finite. As we discussed in Chapter 2, there is agreement that returns of many securities have finite variance, and that over longer time scales log returns resemble Gaussian returns. The assumption that the can approximate utility with a quadratic function is not unrealistic, since the optimization horizon in quantitative investing is short and the payoffs are small. Globally, utility is not quadratic, and this can be described by the parameters used in the utility approximation formula. A higher value of the ratio b/a can be interpreted as penalizing more the uncertainty of payoffs relatively to their expected values, i.e., in being more risk-averse. I am not discussing this further, since the topic is covered extensively in textbooks (see Section 7.5 for references), and is not essential to the remainder of this book.

4.5.3 Sharpe Ratio

The *Sharpe Ratio* is defined as the ratio of expected returns of a strategy to its volatility. It combines the previous two quantities by measuring returns in units of volatility over a certain period of time. If we assume that the returns⁶ of a strategy identically distributed and independent, the Sharpe Ratio is the same as the t-statistic of the mean of the return distribution. In finance, the Sharpe Ratio is named after William F. Sharpe, one of the authors for the Capital Asset Pricing Model (CAPM). The Sharpe Ratio has drawbacks and advantages. Its drawbacks are two. First, it is not quite justifiable as a metric that ranks uncertain outcomes. Aside from decision-theoretic considerations⁷ the Sharpe Ratio of a portfolio with negative expected return of -5% and volatility 5% is *higher* than the Sharpe Ratio with the same negative return, and volatility 10%. This is unintuitive at best and wrong at worst. Secondly, it inherits the limitation of volatility as a measure of risk. It is possible to replace the denominator with

⁶Or, more commonly, the *excess returns*, i.e., the returns of a strategy in excess of the risk-free rate, often the three-month Treasury yield. This is the return of holding a self-financed security: we borrow one dollar in the first period at the risk-free rate, and buy one dollar of the security. In the second period, we receive the security return, and pay off the loan. See also Section ??, where excess returns enter naturally.

⁷For these, see Huang and Litzenberger (1988).

one's favorite risk measure, of which there is a near-infinite supply⁸. There are advantages, however. First, the Sharpe Ratio is intuitive: return in units of "risk". Second, it comes with a rich arsenal of theoretical results. We have confidence intervals and characterizations of its empirical properties, theoretical results, like relationship between cross-sectional regressions and Sharpe Ratio. The Sharpe Ratio also implies a bound on the probability of incurring a certain loss. This follows from Cantelli's inequality. For a random variable ξ with mean μ and standard deviation σ , this inequality states that

$$P(\xi < \mu - \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2} \quad \text{or} \quad P(\xi < -\lambda) \leq \frac{\sigma^2}{\sigma^2 + (\lambda + \mu)^2} \quad (4.14)$$

If ξ is the annual return of a strategy, and s is the annualized Sharpe Ratio of the strategy, and the loss is expressed as a multiple of standard deviations $-L\sigma$, as practitioners often do, then the inequality is

$$P(\xi < -L\sigma) \leq \frac{1}{1 + (L + SR)^2} \quad (4.15)$$

This holds for *any* distribution of returns with Sharpe Ratio SR. For example, consider an annualized Sharpe Ratio of 3 and an annualized volatility of \$50M. The probability of a \$100M loss is no greater than two standard deviations is not greater than 3.8%. This a much higher value than we would obtain under the assumption of normal returns. In that case, the probability of a loss would be 2.9E-7.

4.5.4 Capacity

Whereas the return and Sharpe Ratio are well known and defined, the *capacity* of a strategy is not unequivocally defined. An informal definition of capacity is "the highest expected PnL that a strategy is able to produce over a certain horizon". You may ask, "but isn't expected PnL just equal to Sharpe Ratio times dollar volatility? Capacity is essentially the maximum volatility at which we can run a strategy". This would be true if Sharpe Ratio were independent of volatility, and in that case, why not run a strategy to infinity volatility, or at least to the proverbial 11? Sharpe, however, is almost always a decreasing function of volatility. For a large enough volatility, the Sharpe Ratio becomes zero, and beyond this threshold the strategy is unprofitable. The capacity of a strategy can be defined as the maximum PnL that can be attained, subject to

⁸See, for example, Bacon (2005) for a list of risk metrics, both theoretically justified and heuristic.

FAQ 4.1: *What are the dimensions of the Sharpe Ratio?*

Return, volatility and Sharpe Ratio depend on the time horizon over which they are measured. Daily return r is daily PnL on capital; annualized return, where we assume the same daily PnL over T trading days in a year, is rT . We say that returns have dimension $[\text{time}]^{-1}$. An example: a strategy has a return of $10\%/(1 \text{ year}) = 10\%/(T \text{ days}) = (10\%/T)/\text{days}$. Provided that returns are serially uncorrelated (see Subsection 2.1.5), variance is also linear in time, because the variance of returns over a year is the sum of T daily variances, so its dimension is $[\text{time}]^{-1}$. Volatility is the square root of variance and has the dimension of $[\text{time}]^{-1/2}$. The Sharpe Ratio has the dimension $[\text{return}]/[\text{volatility}] = [\text{time}]^{-1}/[\text{time}]^{-1/2} = [\text{time}]^{-1/2}$. When converting the horizon of a Sharpe Ratio for an equity strategy from a daily horizon to a monthly one, we multiply the daily Sharpe Ratio by $\sqrt{21}$, where 21 is the number of trading days in a month. The conversion factor to an annualized Sharpe Ratio in the US is $\sqrt{252}$.

a constraint that the Sharpe Ratio exceed an acceptable level. Alternatively, we could require a minimum bound on the expected return on capital. Defined this way, the capacity is an important parameter for hedge fund managers and portfolio managers alike. A strategy may have attractive return and Sharpe Ratio when run at low volatility. If it can yield only a modest PnL, it will be economically unattractive.

FAQ 4.2: *What is the confidence interval for the Sharpe Ratio?*

Suppose you observe T consecutive returns (or PnL), and estimate the Sharpe Ratio from these data. What is the confidence interval of this estimator? First, the Sharpe Ratio estimator is

$$\hat{\mu} := \frac{1}{T} \sum_{t=1}^T r_t \quad \hat{\sigma} := \sqrt{\frac{1}{T} \sum_{t=1}^T (r_t - \hat{\mu})^2}$$

$$\widehat{\text{SR}} := \frac{\hat{\mu}}{\hat{\sigma}}$$

For excess returns r_t that are iid and with finite variance, the estimator is normally distributed in the limit $T \rightarrow \infty$, with standard error

$$\text{SE}(\widehat{\text{SR}}) = \sqrt{\frac{1 + \text{SR}^2 / 2}{T}}$$

Compare this to the case in which we knew in advance the standard deviation of the returns. The Sharpe Ratio is then $\widehat{\text{SR}} := \hat{\mu}/\sigma$, and the SE is $\sqrt{1/T}$.

In the case of autocorrelated returns with $\text{cor}(r_s, r_t) = \rho^{|t-s|}$, the Sharpe Ratio estimator and the asymptotic confidence interval are, for small values of $|\rho|$,

$$\widehat{\text{SR}} := \frac{\hat{\mu}}{\hat{\sigma}} \sqrt{\frac{1 - \rho}{1 + \rho}} \simeq \frac{\hat{\mu}}{\hat{\sigma}} \sqrt{\frac{1 - \rho}{1 + \rho}} \simeq \frac{\hat{\mu}}{\hat{\sigma}} (1 - \rho)$$

4.6 ★Appendix

4.6.1 Convex Optimization

This appendix collects some basic facts about optimization of convex functions that are used in the book. Extensive treatments of convex optimization are in [Boyd and Vandenberghe \(2004\)](#), [Luenberger and Ye \(2008\)](#) and [Bazaraa et al. \(2006\)](#). Convexity plays a central role in our treatment. Fig. 4.1 illustrates the concept.

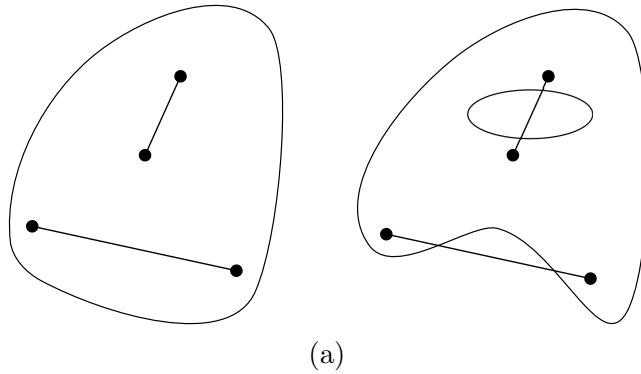


Figure 4.1: A convex set (left) and a non-convex set (right).

Definition 4.1. A convex set \mathcal{D} is a set such that for any $x, y \in \mathcal{D}$ and $t \in [0, 1]$, $tx + (1 - t)y \in \mathcal{D}$.

\mathcal{D} be a convex set in \mathbb{R}^n and $f : \mathcal{D} \rightarrow \mathbb{R}$.

Definition 4.2. f is convex is its graph is convex, i.e., for any pair $x, y \in S$, $f(tx + (1 - t)y) \geq tf(x) + (1 - t)f(y)$. It is strictly convex if the inequality is strict.

We assume below that f is convex and differentiable, and that the set S is convex, closed and bounded.

Theorem 4.1. The intersection of an arbitrary number of convex sets is a convex set.

Theorem 4.2. Let f be a convex function. The set $\{x | f(x) \leq 0\}$ is a convex set.

Definition 4.3.

1. A global minimum $x^* \in \mathcal{D}$ is a point such that $f(x^*) \geq f(x)$ for all $x \in \mathcal{D}$. It is strict if the inequality is strict.
2. A local minimum $x^* \in S$ is a point such that there is $\epsilon > 0$ such that $f(x^*) \geq f(x)$ for all $x \in S \cap B_\epsilon(x^*)$, where $B_\epsilon(x^*)$ is a ball of radius ϵ centered at x^* . It is strict if the inequality is strict.

Theorem 4.3.

1. A global maximum for f exists.
2. If x^* is a local minimum, then it is a global minimum.
3. if f is strictly convex, then it is the only global minimum and it is strict.

In this book, we are interested in characterizing the solutions of the problem

$$\begin{aligned} & \min f(x) \\ \text{s.t. } & g_i(x) \leq c_i \quad i = 1, \dots, m \\ & b'_j x = a_j \quad j = 1, \dots, p \end{aligned} \tag{4.16}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable and convex, e.g., $f(x) = -\alpha'x$ or strictly convex, e.g., $f(x) = \alpha'x + \kappa x'\Omega x$. g_i are differentiable convex functions⁹. Their sublevel sets $\{x | g_i(x) - c_i \leq 0\}$ are convex. The set $\{x | b'_i x = a_i\}$ where $b_i \in \mathbb{R}^n$, $a_i \in \mathbb{R}$, is also convex,. The set of feasible points for the optimization problem (also called *feasible region*) is the intersection of convex sets and is therefore convex. The maximization of a concave function is equivalent to the maximization of a convex function. Therefore the conditions of Th. 4.3 apply. Denote the optimal value of the objective function p^* .

4.6.2 Duality

Definition 4.4. The Lagrangian of the problem 4.16 is a function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$.

$$L(x, \lambda, \mu) := f(x) + \sum_{i=1}^n \lambda_i(g_i(x) - c_i) + \sum_{j=1}^p \mu_j(b'_j x - a_j)$$

The variables λ, μ are the Lagrange multipliers of Prob. (4.16).

⁹You may observe that the formulation is redundant. We can do away with the linear equalities and replace them with two inequality constraints $b'_j x \leq a_j$, $-b'_j x \leq -a_j$. Linear constraints are frequently found in practice, and the formulation we present simplifies the interpretation of solutions in the presence of linear constraints.

The Lagrange can be viewed as a penalized objective function. If λ and μ are sufficient large and, in the case of μ , have the correct sign, you can see the optimal solution x_L^* of L approach the optimal solution x^* of Prob. (4.16), because these large values “discourage” constraint infraction. The theorems below identify the limits of the intuition: minimizing L is indeed an intermediate step toward solving the problem is a seemingly different way.

Definition 4.5. *The dual function is defined as*

$$g(\lambda, \mu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \mu)$$

Theorem 4.4 (Weak Duality). *For each $\lambda \geq 0$, $\mu \in \mathbb{R}^p$*

$$g(\lambda, \mu) \leq p^*$$

We can take the sup of g ; since it is bounded above and λ, μ take value in a closed set, the sup is attained on the feasible set. Denote (λ^*, μ^*) the optimum point.

$$\begin{aligned} q^* &:= \max g(\lambda, \mu) \\ \text{s.t. } \lambda &\geq 0 \end{aligned}$$

$q^* \leq p^*$. When does the inequality become an equality? Even when the original problem is convex, the equality does not always hold. An additional condition is needed, called *constraint qualification*. More than a single condition, it is a family of requirements for the constraints that, when satisfied, ensure that weak duality becomes strong. Roughly speaking, these conditions ensure that constraints are not redundant. A common constraint qualification is *Slater’s Condition*: there exists x in the feasible region of Prob. (4.16) for which all the inequality constraints are strictly satisfied: $g_i(x) < 0$ for all $i = 1, \dots, m$.

Theorem 4.5 (Strong Duality). *If Slater’s condition holds, then $q^* = p^*$.*

From strong duality, it follows that

$$\begin{aligned} f(x^*) &= q^* = p^* = f(x^*) + \sum_{i=1}^n \lambda_i(g_i(x^*) - c_i) + \sum_{j=1}^p \mu_j(b'_j x^* - a_j) \\ &\Rightarrow \sum_{i=1}^n \lambda_i(g_i(x^*) - c_i) = 0 \\ &\Rightarrow \lambda_i(g_i(x^*) - c_i) = 0 \quad i = 1, \dots, m \end{aligned}$$

where the last term comes from the fact that each term is non-positive. This condition is termed *complementary slackness conditions*.

We collect the necessary conditions for optimality of the point (x^*, λ^*, μ^*) :

$$\nabla_x f(x) + \sum_{i=1}^n \lambda_i \nabla_x g_i(x) + \sum_{j=1}^p \mu_j b_j = 0 \quad (4.17)$$

$$g_i(x) \leq c_i \quad i = 1, \dots, m \quad (4.18)$$

$$b'_j x = a_j \quad j = 1, \dots, p \quad (4.19)$$

$$\lambda_i \geq 0 \quad i = 1, \dots, m \quad (4.20)$$

$$\lambda_i(g_i(x) - c_i) = 0 \quad i = 1, \dots, m \quad (4.21)$$

The first equation is the first-order necessary condition for the maximum of the Lagrangian. The second and the third equations are the primal feasibility conditions. The fourth equation is the dual feasibility condition. The last one is the complementary slackness condition. Considered together, Eqs. (4.17-4.21) are referred to as the *Karush-Kuhn-Tucker conditions*, or KKT conditions. A central result in optimization theory is that these conditions are also sufficient.

Theorem 4.6 (KKT conditions). *If a point x^* satisfies the KKT condition, then it is an optimal point for pb. (4.16).*

There are many interpretations of the dual, and for an extensive, clear treatment I recommend Boyd and Vandenberghe [Boyd and Vandenberghe \(2004\)](#). There are several interpretations and uses for the values (λ^*, μ^*) . We mention two.

Equivalent Penalized Problem. Say we solve the unconstrained problem

$$\min_x L(x, \lambda^*, \mu^*)$$

where λ^*, μ^* solve the KKT conditions. Then the solution satisfies 4.17. Hence the optimum is reached at x^* . λ^*, μ^* are the penalization constants such that an unconstrained penalized optimization problem is equivalent to a constrained one.

4.6.3 Local Analysis

It is possible to fully characterize the sensitivity of the solution of Pb. 4.16 to changes in the input parameters without having to solve the problem anew.

Envelope Theorem. Consider a variant of Pb. (4.16) in which, without loss of generality, we omit the linear equality constraints:

$$V(\theta) := \min f(x, \theta) \quad (4.22)$$

$$\begin{aligned} \text{s.t. } g_i(x, \theta) &\leq 0 & i = 1, \dots, m \\ x &\in \mathbb{R}^n \end{aligned} \quad (4.23)$$

Both f and the functions g_i are differentiable functions of a parameter vector $\theta \in \mathbb{R}^p$. Fix θ^* and let x^*, λ^* the optimum point and Lagrangian multipliers.

Theorem 4.7.

$$\frac{dV(x^*, \theta)}{d\theta_i} \Big|_{\theta=\theta^*} = \frac{\partial f(x^*, \theta)}{\partial \theta_i} \Big|_{\theta=\theta^*} + \sum_{i=1}^m \lambda_i^* \frac{\partial g_i(x^*, \theta)}{\partial \theta_i} \Big|_{\theta=\theta^*}. \quad (4.24)$$

Proof. By the Implicit Function Theorem, the function $x^*(\theta), \lambda^*(\theta)$ are differentiable.

$$\begin{aligned} \frac{\partial V}{\partial \theta_i} &= \frac{\partial f}{\partial \theta_i} + \sum_{k=1}^n \frac{\partial f}{\partial x_k} \frac{\partial x_k}{\partial \theta_i} + \sum_{j=1}^m \lambda_j^* \frac{\partial g_j}{\partial \theta_i} + \sum_{j=1}^m \sum_{k=1}^n \frac{\partial \lambda_j^*}{\partial x_k} \frac{\partial x_k}{\partial \theta_i} g_j \\ &= \frac{\partial f}{\partial \theta_i} + \sum_{j=1}^m \lambda_j^* \frac{\partial g_j}{\partial \theta_i} + \sum_k \frac{\partial x_k}{\partial \theta_i} \frac{\partial}{\partial x_k} \left(f + \sum_j \lambda_j^* g_j \right) \\ &= \frac{\partial f}{\partial \theta_i} + \sum_{j=1}^m \lambda_j^* \frac{\partial g_j}{\partial \theta_i} \end{aligned}$$

the last equality holds because of Eq. (4.17). \square

Shadow Prices. The magnitude of the Lagrange multipliers can be interpreted as a measure of the impact that a constraint exerts on the solution. A large multiplier corresponds to high penalty associated to the constraint. If we removed the constraint, the solution would likely be improved. A constraint that is not binding has no impact on the final solution; by complementary slackness, the associated Lagrange multiplier must be zero. There is a well-established quantitative relationship between Lagrange multipliers and marginal impact of the constraint. In Pb 4.16, let $V(a, c)$ be the solution, which we parametrize as function of a, c ; and correspondingly, let $\mu(a, c), \lambda(a, c)$ be the Lagrange

multipliers. It is a direct corollary of the Envelope Function Theorem that

$$\frac{\partial V}{\partial a_i} = \mu_j \quad i = 1, \dots, p \quad (4.25)$$

$$\frac{\partial V}{\partial c_j} = \lambda_j \quad j = 1, \dots, m \quad (4.26)$$

We can interpret the RHS parameters as upper bounds on quantities. For example, a linear constraint may be an upper bound on total exposure of a portfolio to a certain factor. The Lagrange multiplier is then the *price* of that exposure: if we relaxed the exposure by one unit, we would experience an improvement in the overall performance equal to the Lagrange multiplier. For this reason, Lagrange multipliers are also termed *shadow prices*.

Local Change of optimal point. A related question is that of expressing the changes in the optimal point $x^*(\theta)$ as the parameter θ changes. We limit our analysis only to the unconstrained case and $\theta \in \mathbb{R}$, as it is the one needed in this book; the multivariate and constrained one follows along the same lines. The first-order conditions are

$$\frac{\partial f(x^*(\theta), \theta)}{\partial x_i} = 0 \quad \text{for all } \theta \quad (4.27)$$

Define $H \in \mathbb{R}^{n \times n}$ and $g \in \mathbb{R}^n$ as

$$(H)_{ij} := \frac{\partial^2 f}{\partial x_j \partial x_k}(x^*(\theta), \theta) \quad (4.28)$$

$$g_i := \frac{dx_i^*}{d\theta} \quad (4.29)$$

$$b := \frac{\partial}{\partial \theta} \nabla_x f(x^*(\theta), \theta) \quad (4.30)$$

Then

$$0 = \frac{d}{d\theta} \left(\frac{\partial f(x^*(\theta), \theta)}{\partial x_i} \right) = 0 \quad (4.31)$$

$$= \sum_k H_{jk} g_k + b = 0 \quad (4.32)$$

$$\Rightarrow g(\theta^*) = -H^{-1}(\theta^*)b(\theta^*) \quad (4.33)$$

Parametrized Optimization Problems. Consider now the case in which we want to optimize a function $f(x, \theta)$, which depends on a parameter $\theta \in \mathbb{R}$. The

parameter θ may model an input to the model, but may also be used to describe a family of functions. Let $x^*(\theta) := \arg \min\{f(x, \theta) : x \in \mathbb{R}^n\}$. Consider the case in which we solve the problem for θ_0 . This may correspond to our best estimate of an input parameter; or to a functional form which we are able to solve exactly. Let $\delta(\theta_0, \theta)$ be the difference between the value of the objective function at $x^*(\theta_0)$ when the value of the parameter is θ , and the optimal value of the objective function when the parameter is θ_1 :

$$\delta(\theta) := f(x^*(\theta_0), \theta) - f(x^*(\theta), \theta) \quad (4.34)$$

This can be interpreted as an “optimization” error when the parameter choice is not correct. From the definition of x^* , $\delta \geq 0$. We would like to have an estimate for the upper bound of δ .

$$\frac{d\delta}{d\theta} = \frac{\partial f}{\partial \theta}(x^*(\theta_0), \theta) - \langle \nabla_x f(x^*(\theta), \theta), \frac{dx^*}{d\theta}(\theta) \rangle - \frac{\partial f}{\partial \theta}(x^*(\theta_0), \theta) \quad (4.35)$$

$$= - \langle \nabla_x f(x^*(\theta), \theta), \frac{dx^*}{d\theta}(\theta) \rangle \quad (4.36)$$

Since $\langle \nabla_x f(x^*(\theta_0), \theta_0) \rangle = 0$, $d\delta/d\theta$ at θ_0 is zero. The second derivative is given by

$$\frac{d^2\delta}{d\theta^2} = - \langle \nabla_x f(x^*(\theta), \theta), \frac{dx^*}{d\theta}(x^*(\theta_0), \theta) \rangle \quad (4.37)$$

$$= g' H g \quad (4.38)$$

$$- \langle \frac{\partial}{\partial \theta} \nabla_x f(x^*(\theta), \theta), \frac{dx^*}{d\theta}(\theta) \rangle \quad (4.39)$$

$$- \langle \nabla_x f(x^*(\theta), \theta), \frac{d^2x^*}{d\theta^2}(\theta) \rangle \quad (4.40)$$

at $\theta = \theta_0$ the gradient in the third term is zero, and we use

$$\frac{d^2\delta}{d\theta^2} = b' H^{-1} b - \langle b, -H^{-1} b \rangle = 2b' H^{-1} b \quad (4.41)$$

As an application of the formula above, we estimate the optimization error when we approximate convex differentiable g function with a quadratic function h such that, for some point x_0 , $g(x_0) = gh(x_0)$, $\nabla_x g(x_0) = \nabla_x h(x_0)$ and $H_g(x_0) = H_h(x_0)$. The parametrized function is $f(x, \theta) = \theta g(x) + (1 - \theta)h(x)$. We optimize the quadratic function, corresponding to $\theta_0 = 0$. Let the optimal point be x^* . The vector b is

$$b = \nabla_x(g(x^*) - h(x^*)) \quad (4.42)$$

which we can approximate, since $\nabla_x(g(x^*) - h(x^*)) \simeq \nabla_x(g(x_0) - h(x_0)) + H_{g-h}(x_0)(x^* - x_0) = H_{g-h}(x_0)(x^* - x_0) = 0$ since $H_{g-h}(x_0) = 0$. It follows that, to a first approximation, $d^2\delta/d\theta^2 = 0$. The error is cubic in θ .

4.6.4 Solutions To Specific Optimization Problems

Maximize expected return subject to a vol constraint and linear homogeneous equalities.

$$\max \boldsymbol{\alpha}' \mathbf{w} \quad (4.43)$$

$$\text{s.t. } \mathbf{B}' \mathbf{w} = 0 \quad (4.44)$$

$$\mathbf{w}' \boldsymbol{\Omega} \mathbf{w} \leq \sigma^2 \quad (4.45)$$

The solution w^* to this problem is given by

$$\begin{aligned} \boldsymbol{\Pi} &:= I_n - \mathbf{B}(\mathbf{B}' \boldsymbol{\Omega}^{-1} \mathbf{B})^{-1} \mathbf{B}' \boldsymbol{\Omega}^{-1} \\ \tilde{\boldsymbol{\alpha}} &:= \boldsymbol{\Pi} \boldsymbol{\alpha} \\ \mathbf{w}^* &= \frac{\sigma}{\sqrt{\tilde{\boldsymbol{\alpha}}' \boldsymbol{\Omega}^{-1} \tilde{\boldsymbol{\alpha}}}} \boldsymbol{\Omega}^{-1} \tilde{\boldsymbol{\alpha}} \end{aligned}$$

Maximize expected returns using Sharpe Ratio and Asset Correlation Matrix. We have a standard MVO problem, but the available data are the Sharpe Ratios of the assets and their correlations. The solution can be expressed as an intuitive function of these data. The MVO portfolio is proportional to the portfolio $\boldsymbol{\Omega}_r^{-1} \boldsymbol{\alpha}$. We write $\boldsymbol{\Omega}_r = \mathbf{V} \mathbf{C} \mathbf{V}$ where \mathbf{V} is the diagonal matrix containing asset volatilities, and \mathbf{C} is the correlation matrix. The Sharpe Ratio of asset i is $s_i := \alpha_i / \sigma_i$.

$$\begin{aligned} \mathbf{w}^* &= (\mathbf{V} \mathbf{C} \mathbf{V})^{-1} \boldsymbol{\alpha} \\ &= \mathbf{V}^{-1} \mathbf{C}^{-1} (\mathbf{V}^{-1} \boldsymbol{\alpha}) \\ &= \mathbf{V}^{-1} \mathbf{C}^{-1} \mathbf{s} \\ \Rightarrow \mathbf{V} \mathbf{w}^* &= \mathbf{C}^{-1} \mathbf{s} \end{aligned}$$

The Sharpe Ratio of the MVO-optimal portfolio is $\sqrt{\mathbf{s}' \mathbf{C}^{-1} \mathbf{s}}$. Hence the optimal dollar volatility is a function of correlation and sharpe only. For uncorrelated assets, the optimal dollar volatility is proportional to the Sharpe Ratio.

Chapter 5

MVO and Its Discontents

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

5.1 Shortcomings of Naïve MVO

Before introducing more complex optimizations, let’s work through a simple example—maybe the simplest instance of the simplest optimization problem—to illustrate the implications of MVO. We have just two assets, with non-negative Sharpe Ratios s_1, s_2 . Their returns have correlation ρ . The inverse of the covariance matrix is

$$\mathbf{C}^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

So by Equation (4.9), the optimal volatility allocation is

$$v_1^* = \frac{\kappa}{1 - \rho^2} (s_1 - \rho s_2)$$
$$v_2^* = \frac{\kappa}{1 - \rho^2} (s_2 - \rho s_1)$$

If $s_2/s_1 < \rho$, then we short asset 2. Consider first the case where $s_2 = 0$ and $\rho > 0$. In this case, we always short the asset. Asset 2 acts as *hedge*. Shorting it is beneficial because a) it has no cost (zero expected return); b) it reduces the volatility of the portfolio, since it is positively correlation to asset 1. When the Sharpe Ratio of asset 2 is positive, then there is a cost to shorting, and for the hedge to be beneficial, the correlation must exceed a threshold s_2/s_1 .

Even though the recommendation to short an asset with positive returns is explainable, it is probably at odds with the intuition of many readers. If two assets are very correlated, wouldn't it be preferable to go long both, thus averaging out the signal error? We can make this reasoning more rigorous by assessing the impact of estimation error on expected returns and on the correlation.

Impact of errors in forecasted Sharpe Ratios. We denote the *true* Sharpe Ratios \tilde{s}_i , and assume that the error between the true and forecasted Sharpe Ratios is bounded: $\|\tilde{\mathbf{s}} - \mathbf{s}\| \leq \epsilon$. The realized expected return is

$$E(\text{PnL}) = \frac{\kappa}{1 - \rho^2} [(s_1 - \rho s_2) \tilde{s}_1 + (s_2 - \rho s_1) \tilde{s}_2]$$

Insight 5.1: A Simple Linear-Quadratic Problem

Let $\mathbf{a}, \mathbf{x}_0 \in \mathbb{R}^n$. The problem $\min\{\langle \mathbf{a}, \mathbf{x} \rangle \mid \|\mathbf{x} - \mathbf{x}_0\|^2 \leq \epsilon^2\}$ has solution

$$\begin{aligned} \mathbf{x}^* &= \mathbf{x}_0 - \frac{\mathbf{a}}{\|\mathbf{a}\|} \epsilon \\ \frac{\langle \mathbf{a}, \mathbf{x}^* \rangle}{\langle \mathbf{a}, \mathbf{x}_0 \rangle} - 1 &= - \frac{\|\mathbf{a}\|}{\langle \mathbf{a}, \mathbf{x}_0 \rangle} \epsilon \end{aligned}$$

In the worst case, we solve the problem $\min\{E(\text{PnL}) \mid \|\tilde{\mathbf{s}} - \mathbf{s}\| \leq \epsilon\}$. I leave the solution as an exercise (also, see Insight 5.1); the relative reduction in PnL is

$$-\frac{\sqrt{(s_1 - \rho s_2)^2 + (s_2 - \rho s_1)^2}}{(s_1 - \rho s_2)s_1 + (s_2 - \rho s_1)s_2} \epsilon$$

This is also the relative loss in Sharpe, since the volatility of the portfolio is unaffected by return forecast error. Figure 5.1 shows numerical results for two assets, assuming an error $\epsilon = 0.5$ and varying levels of correlation. An error of 0.5 is perhaps conservative; actual differences in forecasted versus realized Sharpe Ratios are higher. Notice that high correlation makes things worse. In all scenarios, the percentage in efficiency is significant. It is of course lower for higher Sharpes (because the relative forecasting error is smaller); and is higher for higher correlations. In all cases it exceeds 10% and can be as high as 50%.

Impact of errors in correlation among assets. We denote the true correlation $\tilde{\rho}$ and assume that the estimation error is bounded: $|\tilde{\rho} - \rho| \leq \epsilon$. The

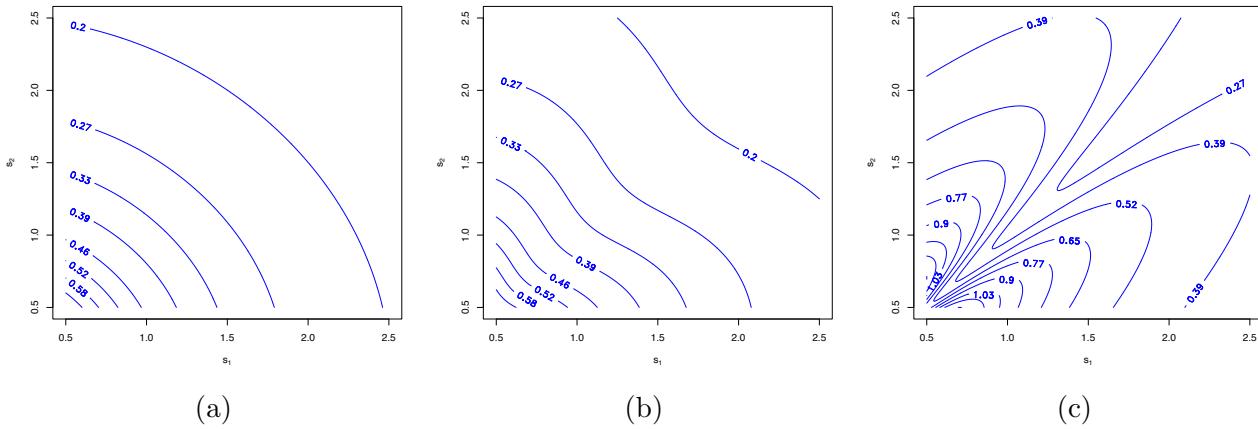


Figure 5.1: Level plots of the loss of PnL (and Sharpe Ratio) as a function of the Sharpe Ratio of two assets, assuming a maximum error ϵ in the Sharpe Ratio norm. Parameters: $\epsilon = 0.5$; Correlation: (a) $\rho = 0.1$, (a) $\rho = 0.5$, (c) $\rho = 0.9$.

error in estimated correlation affects the volatility. We solve the problem $\max\{(\mathbf{V}^*)'\tilde{\mathbf{C}}^{-1}\mathbf{V}^* | |\tilde{\rho} - \rho| \leq \epsilon\}$. In this case the worst-case realized relative volatility increase (exercise!) is

$$-\frac{2|(s_1 - \rho s_2)(s_2 - \rho s_1)|\epsilon}{\sqrt{(\mathbf{V}^*)'\mathbf{C}^{-1}\mathbf{V}^*} + 2|(s_1 - \rho s_2)(s_2 - \rho s_1)|\epsilon}$$

I am showing the impact of the error in Figure 5.2, for a reasonable error in correlation estimate of 0.1. However, in periods of crisis, the error can be larger (albeit not dramatically so). In Figure 5.3 I show the impact of correlation error on Sharpe.

Insight 5.2: Degradation in Performance due to Forecasting Error

When we use Naïve MVO optimization, the degradation in Sharpe Ratio arising from forecasted (*ex ante*) parameters for volatilities and returns vs realized values (*ex post*) can easily range in the 10-50%.

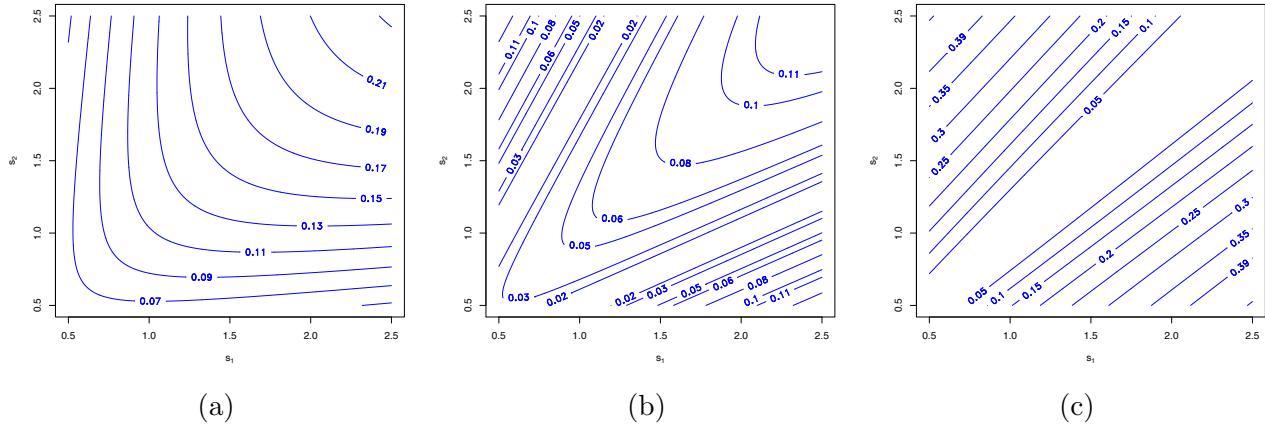


Figure 5.2: Level plots of the loss of PnL (and Sharpe Ratio) as a function of the Sharpe Ratio of two assets, assuming a maximum error ϵ in the correlation. Parameters: $\epsilon = 0.1$; Correlation: (a) $\rho = 0.1$, (b) $\rho = 0.5$, (c) $\rho = 0.9$.

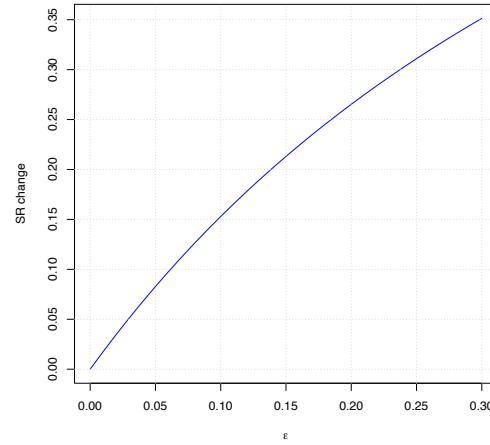


Figure 5.3: fraction loss in Sharpe Ratio for two strategies with Sharpe Ratios of 3 and 2, return correlation $\rho = 0.3$, and error ϵ ranging from 0 to 0.3.

5.2 Constraints and Modified Objectives

Equation (4.2) is the starting point for more complex optimization problems. They reflect the detailed preferences of the investors, short-term concerns, regulatory constraints, and implementation considerations. In applications, optimization formulations differ widely because they address a wide range of concerns:

- *Investor's preferences*: “Keep medium-term momentum exposure exactly equal to zero.”
- *Tactical considerations*: “Don't trade this stock because it could be acquired tomorrow” or ”liquidate this stock because it could be acquired tomorrow”; both are valid, if incompatible, concerns.
- *Regulatory considerations*: “The portfolio must be long only”.
- *Fiduciary considerations*: “The portfolio must *track* a benchmark, i.e., the difference in returns between the portfolio's returns and the benchmark's cannot exceed a certain *tracking volatility*”.
- *Implementation considerations*: “The objective function must include the trading costs.”

From a modeling viewpoint, constraints can take several forms. We introduce those first, and then we map them to the applications at hand. The “mapping” part will be either instructive if you have never been exposed to it, or terminally boring if you have worked in portfolio management for a few years. Rejoice with the former group, and commiserate with the latter.

5.2.1 Types of Constraints

Although one can imagine infinite types of constraints, some of them are much more common than others. We review them below.

- *Linear constraints*. These can be inequality or equality constraints:

$$\begin{array}{ll} \mathbf{A}'\mathbf{w} \leq \mathbf{c} & (\text{Inequality constraints}) \\ \mathbf{A}'\mathbf{w} = \mathbf{c} & (\text{Equality constraints}) \end{array} \quad (5.1)$$

These are perhaps the most common constraints in financial optimization, because they are used to address several of the concerns listed at the

beginning of the section. For example, some strategies are required to be *long-only*. The constraint is simply

$$\mathbf{w} \geq 0 \quad (\text{Long-Only constraint})$$

Extending this to a bound on maximum short and long size for a single position is only a small step. The rationales for such *box constraints* are many. There are natural limits due to maximum institutional ownership of a stock (say, no more than 5% of the outstanding stocks); or to the maximum risk concentration in a stock: the idiosyncratic variance of a stock may not exceed a certain percentage of the total idiosyncratic variance, which translates to a linear constraint. Further still, we may impose a maximum liquidation cost requirement on all stocks; which also becomes a constraint on single position size.

A slightly more complex constraint, which does not seem linear at first sight, is on Gross Market Value: $\sum_i |w_i| \leq G$. This constraint may originate on limits on financial leverage that the fund wants to apply to its managed assets. The constraint may be turned into a linear one¹ by introducing ancillary variables representing the long and short side of a position, and additional constraints:

$$\mathbf{x} \geq 0 \quad (\text{GMV constraint}) \tag{5.2}$$

$$\mathbf{y} \geq 0 \tag{5.3}$$

$$\mathbf{w} = \mathbf{x} - \mathbf{y} \tag{5.4}$$

$$\sum_i (x_i + y_i) \leq G \tag{5.5}$$

A similar constraint is on the long vs short ratio². If we want the long/short ratio to be equal to a certain value, then the constraint is $\sum_i w_i^+ = K \sum_i w_i^-$. This constraint is the same as the GMV constraint, with the exception of Equation (5.5), which we replace with

$$\sum_i x_i = K \sum_i y_i \quad (\text{Long/Short ratio constraint})$$

Yet another class of constraint is that on factor model exposures, and on exposures to other asset characteristics not in the model. An example

¹Before rediscovering the wheel, know that some financial optimization packages abstract the modeling of the GMV constraint, so that you just have to specify it.

²For example a few years ago, *130/30 portfolios* were popular. These strategies managed net-long portfolios, with a 30% of Net Market Value invested in shorts and 130 % invested in longs

is the constraint on historical market betas β_i . The constraint then is $\sum_i \beta_i w_i = b_0$. The general form of factor exposure is *verbatim* that of Equations (5.1).

A constraint on maximum portfolio turnover takes a similar form to the previous constraints that use absolute values. I am leaving it as an exercise to the reader. The turnover constraint may be either (poorly) justified to control costs, or by fiduciary requirements on portfolio turnover. A better way to model costs takes us in the domain of non-linear constraints.

- *Non-linear constraints.* A constraint of a different nature is trading-related. Trading occurs over many periods, and one approach to control excessive trading is to limit the traded capital, possibly weighted to account for asset-specific trading const, in each portfolio rebalancing. This is equivalent to assuming linear transaction costs. We generalize this at little cost, and model trading costs as superlinear in traded amount but growing at a quadratic rate or less: $c_i |\Delta w_i|^\gamma$, where $\gamma \in [1, 2]$. The constraint takes the form:

$$\sum_i c_i |w_i - w_i^{\text{start}}|^\gamma \leq C \quad (\text{Trading cost constraint})$$

where $\mathbf{w}^{\text{start}}$ is the portfolio held at the beginning of the period. The constraint is convex, so that the portfolio optimization problem has a unique solution.

Quadratic constraints appear naturally when we want to control risk at a finer resolution than that on total variance. For example , let Ω_f^{style} be the principal submatrix in the factor covariance matrix, and let $\mathbf{b}^{\text{style}} = (\mathbf{B}^{\text{style}})' \mathbf{w}$ be the vector of style factor exposures. Then a constraint on the maximum style-factor risk becomes

$$\begin{aligned} \mathbf{b}^{\text{style}} &= (\mathbf{B}^{\text{style}})' \mathbf{w} \\ (\mathbf{b}^{\text{style}})' \Omega_f^{\text{style}} \mathbf{b}^{\text{style}} &\leq \sigma_{\text{style}}^2 \quad (\text{Style factor vol constraint}) \end{aligned}$$

Risk constraint are often not only applied to the positions of a portfolio, but to the *active* positions of the portfolio itself. For example, consider a long-only portfolio with a GMV of \$1B, and let $\mathbf{w}^{\text{bench}}$ be the positions of a portfolio with the same GMV, with weights proportional to those of the SP500 benchmark. The *active holdings* are $\mathbf{w}^a = \mathbf{w} - \mathbf{w}^{\text{bench}}$. *tracking error* is the volatility of the active portfolio, and is a measure of the

freedom the portfolio manager has in selecting stocks. A constraint on the tracking error is

$$(\mathbf{w}^a)' \boldsymbol{\Omega}_r \mathbf{w}^a \leq \sigma_a^2 \quad (\text{tracking error constraint})$$

- *Non-convex constraints.* Finally, there are a few constraint types that lead to a non-convex feasible region. Finding a global optimum is in general NP-hard. Convex solvers may either not accept such constraints, or may not converge. I would argue that, in most cases, these constraints should *not* be used on grounds of sensible modeling. I am presenting them both for completeness and as a cautionary tale.

The first constraint type is on the maximum number N_{\max} of assets in the portfolio. This is usually implemented by introducing 0/1 variables x_i , and by setting a maximum (large) absolute position size M . The constraint becomes

$$|w_i| \leq M x_i \quad i = 1, \dots, n \quad (\text{Max number of positions}) \quad (5.6)$$

$$\sum_{i=1}^n x_i \leq N_{\max} \quad (5.7)$$

$$x_i \in \{0, 1\} \quad i = 1, \dots, n \quad (5.8)$$

The rationale for this constraint is that a very broad portfolio may be too burdensome to trade or manage. This combinatorial constraint can be handled by some commercial solvers for realistic problem instances with thousands of assets. However, its utility is limited. It is usually preferable to model trading costs directly, and either not include a constraint at all, or have a threshold for trading below which the trades of the optimal solution are set to zero. This usually has a negligible impact on optimality.

A very different type of constraint is on percentage of idio variance. We have mentioned this metric in Section 3.5.2. It is tempting to include a constraint of the form

$$\mathbf{w}' \boldsymbol{\Omega}_\epsilon \mathbf{w} \geq p_{\text{idio}} \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \quad (5.9)$$

or, equivalently,

$$\mathbf{w}' [p_{\text{idio}} \mathbf{B} \boldsymbol{\Omega}_f \mathbf{B}' - (1 - p_{\text{idio}}) \boldsymbol{\Omega}_\epsilon] \mathbf{w} \leq 0 \quad (5.10)$$

The problem is that the matrix $p_{\text{idio}} \mathbf{B} \boldsymbol{\Omega}_f \mathbf{B}' - (1 - p_{\text{idio}}) \boldsymbol{\Omega}_\epsilon$ is in general not positive definite, and therefore the constraint is not convex (exercise: prove it).

A constraint type with a similar objective is to require a minimum idiosyncratic dollar volatility: $\mathbf{w}' \boldsymbol{\Omega}_\epsilon \mathbf{w} \geq \sigma_{\text{idio}}^2$. This is obviously a non-convex constraint, and its proponents should be excommunicated from the Orthodox Church of Optimization. An sensible approach is to simply upper bound the factor variance, or impose bounds of factor exposures, and test the impact of the bound on the portfolio's performance.

Yet another excommunicable offense is imposing a lower bound on total volatility. I would not mention it, had I not witnessed actual humans proposing it.

In the same spirit, i.e., the goal of ensuring that the portfolio meets a minimum size, is a lower bound on Gross Market Value. The answer to these constraints is that they are usually ill-conceived. If, after accounting for excess return forecasts, trading costs, and risk constraints, the optimal portfolio is small, then maybe it should stay small. And if one really wants to make it bigger (again, not advisable), one could loosen the upper bounds on risk or underestimate the transaction costs.

5.2.2 Do Constraints Improve or Worsen Performance?

The naïve answer to the title of this section is that—of course!—they worsen performance. If you reduce the feasible region of your optimization problem by adding a constraint, you will not get a better optimum. Specifically, if we maximize the Sharpe Ratio, adding constraints will degrade the Sharpe Ratio³ This is true if the data in the problem, i.e. covariance matrix and expected returns, are estimated correctly. If we take estimation error into account, however, constraints may help. The next section interprets constraints as regularization terms for parameters entering in the optimization⁴.

³For example, see [Clarke et al. \(2002\)](#).

⁴The academic literature on this subject is not very large. See [Jagannathan and Ma \(2003\)](#) for an early contribution to the analysis of long-only constraints; The work by [DeMiguel et al. \(2009a,b\)](#) on trading penalties; [Fan et al. \(2012\)](#) on GMV constraints and [Ceria et al. \(2012\); Saxena and Stubbs \(2013\)](#) on penalties on the factor covariance matrix.

5.2.3 Constraints as Penalties

One alternative way to interpret a constraint in portfolio optimization is as a penalty term added to the objective function: given a problem

$$\begin{aligned} & \max f(\mathbf{x}) \\ \text{s.t. } & g(\mathbf{x}) \leq a \end{aligned}$$

with optimal solution $\mathbf{x}^*(a)$, there is a $\lambda^*(a) > 0$ such that

$$\max f(\mathbf{x}) - \lambda^*(a)g(\mathbf{x})$$

has the same solution $\mathbf{x}^*(a)$. We used this result at the beginning of the chapter. The parameter $\lambda^*(a)$ can also be interpreted as a sensitivity to the constraint's right-hand side parameter a . The variable λ is the marginal change in the optimum when we increase (or “relax”) a : $df(\mathbf{x}^*(a))/da = \lambda(a)$. Since a commercial solver returns both \mathbf{x}^* and λ^* , this means that we get sensitivities at zero additional cost. This results also opens up a different modeling approach. What if we converted constraints into penalties? We now know that the outcome, for the appropriate penalizing coefficient, is the same. Does this mean that the approaches are equivalent? The answer is no, and the remainder of this section is devoted to illustrating the difference.

First, let us focus our attention on the meaning of constraints and penalties. There are constraints that are commensurable with the objective, and that are naturally expressed as penalties. For example, you could put a constraint on maximum trading costs. However, costs and expected PnL in the objective have the same unit (dollar) and it makes more sense to express the objective function as the difference of PnL and trading cost. The penalty parameter is simply one. What about risk? If we fix the time interval, the variance constraint has the dimension of dollar squared, and is therefore not commensurable to PnL in the objective. What we could add to the objective function is $\sqrt{\mathbf{w}'\Omega_r\mathbf{w}}$. This is possible in some optimization packages⁵. However, if we know the approximate value σ_0 of final volatility, we can choose a penalty parameter such that the adding a volatility term or a variance one gives a similar result. We do so by

⁵A volatility constraint or penalty is in practice computationally more burdensome to solve than a variance constraint or penalty.

linearizing in the region of the optimum portfolio:

$$\begin{aligned} -\lambda \sqrt{\sigma_0^2 + (\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} - \sigma_0^2)} &\simeq -\frac{\lambda}{\sigma_0} (\mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} - \sigma_0^2) \\ &= \lambda \sigma_0 - \tilde{\lambda} \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \\ \tilde{\lambda} &:= \frac{\lambda}{\sigma_0} \end{aligned}$$

The constant term is irrelevant to the optimization problem, and the volatility is locally approximated by a variance.

A second class of constraints does not have an obvious interpretation. Should we add the constraint on GMV as a penalty? Or long-only constraints? The answer, somewhat surprisingly, is that adding those constraints as penalty may actually help the performance of the optimized portfolio, when the parameters in the model are not accurately estimated.

Let us start with an augmented version of Problem (4.6):

$$\begin{aligned} \max \boldsymbol{\alpha}' \mathbf{w} - \lambda \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} \\ \text{s.t. } \|\mathbf{w}\|^2 \leq G \end{aligned} \tag{5.11}$$

whose penalized version is

$$\max \boldsymbol{\alpha}' \mathbf{w} - \lambda \mathbf{w}' \boldsymbol{\Omega}_r \mathbf{w} - \nu \|\mathbf{w}\|^2 \tag{5.12}$$

This problem can be interpreted in many different ways. The first one is a simple rewriting of the quadratic terms as $\lambda \mathbf{w}' (\boldsymbol{\Omega}_r + (\nu/\lambda) \mathbf{I}_n) \mathbf{w} =: \mathbf{w}' \tilde{\boldsymbol{\Omega}}_r \mathbf{w}$. The problem then is a MVO problem with a modified covariance matrix. The correlations of the original covariance matrix have been reduced by a factor $\lambda/(\lambda+\nu)$. The asset variances have been increased, and are more similar to each other; in the limit $\nu \rightarrow \infty$ they are identical. The norm constraint therefore has a “regularizing” effect on the solution. There are different optimization formulations that lead to the same solution of the optimization problem (5.12).

Uncertain Alpha. Let us start with the assumption that the vector $\boldsymbol{\alpha}$ is not known with accuracy. We instead have the knowledge that the vector is distributed according to a multivariate Gaussian: $\boldsymbol{\alpha} \sim N(\boldsymbol{\alpha}_0, \tau^2 \mathbf{I}_n)$. We still solve a MVO, taking into account alpha uncertainty:

$$\text{var}(\mathbf{r}' \mathbf{w}) = \text{var}(\boldsymbol{\alpha}' \mathbf{w}) + \text{var}(\mathbf{r} - \boldsymbol{\alpha})' \mathbf{w} = \mathbf{w}' (\tau^2 \mathbf{I}_n + \boldsymbol{\Omega}_r) \mathbf{w}$$

The MVO formulation is again the same as that of Equation (4.6), but with a modified covariance matrix. As in the case of Equation (5.12), the variances are made more equal, and correlations are shrunk toward zero.

Robust Alpha. Instead modelling alphas' imperfect estimation by assuming that we know their distribution, we model their error deterministically, and adversarially: we know that the true alphas are within a certain distance d from our estimate and, as we did at the beginning of the chapter, we look at the worst case, i.e., the realized alpha is worst possible one among the admissible realizations. In formulas, we solve

$$\max \mathbf{a}'\mathbf{w} - \lambda \mathbf{w}'\boldsymbol{\Omega}_r \mathbf{w} \quad (5.13)$$

$$\text{s.t. } \mathbf{a} = \arg \min_{\mathbf{x}} \{ \mathbf{x}'\mathbf{w} \mid \|\mathbf{x} - \boldsymbol{\alpha}\| \leq d \} \quad (5.14)$$

We know what the solution to the nested problem (5.14): from Insight 5.1, it is equal to $\mathbf{a} = \boldsymbol{\alpha} - d\mathbf{w}/\|\mathbf{w}\|$. Hence we solve

$$\max \boldsymbol{\alpha}'\mathbf{w} - \lambda \mathbf{w}'\boldsymbol{\Omega}_r \mathbf{w} - d \|\mathbf{w}\| \quad (5.15)$$

This is similar, but not identical, to Equation (5.12): the norm penalty term is not squared. The same argument can be made to show that the norm and the norm squared are interchangeable, once the penalty constant d is rescaled: $d \|\mathbf{w}\| \simeq (d/\|\mathbf{w}_0\|) \|\mathbf{w}\|^2$, for a $\|\mathbf{w}_0\|$ close to $\|\mathbf{w}\|$ of the final solution.

Robust Factors. We consider another instance of constrained optimization. A recurrent theme in this book is model misspecification. Factor models can be misspecified (both in their factor structure and in their expected returns), but they also offer remedies. Consider the case of an omitted factor. As a special case of misspecification, its effect is to worsen the Sharpe Ratio of the MVO portfolio. In order to reduce the impact, let us consider again an adversarial approach. Assume that there is a hidden factor, whose loadings we do not know, but whose volatility τ is given. We use this as a parameter to quantify the importance of the omitted factor.

The new factor model an additional factor loading \mathbf{v} orthogonal to \mathbf{B} . The covariance matrix is

$$\tilde{\boldsymbol{\Omega}}_r = \boldsymbol{\Omega}_r + \tau^2 \mathbf{v} \mathbf{v}'$$

We solve

$$\begin{aligned} & \max_{\mathbf{w}} \min_{\|\mathbf{v}\| \leq 1} \boldsymbol{\alpha}'\mathbf{w} - \lambda \mathbf{w}'(\boldsymbol{\Omega}_r + \tau^2 \mathbf{v} \mathbf{v}') \mathbf{w} \\ & \max_{\mathbf{w}} \boldsymbol{\alpha}'\mathbf{w} - \lambda \mathbf{w}'(\boldsymbol{\Omega}_r + \frac{\tau^2}{\|\mathbf{w}\|^2} \mathbf{w} \mathbf{w}') \mathbf{w} \\ & \max_{\mathbf{w}} \boldsymbol{\alpha}'\mathbf{w} - \lambda \mathbf{w}'(\boldsymbol{\Omega}_r + \tau^2 \mathbf{I}_m) \mathbf{w} \end{aligned}$$

So, yet again, we are solving an optimization problem with a penalized covariance matrix.

Robust Asset correlations. Another case of adversarial modeling that is expressed as a penalization term. Assume that we estimate the asset correlation matrix terms with some error independent of the asset pair, so that the difference between estimated correlation between and true correlation is at most $|\rho_{i,j} - \hat{\rho}_{i,j}| \leq d$. The adversarial model looks for a solution to the MVO problem, where Nature chooses the covariance matrix with the highest variance compatible with the error bound:

$$\max \mathbf{a}'\mathbf{w} - \lambda\sigma^2 \quad (5.16)$$

$$\text{s.t. } \sigma^2 = \arg \max_{\Delta} \left\{ \mathbf{w}'(\boldsymbol{\Omega}_r + \Delta)\mathbf{w} | [\Delta]_{i,j} \leq d^2[\boldsymbol{\Omega}]_{i,i}[\boldsymbol{\Omega}]_{j,j}, i, j = 1, \dots, n \right\} \quad (5.17)$$

The objective of the nested problem is equivalent to

$$\mathbf{w}'\Delta\mathbf{w} = \sum_{i,j} w_i w_j [\boldsymbol{\Omega}]_{i,i} [\boldsymbol{\Omega}]_{j,j} \rho_{i,j} \quad (5.18)$$

Every term is maximized when $\rho_{i,j} = d^2 \times \text{sgn}(w_i w_j)$, and the objective function value is

$$(\mathbf{w}^*)'\Delta\mathbf{w}^* = d^2 \sum_{i,j} |w_i w_j| [\boldsymbol{\Omega}]_{i,i} [\boldsymbol{\Omega}]_{j,j} \quad (5.19)$$

$$= d^2 \left(\sum_i |w_i| [\boldsymbol{\Omega}]_{i,i} \right)^2 \quad (5.20)$$

$$= d^2 \|\Lambda\mathbf{w}\|_1^2 \quad (5.21)$$

Where Λ is a diagonal covariance matrix whose i th diagonal term is the variance of asset i . Let us plug this back in the original problem:

$$\max \mathbf{a}'\mathbf{w} - \lambda\mathbf{w}'\boldsymbol{\Omega}_r\mathbf{w} - \lambda d^2 \|\Lambda\mathbf{w}\|_1^2 \quad (5.22)$$

And we have yet again a penalization term, which is, in this case, the square of an L1 norm of the portfolio weights. The function $\|\Lambda\mathbf{w}\|_1^2$ is convex, so the optimization problem is tractable. I am summarizing the penalization approaches in the table below:

Robust Covariance Matrix. Consider a different starting point to model robust covariance optimization. We assume that the adversary has a budget for the maximum cumulative squared error of the asset covariances: $\sum_{i,j} [\Delta]_{i,j}^2 \leq d^2$.

This is the same as a bound on the Frobenius norm of the error, $\|\Delta\|_F$. The robust problem formulation is similar to the previous one:

$$\begin{aligned} & \max \mathbf{a}'\mathbf{w} - \lambda\mathbf{w}'\Omega_r\mathbf{w} - \lambda\sigma^2 \\ \text{s.t. } & \sigma^2 = \arg \max_{\Delta} \left\{ \mathbf{w}'\Delta\mathbf{w} \mid \|\Delta\|_F^2 \leq d^2 \right\} \end{aligned}$$

The strategy to solve this problem is similar to previous cases: the adversary maximizes a linear objective function with a norm constraint; see Insight 5.1 for the solution. In this case, $(\sigma^*)^2 = d^2 \|\mathbf{w}\|^2$, yet again, and the problem becomes an MVO with a quadratic penalization term.

Approach	Penalty	Parameter Interpretation
Uncertain Alpha	$\tau^2 \ \mathbf{w}\ ^2$	std.error of $\hat{\alpha}$
Robust Alpha	$d \ \mathbf{w}\ $	max distance $\ \alpha - \hat{\alpha}\ $
Robust Factor	$\lambda\tau^2 \ \mathbf{w}\ ^2$	volatility of a missing factor
Robust Correlations	$\lambda d \ \Lambda\mathbf{w}\ _1^2$	max distance $ \rho_{i,j} - \hat{\rho}_{i,j} $
Robust Covariance	$\lambda d^2 \ \mathbf{w}\ ^2$	max distance $\ \Omega_r - \hat{\Omega}_r\ $

Table 5.1: Summary of the penalties/constraints introduced to address model misspecification. References: Alpha Uncertainty ([Stubbs and Vance, 2005](#)), Adversarial Alpha: ([Pedersen et al., 2021](#)), , Adversarial Correlation: ([Boyd et al., 2016](#)), Adversarial Factor: ([Ceria et al., 2012](#)), Covariance Uncertainty: ([Ledoit and Wolf, 2004](#)).

Exercise 5.1. (30) Define the norm $\|\mathbf{x}\|_{\Lambda,p} := \|\Lambda^{-1}\mathbf{x}\|_p$. Extend Problem (5.12) to this norm. Read ([Olivares-Nadal and DeMiguel, 2018](#)) for additional interpretations of this penalty, and discuss their applicability to real-world settings.

5.3 How Does Estimation Error Affect Sharpe Ratio?

An investor starts with estimates of expected returns and of the covariance matrix⁶. We denote them with $\hat{\alpha}$ and $\hat{\Omega}_r$ respectively. The MVO portfolio is proportional to $\hat{\Omega}_r^{-1}\hat{\alpha}$; the proportionality constant is irrelevant for the Sharpe

⁶The third leg of the trading stool is a model for trading cost. We will cover this in later chapters

Insight 5.3: *The Distinction Between Constraints and Penalties*

Although they can yield the same optimal portfolio, the constrained and penalty version differ in two important ways. The first one is that the shadow price of the constraint is not known before the optimization is run. This means that the solution can be very sensitive to the choice of the right-hand side of the constraint: we don't know the trade-off between constraint limit and optimum value. This is not the case with a penalty: we *set* the price, and the price has often a straightforward interpretation (like the price for risk). In successive optimizations, this price is unchanged making comparisons easier. When the interpretation is clear, penalties are preferable. The second difference is almost a corollary of the first one: in the constrained formulation, we may have no feasible solution, which is, in a loose sense, like saying that the price of the constraint is infinite. This is never the case with a penalized formulation, which is always feasible.

Ratio. The realized Sharpe Ratio, however, is a function of the true expected returns and covariance matrix $\boldsymbol{\alpha}$, Ω_r :

$$\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\Omega}_r) = \frac{\boldsymbol{\alpha}'(\hat{\Omega}_r^{-1}\hat{\boldsymbol{\alpha}})}{\sqrt{(\hat{\Omega}_r^{-1}\hat{\boldsymbol{\alpha}})' \Omega_r (\hat{\Omega}_r^{-1}\hat{\boldsymbol{\alpha}})}}$$

We compare the realized Sharpe Ratio to the best Sharpe ratio, based on the true values of $\boldsymbol{\alpha}$ and Ω_r , given by Equation (4.5):

$$\frac{\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\Omega}_r)}{\text{SR}(\boldsymbol{\alpha}, \Omega_r)}$$

We call this the *Sharpe Ratio Efficiency* (SRE). It is important to study this quantity, because we want to know, at all times, whether we are losing a great deal of performance from inaccurate parameter estimation or large transaction costs. We will ask a few qualitative and quantitative questions, and see how far can the analysis take us⁷.

The first fact is intuitive, but still needs to be proved. Incorrect estimates worsen performance.

⁷Early papers on model estimation error, and the relative impact of alpha and estimation error, are [Michaud \(1989\)](#); [Shephard \(2009\)](#); [Chopra and W.Ziemba \(1993\)](#).

Theorem 5.1. *The Sharpe Ratio Efficiency is less or equal than one, and if it equal to one if and only if $\Omega_r^{-1/2}\alpha$ and $\Omega_r^{1/2}\hat{\Omega}_r^{-1}\hat{\alpha}$ are collinear.*

Proof. The SRE is

$$\frac{SR(\hat{\alpha}, \hat{\Omega}_r)}{SR(\alpha, \Omega_r)} = \frac{\alpha' \hat{\Omega}_r^{-1} \hat{\alpha}}{\sqrt{\hat{\alpha}' \hat{\Omega}_r^{-1} \Omega_r \hat{\Omega}_r^{-1} \hat{\alpha}}} \frac{1}{\sqrt{\alpha' \Omega_r^{-1} \alpha}} \quad (5.23)$$

Let⁸

$$\mathbf{a} := \Omega_r^{-1/2} \alpha \quad (5.24)$$

$$\mathbf{b} := \Omega_r^{1/2} \hat{\Omega}_r^{-1} \hat{\alpha} \quad (5.25)$$

so that

$$\frac{SR(\hat{\alpha}, \hat{\Omega}_r)}{SR(\alpha, \Omega_r)} = \frac{\mathbf{a}' \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

The Sharpe Ratio efficiency is always less than one because of Cauchy-Schwartz inequality⁹, unless $\Omega_r^{-1/2}\alpha$ and $\Omega_r^{1/2}\hat{\Omega}_r^{-1}\hat{\alpha}$ are collinear. \square

5.3.1 The Impact of Alpha Error

It is more useful is to derive lower bounds on performance inefficiency, based on the estimation error of either expected returns or covariance.

We need to introduce a few basic results. Let the norm of a matrix be defined as the operator norm. Define the relative alpha error as

$$\left\| \frac{\alpha}{\|\alpha\|} - \frac{\hat{\alpha}}{\|\hat{\alpha}\|} \right\| \leq \delta_{\text{alpha}}$$

In the Appendix (Section 5.5) I prove the following result:

$$\frac{SR(\hat{\alpha}, \hat{\Omega}_r)}{SR(\alpha, \Omega_r)} \geq 1 - \|\Omega_r^{-1}\|_2 \|\Omega_r\|_2 \delta_{\text{alpha}}^2$$

⁸Let \mathbf{H} be a symmetric positive definite matrix and let $\mathbf{V}\Lambda\mathbf{V}'$ be its Singular Value Decomposition. Define $\mathbf{H}^{1/2} := \mathbf{V}\Lambda^{1/2}\mathbf{V}'$. Then $\mathbf{H}^{1/2}\mathbf{H}^{1/2} = \mathbf{H}$ and $\|\mathbf{H}^{1/2}\|_{\text{op}}^2 = \|\mathbf{H}\|_{\text{op}}$.

⁹Which can be found in almost any linear algebra book. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then $|\mathbf{a}'\mathbf{b}| \leq \sqrt{\mathbf{a}'\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{b}}$, with the equality holding only if $\mathbf{b} = \kappa\mathbf{a}$.

5.3.2 The Impact of Risk Error

Theorem 5.2 (Misspecification of Risk). *If there is $\kappa > 0$ such that*

$$\left\| \Omega_r^{1/2} \hat{\Omega}_r^{-1} \Omega_r^{1/2} - \kappa \mathbf{I} \right\|_2 \leq \delta \quad (5.26)$$

Then

$$\frac{\text{SR}(\hat{\alpha}, \hat{\Omega}_r)}{\text{SR}(\alpha, \Omega_r)} \geq 1 - \frac{2\delta}{\kappa + \delta} \quad (5.27)$$

This formula follows directly from the Equation (5.23). At first sight, what is interesting about this result is how weak it is. Let us consider a few special cases. We define $\mathbf{H} := \Omega_r^{1/2} \hat{\Omega}_r^{-1} \Omega_r^{1/2}$.

1. If the estimated covariance matrix is biased, but uniformly so, i.e., $\hat{\Omega}_r = \kappa \Omega_r$, then $\mathbf{H} = \kappa^{-1} \mathbf{I}$, and there is no efficiency loss. We knew this already from the previous chapter. What happens in practice is that we would deploy a portfolio with the highest Sharpe Ratio, but incorrect volatility.
2. Say, however, that we *really* estimate the covariance matrix incorrectly, so that $\mathbf{H} \not\propto \mathbf{I}$. It can still happen that we have a SRE of one! This will happen if $\hat{\alpha}$ is proportional to an eigenvector of \mathbf{H} with a positive eigenvalue. Say the associated eigenvalue is γ . Then, use directly Equation (5.23)

$$\text{SRE} = \frac{\hat{\alpha}'(\gamma \hat{\alpha})}{\|\hat{\alpha}\|^2} \sqrt{\frac{\|\hat{\alpha}\|^2}{\hat{\alpha}'(\gamma^2 \hat{\alpha})}} = \text{sgn}(\gamma)$$

Even more pathologically, though, this also implies that if our $\hat{\alpha}$ is proportional to an eigenvector with *negative* eigenvalue, then the Sharpe Ratio Efficiency is -1. Incidentally, \mathbf{H} is neither necessarily symmetric nor positive definite, so a negative eigenvalue is indeed a possibility.

3. But, you may argue, this is an exceptional circumstance. Consider a simpler but instructive case. We make the assumption that $\hat{\Omega}_r$ has the same eigenvectors as Ω_r . In other words, the Singular Value Decompositions only differ because of the Singular Values.

$$\begin{aligned} \Omega_r &= \mathbf{U} \Lambda \mathbf{U}' \\ \hat{\Omega}_r &= \mathbf{U} \hat{\Lambda} \mathbf{U}' \end{aligned}$$

so that $\mathbf{H} = \mathbf{U}\Lambda^{1/2}\mathbf{U}'\mathbf{U}\hat{\Lambda}^{-1}\mathbf{U}'\mathbf{U}\Lambda^{1/2}\mathbf{U}' = \mathbf{U}\Lambda\hat{\Lambda}^{-1}\mathbf{U}'$; a great simplification. Denote the eigenvalue ratio $\nu_i := \lambda_i/\hat{\lambda}_i$. What is the lower bound on the SRE in this case? We solve for κ :

$$\begin{aligned}\delta &:= \min_{\kappa} \left\| \mathbf{U}(\Lambda\hat{\Lambda}^{-1} - \kappa\mathbf{I})\mathbf{U}' \right\|_2 \\ &= \min_{\kappa} \sqrt{\max_i (\nu_i - \kappa)^2} \\ &= \frac{1}{2}(\max_i \nu_i - \min_i \nu_i)\end{aligned}$$

For $\kappa^* = (\max_i \nu_i + \min_i \nu_i)/2$. We use these values in Equation (5.27) to obtain

$$\text{SRE} \geq 1 - \frac{\max_i \nu_i - \min_i \nu_i}{\max_i \nu_i} = \frac{\min_i \nu_i}{\max_i \nu_i}$$

Hence the loss in efficiency arises from the fact that we estimate unevenly the volatilities of the eigenvectors of the asset covariance matrix. If we underestimate them (or overestimate them) by the same constant, then we lose nothing, as noted in the first point above. Let us think of an adverse case. Say we estimate all volatilities exactly ($\nu_i = 1$) except for one, which we underestimate by 50%. Then the worst-case loss in Sharpe Ratio can be in 50%.

5.4 Trading Sharpe For Capacity

Exercise 5.2 (The trade-off between sharpe ratio and absolute returns). (30)
Consider the problem

$$\begin{aligned}V(r) &:= \min \mathbf{w}' \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w} \\ \text{s.t. } A' \mathbf{w} &\leq c \\ \alpha' \mathbf{w} &\geq r\end{aligned}\tag{5.28}$$

and define the associated Sharpe Ratio $SR(r) := r/\sqrt{V(r)}$. This is a generalization of the dual of the original unconstrained problem

$$\begin{aligned}\tilde{V}(r) &:= \min \mathbf{w}' \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w} \\ \alpha' \mathbf{w} &\geq r\end{aligned}$$

One special instance is the problem

$$\begin{aligned} V(r) := \min \mathbf{w}' \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w} \\ \text{s.t. } \sum_i |\mathbf{w}_i| \leq 1 \\ \alpha' \mathbf{w} \geq r \end{aligned}$$

In this instance, r is a minimum required return on Gross Market Value of the portfolio and $\sqrt{V(r)}$ is the smallest achievable volatility.

1. Prove that $V(r) \geq \tilde{V}(r)$ for all $r \in \mathbb{R}$;
2. Prove that $\sqrt{V(r)}$ is increasing and convex;
3. Prove that $\sqrt{\tilde{V}(r)}$ is linear and increasing;
4. Prove that $SR(r)$ is non-increasing and trivially bounded by \widetilde{SR} .

This exercise shows that a high Sharpe Ratio can be traded off for higher payoffs, included higher returns on GMV. For example, while the MVO portfolio may have a high Sharpe but low return on GMV, a constrained version can achieve higher return, but at the cost of a lower risk-adjusted performance.

5.5 ★Appendix: Theorems on Sharpe Efficiency Loss

These theorems are informally introduced in Section 5.3.

We recall that

$$\|\mathbf{H}^{-1}\|_2^{-1} \|\mathbf{x}\| \leq \|\mathbf{Hx}\| \leq \|\mathbf{H}\|_2 \|\mathbf{x}\|$$

and

$$\|\mathbf{Hy}\| \leq \|\mathbf{Hx}\| + \|\mathbf{H}(\mathbf{y} - \mathbf{x})\| \leq \|\mathbf{Hx}\| + \|\mathbf{H}\|_2 \|\mathbf{x} - \mathbf{y}\|$$

so that

$$| \|\mathbf{Hx}\| - \|\mathbf{Hy}\| | \leq \|\mathbf{H}\|_2 \|\mathbf{x} - \mathbf{y}\|$$

Also, use the cosine rule:

$$\begin{aligned} \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\|^2 &= 2 \left(1 - \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \right) \\ \Rightarrow \frac{\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Omega}}_{\mathbf{r}})}{\text{SR}(\boldsymbol{\alpha}, \boldsymbol{\Omega}_{\mathbf{r}})} &= \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \\ &= 1 - \frac{1}{2} \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\|^2 \end{aligned}$$

where \mathbf{a}, \mathbf{b} are defined by Equations (5.24) and (5.25).

Lemma 5.1. Let \mathbf{H} be symmetric positive-definite, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and

$$\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\| \leq \delta$$

Then

$$\left\| \frac{\mathbf{Hx}}{\|\mathbf{Hx}\|} - \frac{\mathbf{Hy}}{\|\mathbf{Hy}\|} \right\| \leq 2 \min\{\|\mathbf{H}\|_2 \|\mathbf{H}^{-1}\|_2 \delta, 1\}$$

Proof. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$.

$$\begin{aligned} \left\| \frac{\mathbf{Ha}}{\|\mathbf{Ha}\|} - \frac{\mathbf{Hb}}{\|\mathbf{Hb}\|} \right\| &= \left\| \frac{\|\mathbf{Hb}\| \mathbf{Ha} - \|\mathbf{Ha}\| \mathbf{Hb}}{\|\mathbf{Ha}\| \|\mathbf{Hb}\|} \right\| \\ &= \left\| \frac{\|\mathbf{Hb}\| \mathbf{H}(\mathbf{a} - \mathbf{b}) - (\|\mathbf{Ha}\| - \|\mathbf{Hb}\|) \mathbf{Hb}}{\|\mathbf{Ha}\| \|\mathbf{Hb}\|} \right\| \\ &\leq \frac{1}{\|\mathbf{Ha}\|} (\|\mathbf{H}(\mathbf{a} - \mathbf{b})\| + |\|\mathbf{Ha}\| - \|\mathbf{Hb}\||) \\ &\leq \frac{1}{\|\mathbf{Ha}\|} (\|\mathbf{H}(\mathbf{a} - \mathbf{b})\| + \|\mathbf{H}\|_2 \|\mathbf{a} - \mathbf{b}\|) \\ &\leq \frac{1}{\|\mathbf{Ha}\|} (\|\mathbf{H}\|_2 \|(\mathbf{a} - \mathbf{b})\| + \|\mathbf{H}\|_2 \|\mathbf{a} - \mathbf{b}\|) \\ &\leq \frac{2}{\|\mathbf{Ha}\|} \|\mathbf{H}\|_2 \|(\mathbf{a} - \mathbf{b})\| \\ (\mathbf{a} := \frac{\mathbf{x}}{\|\mathbf{x}\|}, \mathbf{b} := \frac{\mathbf{y}}{\|\mathbf{y}\|}) &\leq \frac{2}{\left\| \mathbf{H} \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\|} \|\mathbf{H}\|_2 \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\| \\ &\leq 2 \|\mathbf{H}\|_2 \|\mathbf{H}^{-1}\|_2 \delta \end{aligned}$$

□

This bound is tight, up to a constant. For an example, consider the case of diagonal¹⁰ $\mathbf{H} := \text{diag}(\lambda_1, \dots, \lambda_n)$, $\mathbf{x} := \epsilon \mathbf{e}_1 + \mathbf{e}_n$, $\mathbf{y} := \mathbf{e}_n$, with $\epsilon \leq \lambda_n/\lambda_1$. We have

$$\begin{aligned} \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\| &\leq \sqrt{\frac{3}{2}}\epsilon =: \delta \\ \left\| \frac{\mathbf{Hx}}{\|\mathbf{Hx}\|} - \frac{\mathbf{Hy}}{\|\mathbf{Hy}\|} \right\| &= \left\| \frac{\lambda_n \mathbf{e}_n + \epsilon \lambda_1 \mathbf{e}_1}{\sqrt{\lambda_n^2 + \epsilon^2 \lambda_1^2}} - \mathbf{e}_n \right\| \\ &= \sqrt{\left(\frac{\lambda_n^2}{\sqrt{\lambda_n^2 + \epsilon^2 \lambda_1^2}} - 1 \right)^2 + \frac{(\epsilon \lambda_1)^2}{\lambda_n^2 + \epsilon^2 \lambda_1^2}} \\ &\geq \sqrt{2 \frac{-\lambda_n^2 (\epsilon \lambda_1 / \lambda_n)^2 / 2 + (\epsilon \lambda_1)^2}{\lambda_n^2 + (\epsilon \lambda_1)^2}} \\ &\geq \sqrt{\frac{(\epsilon \lambda_1)^2}{\lambda_n^2 + (\epsilon \lambda_1)^2}} \\ &\geq \frac{1}{\sqrt{2}} \frac{\lambda_1}{\lambda_n} \epsilon \\ &= \frac{1}{\sqrt{3}} \|\mathbf{H}\|_2 \|\mathbf{H}^{-1}\|_2 \delta \end{aligned}$$

Theorem 5.3 (Misspecification of alpha). *If*

$$\left\| \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|} - \frac{\hat{\boldsymbol{\alpha}}}{\|\hat{\boldsymbol{\alpha}}\|} \right\| \leq \delta$$

Then

$$\frac{\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Omega}}_{\mathbf{r}})}{\text{SR}(\boldsymbol{\alpha}, \boldsymbol{\Omega}_{\mathbf{r}})} \geq 1 - 2 \|\boldsymbol{\Omega}_{\mathbf{r}}^{-1}\|_2 \|\boldsymbol{\Omega}_{\mathbf{r}}\|_2^{-1} \delta^2$$

Proof. From Lemma 5.1:

$$\begin{aligned} \left\| \frac{\boldsymbol{\Omega}_{\mathbf{r}}^{-1/2} \boldsymbol{\alpha}}{\|\boldsymbol{\Omega}_{\mathbf{r}}^{-1/2} \boldsymbol{\alpha}\|} - \frac{\boldsymbol{\Omega}_{\mathbf{r}}^{-1/2} \hat{\boldsymbol{\alpha}}}{\|\boldsymbol{\Omega}_{\mathbf{r}}^{-1/2} \hat{\boldsymbol{\alpha}}\|} \right\| &\leq 2 \|\boldsymbol{\Omega}^{-1/2}\|_2 \|\boldsymbol{\Omega}^{1/2}\|_2 \delta \\ &= 2 \sqrt{\|\boldsymbol{\Omega}_{\mathbf{r}}^{-1}\|_2 \|\boldsymbol{\Omega}_{\mathbf{r}}\|_2} \delta \end{aligned}$$

¹⁰We use the notation $\mathbf{e}_1, \dots, \mathbf{e}_n$ for the standard basis in \mathbb{R}^n .

Then

$$\begin{aligned} \frac{\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Omega}}_r)}{\text{SR}(\boldsymbol{\alpha}, \boldsymbol{\Omega}_r)} &= 1 - \frac{1}{2} \left\| \frac{\boldsymbol{\Omega}_r^{-1/2} \boldsymbol{\alpha}}{\|\boldsymbol{\Omega}_r^{-1/2} \boldsymbol{\alpha}\|} - \frac{\boldsymbol{\Omega}_r^{-1/2} \hat{\boldsymbol{\alpha}}}{\|\boldsymbol{\Omega}_r^{-1/2} \hat{\boldsymbol{\alpha}}\|} \right\|^2 \\ &\geq 1 - \|\boldsymbol{\Omega}_r^{-1}\|_2 \|\boldsymbol{\Omega}_r\|_2 \delta^2 \end{aligned}$$

□

Theorem 5.4 (Misspecification of Risk). *If there is $\kappa > 0$ such that*

$$\left\| \boldsymbol{\Omega}_r^{1/2} \hat{\boldsymbol{\Omega}}_r^{-1} \boldsymbol{\Omega}_r^{1/2} - \kappa \mathbf{I}_n \right\|_2 \leq \delta \quad (5.29)$$

Then

$$\frac{\text{SR}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Omega}}_r)}{\text{SR}(\boldsymbol{\alpha}, \boldsymbol{\Omega}_r)} \geq 1 - \frac{2\delta}{\kappa + \delta} \quad (5.30)$$

Proof. Let $\mathbf{H} := \boldsymbol{\Omega}_r^{1/2} \hat{\boldsymbol{\Omega}}_r^{-1} \boldsymbol{\Omega}_r^{1/2}$ and let $\hat{\boldsymbol{\alpha}} := \boldsymbol{\Omega}_r^{-1/2} \boldsymbol{\alpha}$. Using this notation, the SRE Equation (5.23) and condition (5.29) are

$$\begin{aligned} \frac{\text{SR}(\boldsymbol{\alpha}, \hat{\boldsymbol{\Omega}}_r)}{\text{SR}(\boldsymbol{\alpha}, \boldsymbol{\Omega}_r)} &= \frac{\hat{\boldsymbol{\alpha}}' \mathbf{H} \hat{\boldsymbol{\alpha}}}{\|\hat{\boldsymbol{\alpha}}\|^2} \sqrt{\frac{\|\hat{\boldsymbol{\alpha}}\|^2}{\hat{\boldsymbol{\alpha}}' \mathbf{H}^2 \hat{\boldsymbol{\alpha}}}} \\ \|\mathbf{H} - \kappa \mathbf{I}_n\|_2 &\leq \delta \end{aligned}$$

Let $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n$ eigenvalues of \mathbf{H} . The condition $\|\mathbf{H} - \kappa \mathbf{I}_n\|_2 \leq \delta$ is equivalent to $|\lambda_i - \kappa| \leq \delta$ for all $i = 1, \dots, n$.

$$\begin{aligned} \frac{\hat{\boldsymbol{\alpha}}' \mathbf{H} \hat{\boldsymbol{\alpha}}}{\|\hat{\boldsymbol{\alpha}}\|^2} &\geq \lambda_n \geq \kappa - \delta \\ \frac{\hat{\boldsymbol{\alpha}}' \mathbf{H}^2 \hat{\boldsymbol{\alpha}}}{\|\hat{\boldsymbol{\alpha}}\|^2} &\leq \lambda_1^2 \leq (\kappa + \delta)^2 \\ \Rightarrow \frac{\hat{\boldsymbol{\alpha}}' \mathbf{H} \hat{\boldsymbol{\alpha}}}{\|\hat{\boldsymbol{\alpha}}\|^2} \sqrt{\frac{\|\hat{\boldsymbol{\alpha}}\|^2}{\hat{\boldsymbol{\alpha}}' \mathbf{H}^2 \hat{\boldsymbol{\alpha}}}} &\geq \frac{\kappa - \delta}{\kappa + \delta} = 1 - \frac{2\delta}{\kappa + \delta} \end{aligned}$$

□

The population covariance matrix is not known. A one-period proxy for the covariance matrix is \mathbf{rr}' . The next lemma presents a closed-form expression of the left-hand side of Inequality (5.29) when $\boldsymbol{\Omega}_r$ is replaced by this proxy.

Chapter 6

Evaluating Alpha

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com

Xtwitter: @_paleologo (DM me)

LinkedIn: <https://www.linkedin.com/in/gappy/> (connect, then message)

The task of estimating factor models and testing alphas for systematic strategies usually involves reusing the same historical data. This is somewhat unintuitive. One of the defining features of the past 40 year has been the increased recording of new data sets and its dissemination. Investment firms have budgets of tens of billions of dollars allocated to the purchase of market and alternative data, and to the bespoke collection of data (e.g., via web scrapes). And yet, the characteristics of traded assets, and specifically of companies, do not change on a minute-by-minute basis; and investment strategies with relatively long holding times – of the order of a day or longer – do not necessarily employ tick-by-tick data. If we record prices for a broad local investment universe at five-minute intervals, we collect 60 millions numeric data points¹; including a security identifier and a time stamp, the required storage is of the order of gigabytes. History is not replaceable, and sometimes is not deep. Not replaceable, in the sense that it is not easy to produce a simulated version of the past that provably reproduces all of its features. Not deep, in the sense that we do not live in a stationary world. The pace at which the real world outside of finance changes is breathtaking and accelerating; and markets are a timid reflection of it. Not even taking such change in to account, the introduction of new technology, of new market microstructure designs, of new regulations,

¹Assuming 6.5 trading hours, 252 trading days and 3,000 stocks.

and the ongoing collective learning process of all market participants make the investing world of five years ago very different from today’s. The fact that we have to rely on historical data poses a major challenge. We cannot design experiments. Our studies are observational and repeated. Financial practitioners do not have a shared protocol for experimental analysis. Even if we had one, it is far from obvious that it is the correct one. Well-established disciplines like Medicine and Psychology had shared experimental practices accompanied by experimental design, and yet they have undergone a reevaluation when their practitioners found that most of their results are not replicable (Ioannidis, 2005; Open Science Collaboration, 2015).

This poses a few challenges for us modelers. We have a very large number of signals types, which themselves depend on continuous tuning parameters. This is similar to the situation faced by biostatisticians, who are faced with tens of thousands of simultaneous tests in the form of responses from DNA microarray Dudoit et al. (2003); Huang et al. (2009). The details are quite different. The “response variable” for a DNA microarray is usually discrete (“polytomous”), and responses are uncorrelated or weakly correlated. In quantitative finance, the response variable (be it return or Sharpe) is continuous, and signal correlation plays a decisive role.

This chapter has four sections. First, we list some basic best practices for data preparation and usage. Second, we describe some common backtesting practices and critique them. The third section summarize some important hypothesis testing approaches. The last section is entirely devoted to describing a new backtesting protocol, which offers several advantages over the previous ones: it gives finite-sample uniform probabilistic bounds on the Sharpe Ratio of a large set of strategies.

6.1 Backtesting Best Practices

We review a few best practices for backtesting. They do not originate from some comprehensive theory for them. Unlike Athena, who was born fully formed from the mind of Zeus, it is an ever-incomplete, occasional shallow body of knowledge that has formed by experimentation. Some references covering the approximately the same ground are Arnott et al. (2019); Wang et al. (2014); López de Prado (2020).

Data Sourcing. High-quality data are essential to backtesting, and the search for better data is a never-ending task for a researcher. There are four broad

areas of concern. The first one is data sourcing. There are multiple vendors offering similar data. When comparing them, ask the following questions:

- *Definition and Interpretation.* Perhaps the first and most important question, not only in data sourcing, but in quantitative investing, is *what do the data mean?* What is the exact definition of the data collected? What are their physical units? If the dataset is money-related, it should be unambiguous what is the reference currency (or currencies, for exchange rates). If the data is flow-related (i.e., measuring rates over time), the time unit should be defined. A surprising number of mistakes happens because of unit conversion errors.
- *Provenance.* Where are the data coming from? Does the vendor collect the data themselves (e.g., via web scraping, or internet traffic); more often the vendor serves as an intermediary between a data originator and the client. In the former case, what is the collection criterion? Does the vendor sample data or collects the data exhaustively? Is the population sampling methodology sound? In the latter case, who is originating the data? Are they trustworthy?
- *Completeness.* This leads to the second question: completeness. Are there data that are obviously missing from the data set such as, for example, intermittently missing prices? Are there data that are non-obviously missing, such, as for example, unrecorded consumer credit card transactions? Some of these questions can be answered by performing exploratory analysis on the data themselves, others need to be addressed with the vendor.
- *Quality Assurance.* How does the provider ensure that the data it receives or collects are consistently of good quality? Does it have checks for change points in the data characteristics?
- *Point-in-time vs. Restated Data.* Does the provider offer data collected as of a certain date, without changing them at a later date, based on corrections and company updates? This is an instance of data leakage, which we will cover this in more detail later.
- *Transformations.* Data are almost always transformed by the vendor in some instances. Examples are: imputation of missing data; winsorization and censoring of outliers; end-of-period price calculations (last transaction,

mid bid-ask price, weighted average). These transformations should be documented, evaluated, and if possible verified by the research analyst.

- *Exploring Alternatives and complements.* Always ask the following common-sense questions: can we obtain better data, in three dimensions. First, are there providers offering larger coverage for the same data set (more securities at any point in time, deeper history, more frequent data)? Second, are there providers with better data? For example, if data are collected from broker-dealers, the alternative provider has an agreement with a larger number of participating contributors. Third, can we obtain complementary data? These are data sets that used jointly with the original data set, greatly increase its original value. For example, we may obtain transactional data that help us estimate short-term revenues of a company, in addition to data that give us a good estimate of their costs.

Research Process. Every researcher has their own research process. This is part of their competitive advantage; it's indeed part of *what they are*, of thoughts and learned lessons accumulated over a lifetime of experiences and of studying. It would be futile to superimpose the author's overall research philosophy to that of the reader², just in a few pages. *However*, there are a few steps that are uncontroversial, and are part of basic hygiene. Consider these as the precept to never leave home without wearing underwear.

- *Data Leakage.* The first recommendation is to avoid *data leakage*. The definition of data leakage is the presence in the training data, the data available up to time t , of information contained in the target, i.e., returns in periods $t + 1$ and later. The reference rule is to never use data in a backtest on a certain date that we are not able to use in production today. Detecting data leakage is more art than science, and it requires both a deep knowledge of the data (see above) and of the problem at hand. Below are a few examples.
 - *Survivorship bias.* If we backtest the performance of a strategy over an extended period of time, considering only the stocks that have continuously traded during this period, i.e., the surviving stocks at the end of the backtest, we are subject to survivorship bias. Stocks are most often delisted because they experience large losses, trade at

²Although, you may argue, this whole book is an exposition of my investment philosophy. Point taken, to an extent. I am providing some building blocks, and you are reshaping and assembling them into something sensible.

low unit prices, become illiquid and do not meet the criteria for being listed on exchange. Removing them biases the investment sample toward outperformers with different characteristics from the broader investment universe at a point in time. For example, the Survivors' liquidity, momentum, and size are larger than the universe. This is the simplest and most impactful instance of data leakage. The remedy to this issue is to: a) employ an objective methodology for inclusion that is applicable at any point in time; b) specify a realistic and conservative rule in the backtest for the event of a delisting. For example, one could assume that the entire investment is written off. Note that the methodology in a) should be specified *before* backtests are initiated. Changing the inclusion rules based on the result of backtest is also an instance of data leakage, and it should be avoided. Criteria for inclusion are indeed not straightforward to specify. A common recommendation is to use a prespecified investment universe represented by a benchmark, like Russell 1000, Russell 3000, MSCI benchmarks, or commercial factor models investment universes. Note that benchmark components are always announced before ("announcement date") the effective addition date ("reconstitution date"), and the returns of the stocks are affected by the announcement on the inclusion. In your backtest, you may want to capture this information, in order to assess how much of the performance of your quantitative strategies is affected by recent changes in the investment universe.

- *Financial Statements.* Financial statement information for a given quarter or year should be included in the backtesting data on the day (or the day after) of their public release, not on the last quarter to which the data refer, or a fixed number after the reference quarter.
- *Point-in-Time Data.* Financial data used in the backtest on a given date should always be the most current data available *as of that date*. If a 10Q (a quarterly financial report) is restated because of material error, the backtest should not reflect that. The strategy must be tested by allowing the presence of error in its input data.
- *Price Adjustments.* Shares are regularly split (or reverse-split) into multiple shares. The price of the split share is adjusted accordingly. This occurs when the stock appreciates to the point one share becomes so expensive that it prevents investors from being able to buy it. In order to compute historical returns across long time series, prices and dividend are usually split-adjusted. This introduces a complication. A

low stock price in the distant past indicates that the shares have been split several times in the future, likely because of high returns. The price becomes informative of future performance. The recommended remedy is to use adjusted prices only for return calculation. For feature generation, use as-of-date prices.

- *Missingness.* In certain cases (mostly unstructured data instances), data points are missing either because they were not made available as-of-date or because they contain sensitive information and were redacted. In the latter case, missingness may be suffering from look-ahead bias and is informative of future returns.
- *Avoidable Mistakes.* The amount of silly mistakes (in hindsight) that experienced, effective researchers make never ceases to amaze³. For example, a stock characteristic available in a data set had high Information Coefficient. Upon further investigation, it was a stock split conversion factor (see the previous bullet point). As another example, an erroneous t vs. $t + 1$ convention error caused a research to include the next-day return in a three-month momentum factor definition, also causing a false positive.
- *Strategy Development.* There are some qualitative recommendations that, while missing a solid foundation, are hard to argue against.
 - *Have a theory (if you can).* It is preferable to have a theory for every anomaly, and if we pre-registered the predictions of the theory before the backtest. For example, in their paper on quality, [Asness et al. \(2019\)](#) propose a theory guiding the development of the factor; so do [Frazzini and Pedersen \(2014\)](#) when they analyze the beta anomaly. With a theory as a guide, it is easier to choose a security characteristic among many, therefore reducing the number of strategies being tested; it is possible to interpret the result and believe in it more, and it is possible to critique and revise the characteristic, which is maybe not desirable (it would be nice if we got it right the first time) but necessary.
 - *Enforce reproducibility.* Document All Your Strategies And make sure you can reproduce and rerun them at any time.

³One of the implicit themes of this book is that investors are stupider than they think, and that authors of investing books are even dumber than investors.

- *Use as much as possible the same setting in backtesting and in production.* By this we mean that we should use the same point-in-time data, but also the same optimization formulations, the same market impact model, and the same codebase.
- *Calibrate the market impact model.* When we perform a backtest the market impact model has a “descriptive” role. It estimates the losses in efficiency from actual trading. It is not possible, however, to verify the realized market impact on historical data. In order to align backtest performance to live one, it is important not to take a market impact model at face value, especially one provided by a vendor or a partner. Instead, calibrate its parameters against live performance of the current version of your strategy, so that realized and backtested PnL of your current strategy overlap as much as possible.
- *Include Borrow Costs.* As part of the effort to align production and simulated PnL, one should take into account borrow costs for shorted securities, since they can have a material impact on the profitability of the short side. This has challenges. Historical borrow rates are not readily available historically. The researcher may have to approximate them, or predict them on the basis of security characteristics. Another complication, albeit less impactful, is the tax treatment of dividends. When they are received by the investor, they are subject to taxation. When the investor is short the security, the treatment of dividends is more complex. In practice, tax dividend treatment is complex, and does not make a material difference in backtests, so it is not modeled. Be aware of it, in case you see discrepancies between accounting and simulated PnL.
- *Define beforehand the backtesting protocol.* A backtesting protocol is the sequence of actions and decisions that lead to assessing the performance of a strategy. It is the subject of the next section. For the sake of this list of folk wisdom precepts, it is sufficient to say that the backtesting protocol should be changed a) rarely, b) for a good reason, and c) if it changes, you should rerun and re-evaluate all your strategies under the new protocol.
- *Define beforehand the dataset being used.* If dataset selection is seen as part of the backtesting protocol, the heuristic follows from the previous point. The difference however is that new data become available every day, both in the form of live data, and extensions to

historical dataset. Researchers may be prone to include datasets that confirm their findings, and ignore those that do not. Ignoring new data would be suboptimal, and including them selectively may lead to the wrong conclusions. Use your judgement and research integrity, which no theorems can help.

6.2 The Backtesting Protocol

6.2.1 Cross-Validation and Walk Forward

Evaluating trading strategies bears similarities with statistical model selection (Hastie et al., 2008). We have a family of strategies (in Statistics, a family of models), and a performance measure, such as Sharpe Ratio or return on GMV. The strategies themselves may depend on several parameters. Two evaluation schemes are most common. The first one is cross-validation (Hastie et al. (2008), Ch. 7, and Mohri et al. (2018), Ch. 4). The available data is split into a *training dataset* and a *holdout (or validation) dataset*. Sometimes, based on the estimated time-dependence in the time series, the training and holdout samples are separated by a “buffer” dataset. The cross-validation is split into K equal-

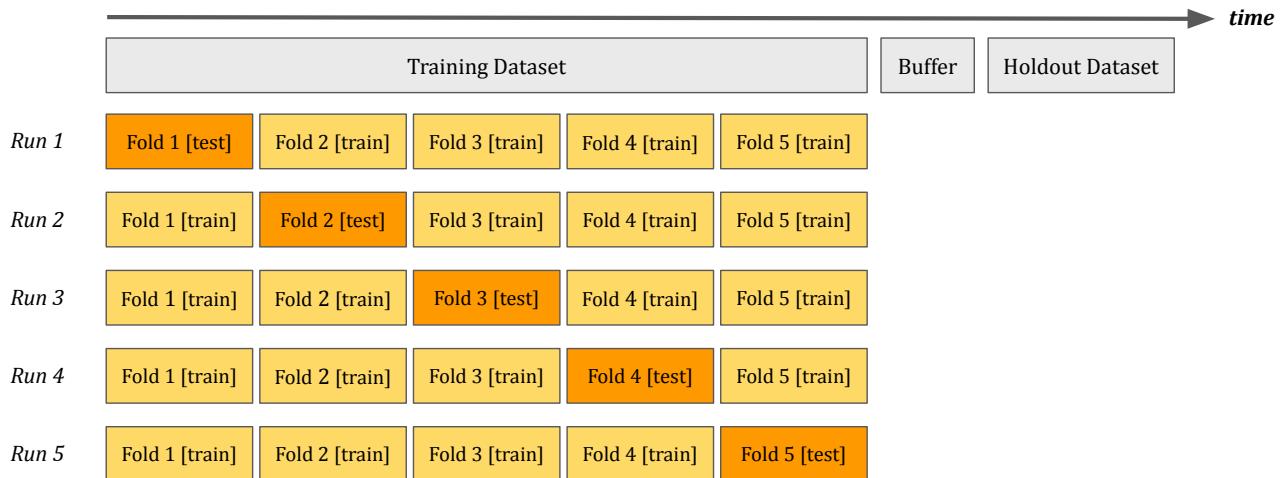


Figure 6.1: A scheme of the cross-validation procedure. Darker boxes are validation folds, while lighter boxes are training folds.

sized samples (“folds”). Then we perform K estimation-evaluation exercises. The parameters are estimated on each of the possible combinations of $K - 1$ folds, and the performance of the model is evaluated of the remaining fold using the optimized parameter; see Figure 6.1. Then estimate the cross-validation

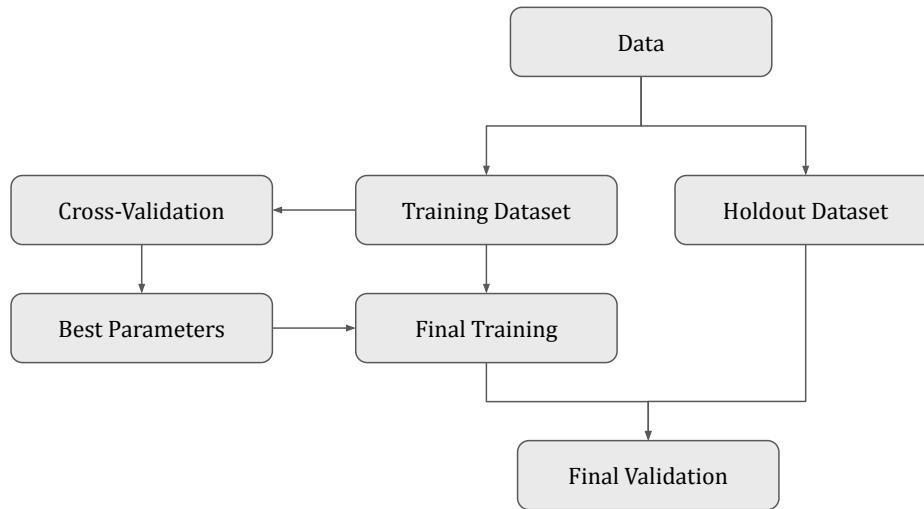


Figure 6.2: A scheme of the cross-validation procedure. Data are split into two sets. Cross-validation is performed on the first one (training dataset), to estimate the expected performance of a strategy. The model is then optimized on the entire training dataset, and validated on the second one (validation dataset).

performance as the average of the single-run performances. Finally, performance is checked against the holdout sample; a scheme is shown in Figure 6.2. There are several contraindications to using cross-validation for financial applications. First, the samples are not independent. The time dependence is reflected in the returns themselves. We know that serial dependence of return is weak and has short memory, while volatility dependence is strong and has long memory. For certain time series, it is possible to remedy this by keeping the order intact in the training folds and the errors are serially uncorrelated [Bergmeir et al. \(2018\)](#); [Cerqueira et al. \(2023\)](#). This is not the only issue faced in financial applications, however. For example, consider the inclusion of security momentum as a predictor. This characteristic uses past returns. Now, if the validation fold precedes temporally the training fold, these past return are in the validation fold and we are incurring in a typical instance of data leakage: the predictors directly contain information about the target. This is just an obvious example; there are subtle ones as well. For example, we could use forward earnings forecast as a predictor. But forward earnings are usually produced by analysts, who base their judgement on past returns. Like momentum, we may have leaked target data into the training set. Besides the temporal dependencies, there is another practical objection to K -fold cross-validation. In their influential book, [Hastie et al. \(2008\)](#) (Section 7.10.2) make a forceful case that the model should *entirely*

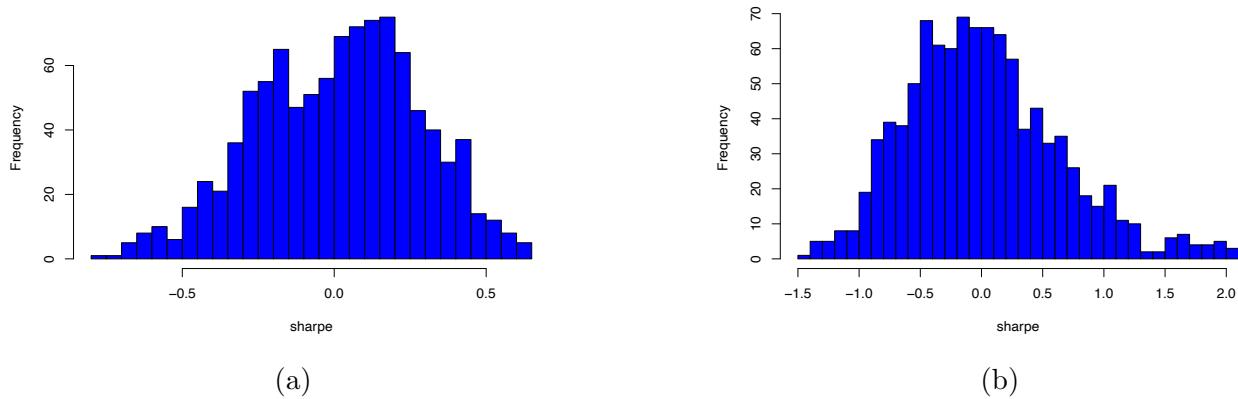


Figure 6.3: Cross-validated Sharpe for (a) Scenario 1, (b) Scenario 2.

be selected by cross-validation. Predictive variables (be they alphas or factors, in our framework) should not be screened in advance. This is not never done in practice. The predictiveness of signals or fully-fledged strategies is tested separately. Perform cross-validation enough times on different classes of models, and you will inevitably obtain favorable results. The holdout dataset is meant to serve as a final check against this “fishing expedition” (Cochrane, 2005). And yet, when the number of raw signals runs in the millions, it is inevitable to cycle through several refinements and model revisions, so that the holdout sample performance becomes just another variable to be optimized, instead of a performance check to be run only once.

An example may help illustrate the perils of cross-validation. We have $N = 1000$ assets. We simulate iid asset returns with $r_{t,i} \sim N(0, \sigma^2)$, with $\sigma = 0.01$. We introduce p random asset characteristics, also iid drawn at random: $[\mathbf{B}]_{i,j} \sim N(0, 1)$. These random features are by design not predictive of returns. The backtest consists in a 5-fold cross-validation, to estimate the performance of the predictors. In each run, we select the best performing factor, based on in-sample IC, and then compute the IC on the test fold. Then we report the average cross-validated IC. We repeat the process on 1000 simulated data sets. Below are the results for two scenarios:

1. The first one is the “many periods, few predictors” case: we set $T = 5000$ (twenty years of daily data) and $p = 2$; two predictors because one would have felt too lonely.
2. The second one is the “not many periods, more predictors” case: we set $T = 1250$ (five years of daily data) and $p = 500$; not nearly as many as we

meet in practice.

The frequency histograms of the simulations are shown in Figure 6.3. Some summary statistics of the simulations are shown in Table 6.1. The averages are close to zero in both case, with a much larger standard deviation for the many factor-case. The percentage of samples whose Sharpe Ratio passes the 1% significant level is shown in the last column of the table⁴.

T	p	Mean(SR)	Stdev (SR)	% passing
5000	2	0.07	0.6	1.2
1250	500	0.04	1.4	19

Table 6.1: Frequency histograms for the two simulated scenarios; the conversion IC to Sharpe Ratio is $SR = IC\sqrt{252N}$.

A remedy to the data leakage issues arising in cross-validation is *walk-forward backtesting* (Pardo, 2007). In this scheme, we use historical data up to period t and target returns for period $t + 1$. The scheme is as close as possible to the production process. It addresses two drawbacks of cross-validation for time-series – serial dependence and risk of data leakage – and it also augments naturally the data set with the arrival of new data. Finally, it is naturally adaptive: it fine-tunes parameters as the environment changes. These advantages are complementary to cross-validation. As a result, it is often the case that signals, or simplified strategies, are first tested using cross-validation, and then tested “out of sample” in a walk-forward test. This is not ideal, however, since it has an opportunity cost caused by the delay in running the strategy in production. Walk-forward has an additional important drawback: it uses less training data than cross-validation. When the set of models and parameters is very large, this limitation could be very severe. On the other side, when the model has been identified, and only a few parameters need to be optimized, then this drawback becomes negligible. Two additional trading settings in which walk-forward does not suffer from data limitations are when a) data are plenty. This is the case of high-frequency trading; b) data are very non-stationary. This is the case, to some extent, of every trading strategy, and this very fact suggests that walk-forward backtesting is, in any event, a necessary step in the validation of a strategy, and in its preparation for production.

Summing up, neither cross-sectional nor walk-forward schemes are without flaws. Ideally, we would like a protocol with the following features.

⁴This is the percentage of simulation samples for which the condition $SR > 2.3\sqrt{(1 + SR^2)/T}$

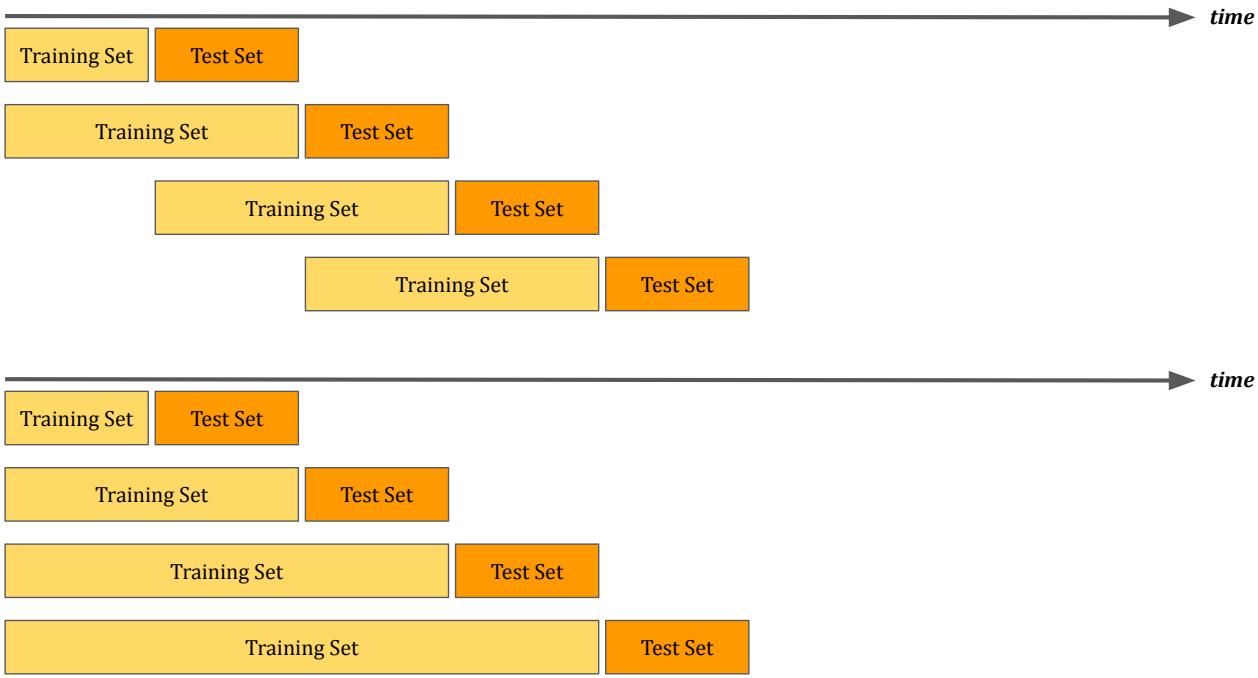


Figure 6.4: Two common walk-forward schemes. The top one uses fixed-length training data, thus preserving the estimation procedure. The bottom one uses all the available data up to certain epoch, possibly weighting data differently based on the interval from the decision epoch.

1. *non-anticipative/immune from data leakage;*
2. *taking into account serial dependency;*
3. *using all data;*
4. *allowing for multiple testing of a very large number of signals;*
5. *Providing a rigorous decision rule.*

Walk forward meets the first two requirements; cross-validation meets the third. Neither meet the last two. The next section introduces a novel backtesting protocol, the *Rademacher Anti-Serum* (RAS) (in short, RAS), which meets these requirements.

6.3 The Rademacher Anti-Serum

6.3.1 Setup

We will be concerned with testing the performance of strategies and signals. The former is simply the time series of the walk-forward simulated returns z-scored by their predicted volatility, so that their average equals the empirical sharpe ratio for strategy n , which we denote by $x_{t,n}$. In this respect, the protocol is similar to walk-forward. When we test signals, we instead consider the Information Coefficient for the signal n at time t . The definitions are below⁵:

$$\begin{aligned} x_{t,n} &:= \frac{\mathbf{w}'_{t,n} \mathbf{r}_t}{\sqrt{\mathbf{w}'_{t,n} \boldsymbol{\Omega}_t \mathbf{w}_{t,n}}} && (\text{Sharpe Ratio}) \\ x_{t,n} &:= \frac{\boldsymbol{\alpha}'_{t,n} \boldsymbol{\epsilon}_t}{\|\boldsymbol{\alpha}_{t,n}\| \|\boldsymbol{\epsilon}_t\|} && (\text{Information Coefficient}) \end{aligned}$$

We also denote $x_{t,n}$ both instances; the interpretation will be clear from the context. In either case, the primitive dataset needed for the analysis is a $T \times N$ matrix \mathbf{X} . Rows denote observations as of a certain timestamp and columns denote strategies, whose set we denote S . For notational simplicity, the t th row of \mathbf{X} is denoted by \mathbf{x}_t , and the n th column \mathbf{x}^n . In the following we make the important assumption that the random vectors \mathbf{x}_t are iid random variable, drawn from a common probability distribution P . We justify this on two ground. The first one is empirical. Serial dependence is small for returns observed at daily frequencies or lower⁶. The second one is that our framework can be extended to the case of time-dependent returns, at a price of weaker, asymptotic results. We recommend to plot the autocorrelation plot of the univariate plot of the series \mathbf{x}^n . If there is sizable autocorrelation up to lag s , then replace the original time series with $\lfloor N/s \rfloor$ non-overlapping, contiguous averages of blocks $(x_{1+ks,n}, \dots, x_{(k+1)s,n})$. We employ the following notation. We let the joint distribution of \mathbf{x}_t be P . Let $D = \bigotimes_{i=1}^N P$ be the joint probability distribution on the space of $T \times N$ matrices in which the element $\mathbf{x}_t \sim P$ has independent, identically distributed (iid) rows, each drawn from P .

We also define a “bootstrap distribution”, i.e., a probability distribution $P^*(\mathbf{X})$. An element \mathbf{Y} is drawn from D^* by sampling with replacement T rows of the data matrix \mathbf{X} . For notational simplicity we drop the subscript \mathbf{X} since we only deal with only one data matrix.

⁵For definitions and uses of $\boldsymbol{\alpha}$, see Sections 3.3 and 4.4.

⁶See Chapter 2 and references therein, for example Cont (2001) and Taylor (2007).

The expected value of \mathbf{x}_t is denoted by $\boldsymbol{\theta} \in \mathbb{R}^N$. This is the true vector of strategy/signal performances. Define $\hat{\boldsymbol{\theta}}(\mathbf{X}) \in \mathbb{R}^N$ as the vector of column averages of \mathbf{X} :

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \quad (6.1)$$

which is the expected value of the row of \mathbf{X} according to the bootstrap distribution.

Let a *Rademacher random vector* $\boldsymbol{\epsilon}$ be a T -dimensional random vectors whose elements are iid and take values in 1 or -1 with probability 1/2. The *Rademacher Complexity* of \mathbf{X} is defined as:

$$\hat{R} = E_{\boldsymbol{\epsilon}} \left(\sup_n \frac{|\boldsymbol{\epsilon}' \mathbf{x}^n|}{T} \right)$$

Before stating a rigorous result linking this quantity to a bound on performance, we focus our attention on its interpretation. Specifically, we can interpret \hat{R} in at least three ways.

- **As the covariance to random noise:** Consider ϵ as a random covariate. We can interpret \hat{R} as the expected value of the highest covariance of the performance measure of a strategy to random noise. If, on average, for every set of +/-1 indicators, there is at least a strategy that covaries with it, then “we can do no wrong”: for every realization of a random series, there’s a strategy that would do well matching it. If we interpret the $x_{t,n}$ as predictions for epoch t , then this means that for every sequence of events ϵ_t we have a strategy that predicts them well.
- **As generalized 2-way cross-validation:** For sufficiently large T , the sets of positive elements in $\boldsymbol{\epsilon}_t$ concentrates around size $T/2$. We denote S^+ the set of $T/2$ periods where $\epsilon_t = 1$, and S^- the other periods. Rewrite the term inside the sup as

$$\frac{|\boldsymbol{\epsilon}' \mathbf{x}^n|}{T} = \frac{1}{2} \left| \frac{2}{T} \sum_{s \in S^+} x_{s,n} + \frac{2}{T} \sum_{s \in S^-} x_{s,n} \right| = \frac{1}{2} |\hat{\theta}_n^+ - \hat{\theta}_n^-|$$

For strategy n , this is the discrepancy in average performance measured on two equal-sized random subsets of the observations. By taking the sup across strategies, we are estimating the worst case: we estimate

performance on a subset, and get a very different result on the remaining subset! And if the discrepancy is high for each random subset, this will indicate that performance is not consistent: there's always at least a strategy that performs comparatively well *somewhere* and poorly in the remaining periods. The associated \hat{R} is high, and means that the set of strategies has unreliable performance.

- **As measure of span over possible performances:** We interpret ϵ as a “random direction” chosen at random in \mathbb{R}^T . The vector has Euclidean norm equal to \sqrt{T} . In the case where the performance measure is the standardized return, $E(\|\mathbf{x}^n\|)$ is also equal to \sqrt{T} , and is strongly concentrated around this value. The empirical Rademacher \hat{R} is then approximately equal to

$$E_\epsilon \left(\sup_n \left| \frac{\epsilon' \mathbf{x}^n}{\|\epsilon\| \|\mathbf{x}^n\|} \right| \right)$$

This can be interpreted in the following way. We have a set of N vectors $\mathbf{x}^1, \dots, \mathbf{x}^N$. We pick a random direction in the ambient space, and observe the maximum collinearity (expressed as the cosine distance) of this random direction to our vectors. The expected value of this collinearity measures how much our set of strategy vectors span \mathbb{R}^T . If we have n vectors that are copies of the same vector, the answer is: not very well. If conversely these vectors are all orthogonal, we have maximum collinearity. The Rademacher complexity is a geometric measure of how much the vectors \mathbf{x}^n “span” \mathbb{R}^T .

One interesting characteristic of the Rademacher complexity is that it takes into account dependence among strategies. If for example we had a billion strategies to our set of candidate strategies, but they are all identical (hence perfectly correlated) we are not increasing the Rademacher complexity. However, if the strategies are uncorrelated from each other, then the Rademacher complexity is high, indicating higher likelihood of overfitting.

6.3.2 Main result and Interpretation

The thrust of our protocol is to provide a uniform additive haircut to the performance statistic. In other terms, for each strategy n we have an empirical performance $\hat{\theta}_n$, by Equation (6.1). In the case of z-scored returns, this is the empirical Sharpe Ratio. Then, we can establish a probabilistic guarantee on the

true Sharpe Ratio: with high probability, say, greater than $1 - \delta$, the Sharpe Ratio of the strategy is greater than $\hat{\theta}_n - \text{“haircut”}$, where the haircut is a function of the Rademacher complexity, the number of samples T , and the parameter δ .

Here, we describe the steps that establish a lower bound for performance. We start with signals. In this case, we have $|x_{t,n}| \leq 1$, because the value is a correlation. For *all* signals, the true performance metric θ_n is bounded below by the empirical performance minus a haircut, with probability greater than $1 - \delta$:

$$\theta_n > \hat{\theta}_n - \underbrace{2\hat{R}}_{(data snooping)} - \underbrace{2\sqrt{\frac{\log(2/\delta)}{T}}}_{(estimation error)} \quad (6.2)$$

The result is described in Procedure 6.1.

Procedure 6.1: Rademacher Anti-Serum for Signals

1. Backtest all the strategies using a walk-forward procedure. Let $\mathbf{X} \in \mathbb{R}^{T \times N}$ be the matrix with Information Coefficients of strategy n at time t .
2. Compute $\hat{\theta}(\mathbf{X})$, as defined in Equation (6.1).
3. Compute $\hat{R}(\mathbf{X})$.
4. For all $n \in 1, \dots, N$

$$\theta_n > \hat{\theta}_n - 2\hat{R} - 2\sqrt{\frac{\log(2/\delta)}{T}}$$

with probability greater than $1 - \delta$.

Now, consider the case for Sharpe analysis. The formula is similar, but with a different estimation error.

$$\theta_n - \hat{\theta}_n \geq - \underbrace{2\hat{R}}_{(data snooping)} - \underbrace{3\sqrt{\frac{2\log(2/\delta)}{T}} - \sqrt{\frac{2\log(2N/\delta)}{T}}}_{(estimation error)} \quad (6.3)$$

The proofs are in the Appendix, Subsection 6.4, Theorems 6.3 and 6.4.

Procedure 6.2: *Rademacher Anti-Serum for Sharpe*

1. Backtest all the strategies using a walk-forward procedure. Let $\mathbf{X} \in \mathbb{R}^{T \times N}$ be the matrix with Information Ratio of strategy n at time t .
2. Compute $\hat{\theta}(\mathbf{X})$, as defined in Equation (6.1).
3. Compute $\hat{R}(\mathbf{X})$.
4. For all $n \in 1, \dots, N$

$$\theta_n - \hat{\theta}_n \geq -2\hat{R} - 3\sqrt{\frac{2 \log(2/\delta)}{T}} - \sqrt{\frac{2 \log(2N/\delta)}{T}}$$

with probability greater than $1 - \delta$.

We focus on the interpretation of the claim. The theorem states that the lower bounds on IC and Sharpe hold *simultaneously* at least with probability $1 - \delta$. Moreover the statement holds for any finite T ; no asymptotic approximation is involved. The true expected performance differs from the empirical performance because of two nonnegative terms:

- The first is the term $2\hat{R}$. This is the *data snooping term*. The larger the number of strategies, the higher the \hat{R} , because sup is strictly increasing in the set of strategies. Moreover, as we discussed, the higher the dependency among strategies, the lower \hat{R} . In the limit case where we test multiple replicas of the same strategy \hat{R} is zero. If the number of periods goes to infinity, $2\hat{R}$ does not go to zero.
- The second is the *estimation term*. It is the unavoidable For some intuition, consider the case of T iid normal random variables θ_t with mean 0 and unit variance. Their average $\hat{\theta}$ is approximately distributed as a normal distribution with standard deviation $1/\sqrt{T}$. What is the δ -quantile of the distribution? There is no closed-formula for it, but we can approximate it using Equation (2.4). For a normal distribution with zero mean and

standard deviation $1/\sqrt{T}$, and Cumulative Distribution Function F ,

$$F^{-1}(\delta) \geq -\sqrt{\frac{2 \log[1/(\sqrt{2\pi}\delta)]}{T}}$$

This is similar, save for constants, to the estimation errors in Equations (6.2) and (6.3). In the limit $T \rightarrow \infty$, The estimation error in both procedures approaches 0.

The procedure is operationally simple: simulate all possible strategies in a walk-forward manner. There should be no look-ahead bias. The strategies should be formulated without looking at the entire data set and their parameters should be tuned based on past history only. As we mentioned in the “best practices” section, all strategies should be documented and should run in parallel to the production strategy. Then, estimate the Rademacher complexity of matrix \mathbf{X} by the expectation in the definition of that statistic. The Rademacher complexity is easy to compute for tens of millions (or more) of strategies, and can be computed for even larger sets of strategies using tools from numerical analysis.

A few more practical remarks are in order, for the application of the formulas:

- The SAR procedure for signals uses the worst case $|x_{t,n}| \leq 1$. In practice, however, it is extremely unlikely to observe ICs close to one; and IC greater than 0.1 is extremely unlikely. If we assume $|x_{t,n}| \leq \kappa < 1$, and apply Theorem 6.3, the estimation term becomes smaller, by a factor κ :

$$\text{“estimation error”} = 2\kappa \sqrt{\frac{\log(2/\delta)}{T}}$$

Consider some realistic parameters: $\kappa = 0.02$, $\delta = 0.01$ and $T = 2500$. Then the estimation error is about 0.002.

- In the SAR procedure for strategies, the formula for the estimation error is a rather simple bound and the constant factor could be probably improved. For realistic parameters, the error is quite large. For example $\delta = 0.01$, $T = 2500$, and $N = 1E6$, the estimation error is 0.31, corresponding to an annualized estimation error of 5.1. This seems a loose bound, compared to the standard formula for the standard error of the Sharpe Ratio (Lo, 2002): for a strategy with Sharpe Ratio equal to 3, the estimation error is “estimation error” = $-F(\delta)\sqrt{(1 + SR^2/2)/T}\sqrt{252} = 1.7$.

6.4 ★Appendix: Proofs

We use some essential inequalities in the proofs. Standard references are [Boucheron et al. \(2013\)](#) and [Vershinini \(2018\)](#).

Theorem 6.1 (McDiarmid's inequality). *Let X_1, \dots, X_n be independent random variables, and $f : \mathbb{R}^n \rightarrow \mathbb{R}$, such that, for each i ,*

$$\sup_{x_i, x'_i} |f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c_i$$

Then, for all $\epsilon > 0$,

$$P(|f - Ef| > \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_i c_i^2}\right)$$

Specifically, if $c_i = c$, and with probability greater than $1 - \delta/2$,

$$f < Ef - \sqrt{\frac{nc^2}{2} \log(\delta/2)}$$

A mean-zero sub-gaussian random variable X is one for which a positive constant σ exists, such that the inequality $P(|X| > \epsilon) \leq 2 \exp(-\epsilon^2/(2\sigma^2))$ holds for all positive ϵ . The parameter σ^2 is the proxy variance.

Theorem 6.2 (Generalized Hoeffding's inequality). *Let X_1, \dots, X_n be i.i.d. random variables with finite sub-gaussian norms and proxy. Then, for all $\epsilon > 0$,*

$$P\left(\frac{1}{n} \sum_{i=1}^n X_i - EX > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) \quad (6.4)$$

Theorem 6.3 (Bounds for Bounded Performance Metrics). *Assume that $|x_{tn}| \leq a$ for all $n = 1, \dots, N, t = 1, \dots, T$. For all $n \in 1, \dots, N$*

$$\theta_n \geq \hat{\theta}_n - 2\hat{R} - 3a\sqrt{\frac{2\log(2/\delta)}{T}} \quad (6.5)$$

Proof. The straightforward inequality holds for all $n = 1, \dots, N$: $\theta_n - \hat{\theta}_n \geq -\sup_n |\hat{\theta}_n - \theta_n|$. Define

$$\Phi := \sup_n |\hat{\theta}_n - \theta_n| \quad (6.6)$$

We claim that with probability greater than $1 - \delta/2$

$$\Phi \leq E_D \Phi + a \sqrt{\frac{2 \log(2/\delta)}{T}} \quad (6.7)$$

This allows one to deal with $E_D \sup_n |\hat{\theta}_n - \theta_n|$, which is easier. To prove the inequality, note that, for all $x_{t,i}, x'_{t,i} \in [-a, a], t = 1, \dots, T, i = 1, \dots, N$,

$$|\hat{\theta}_n(\dots, x_{t,i}, \dots) - \hat{\theta}_n(\dots, x'_{t,i}, \dots)| \leq \frac{2a}{T} \quad (6.8)$$

From which it follows that

$$\sup_{x_{t,n}, x'_{t,n} \in \mathbb{R}^T} |\Phi(\dots, x_{t,n}, \dots) - \Phi(\dots, x'_{t,n}, \dots)| \leq \frac{2a}{T}$$

We apply McDiarmid's inequality to Φ to obtain the result.

In order to obtain a lower bound on θ_n we need an upper bound on $E_D \Phi$. In the equalities below, we introduce a probability measure D' identical to, and independent from, D .

$$\begin{aligned} & E_D \sup_n |\hat{\theta}_n - \theta_n| \\ &= E_D \sup_n |\hat{\theta}_n(\omega) - E_{D'} \hat{\theta}_n(\omega')| \\ &= E_D \sup_n |E_{D'} \hat{\theta}_n(\omega) - E_{D'} \hat{\theta}_n(\omega')| \quad (\text{conditioning}) \\ &\leq E_D E_{D'} \sup_n |\hat{\theta}_n(\omega) - \hat{\theta}_n(\omega')| \quad (\text{Jensen}) \\ &\leq \frac{1}{T} E_D E_{D'} \sup_n \left| \sum_t (x_{t,n}(\omega) - x_{t,n}(\omega')) \right| \\ &= (*) \end{aligned}$$

We introduce an additional source of noise (the ϵ Rademacher vector) and we lose a constant of 2, but gain in tractability. We can change the signs of each summand by multiplying by some arbitrary factor $y_t \in \{+1, -1\}$, since the

terms are exchangeable.

$$\begin{aligned}
(*) &= \frac{1}{T} E_D E_{D'} \sup_n \left| \sum_t y_t(x_{t,n}(\omega) - x_{t,n}(\omega')) \right| \\
&= \frac{1}{T} E_D E_{D'} E_\epsilon \sup_n \left| \sum_t \epsilon_t(x_{t,n}(\omega) - x_{t,n}(\omega')) \right| \\
&\leq \frac{1}{T} E_D E_{D'} E_\epsilon \sup_n \left| \sum_t \epsilon_t(x_{t,n}(\omega)) \right| + E_D E_{D'} E_\epsilon \sup_n \left| \sum_t \epsilon_t x_{t,n}(\omega') \right| \\
&= \frac{1}{T} E_D E_\epsilon \sup_n \left| \sum_t \epsilon_t x_{t,n}(\omega) \right| + \frac{1}{T} E_D E_\epsilon \sup_n \left| \sum_t \epsilon_t x_{t,n}(\omega') \right| \\
&= \frac{2}{T} E_D \hat{R} \\
&= 2R
\end{aligned}$$

Where we defined R as the expected value of the Rademacher complexity over the distribution of performance realizations.

We now use McDiarmid again: for all $x_{t,i}, x'_{t,i}$,

$$|\hat{R}(\dots, x_{t,i}, \dots) - \hat{R}(\dots, x'_{t,i}, \dots)| \leq \frac{2a}{T}$$

Hence, with probability greater than $1 - \delta/2$

$$R \leq \hat{R} + a \sqrt{\frac{2 \log(2/\delta)}{T}} \quad (6.9)$$

Now we employ the union bound on inequalities (6.7) and (6.9) to obtain the claim. \square

A random variable ξ is σ -sub-gaussian if there is a $\sigma > 0$ such that $E \exp(\lambda \xi) \leq (\lambda^2 \sigma^2 / 2)$, or if equivalently, $P(|\xi| > a) < 2 \exp(-a^2 / 2\sigma^2)$.

Theorem 6.4 (Bounds for Sub-Gaussian Performance Metrics). *Assume that $P(|x_{t,n}| > \epsilon) \leq 2e^{-\epsilon^2/2}$ for all $\epsilon > 0$, for all $n = 1, \dots, N, t = 1, \dots, T$. Then, for all $n \in 1, \dots, N$*

$$\theta_n - \hat{\theta}_n \geq -2\hat{R} - 3\sqrt{\frac{2 \log(2/\delta)}{T}} - \sqrt{\frac{2 \log(2N/\delta)}{T}}$$

Proof. Let $a > 0$. We split $\theta_n - \hat{\theta}_n$ into the sum of two terms: $\theta_n - \hat{\theta}_n = g(\mathbf{x}^n, a) + h(\mathbf{x}^n, a)$, where

$$\begin{aligned}\theta_n - \hat{\theta}_n &= g(\mathbf{x}^n, a) + h(\mathbf{x}^n, a) \\ &\geq -\sup_n |g(\mathbf{x}^n, a)| - \sup_n |h(\mathbf{x}^n, a)| \\ g(\mathbf{x}^i, a) &:= E\left[\frac{1}{T} \sum_{t=1}^T x_{t,i} \mathbf{1}(|x_{t,i}| \leq a)\right] - \frac{1}{T} \sum_{t=1}^T x_{t,i} \mathbf{1}(|x_{t,i}| \leq a) \\ h(\mathbf{x}^i, a) &:= E\left[\frac{1}{T} \sum_{t=1}^T x_{t,i} \mathbf{1}(|x_{t,i}| > a)\right] - \frac{1}{T} \sum_{t=1}^T x_{t,i} \mathbf{1}(|x_{t,i}| > a)\end{aligned}$$

We bound $P(\sup_i |h(\mathbf{x}^i, a)| \geq v)$. By symmetrization

$$E|h(\mathbf{x}^i, a)| \leq 2E \sum_{t=1}^T |\epsilon_t x_{t,i} \mathbf{1}(|x_{t,i}| > a)|$$

The random variable $|\epsilon_t x_{t,i} \mathbf{1}(|x_{t,i}| > a)|$ is subgaussian, since it is dominated by $|x_{t,i}|$ with probability 1, and it has the same proxy variance as $|x_{t,i}|$. By the General Hoeffding inequality,

$$\begin{aligned}P\left(\left|\sum_{t=1}^T h(x_{t,i})\right| > v\right) &\leq \exp(-Tv^2/2) \\ P\left(\sup_i \left|\sum_{t=1}^T h(x_{t,i})\right| > v\right) &\leq N \exp(-Tv^2/2) \\ P\left(\sup_i \left|\sum_{t=1}^T h(x_{t,i})\right| > \sqrt{2 \log(2N/\delta)/T}\right) &\leq \delta/2\end{aligned}$$

By the union bound,

$$\theta_n - \hat{\theta}_n \geq -2\hat{R} - 3\sqrt{\frac{2 \log(2/\delta)}{T}} - \sqrt{\frac{2 \log(2N/\delta)}{T}} \quad (6.10)$$

□

Chapter 7

Evaluating Risk

Draft (June 21, 2024). Please read the chapter carefully and send comments and corrections to the author. Any contribution will be acknowledged in the final copy.

Email: paleologo@gmail.com (send email with “EQI” in the title)

In the research process we develop models for expected returns, risks and transaction costs. The identification of an effective model in each category rests of one or more loss functions. The metrics we consider are:

- R squared on total returns;
- R squared on idiosyncratic returns;
- Factor Performance;
- Volatility losses: Qlike, MSE, Sharpe (again);
- Turnover.
- Betas;

7.1 Evaluating Alpha

Given a factor estimation model (characteristic, statistical, time-series), the analyst has for every time period, the tuple $(\alpha_t, \mathbf{B}_t, \mathbf{f}_t, \epsilon_t, \Omega_{\mathbf{f},t}, \Omega_{\epsilon,t})$.

- *R squared on total returns.* A simple measure of the ability of the model to describe returns is the *coefficient of determination*, also called R squared,

defined as 1 minus the ratio of the weighted residual sum of squares and the total sum of squares.

$$R^2 = 1 - \frac{\sum_{t=1}^T \left\| \boldsymbol{\Omega}_{\epsilon,t}^{-1/2} (\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t) \right\|^2}{\sum_{t=1}^T \left\| \boldsymbol{\Omega}_{\epsilon,t}^{-1/2} \mathbf{r}_t \right\|^2}$$

In the second equality we have replaced the residual with the remaining terms from the basic factor model equation 3.1. A high coefficient of determination is interpreted as a positive attribute of the model specification. Keep in mind, however, that R^2 is usually increasing in the number of factors when we use a weighted least squares loss function. In order to estimate the model, we are minimizing the R square loss function, and there is no way to assess the same metric out of sample. As such, the R squared on total returns is only mildly informative and should be used as a descriptive statistic.

- *R squared on idiosyncratic returns.* Assume that we have predictions of alphas. A prediction $\hat{\alpha}_{\perp,t}$ will be functions of data available up to $t-1$. Given residuals $\epsilon_t := \mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t$, we ask what is the performance of these predictions. For simplicity, let $\boldsymbol{\Omega}_{\epsilon,t} = \mathbf{I}$. Let κ a scaling factor for the predicted alpha. We saw in Insight 4.3 that the R squared of the weighted least squares regression of ϵ_t on $\hat{\alpha}_{\perp,t}$ is linked to the Information Ratio by the relationship $IR = \sqrt{R^2 n T}$.

We can classify the loss functions used for multivariate volatility estimation in two broad categories. First, there are diagnostic losses that are used to check the performance of intermediate steps in the covariance matrix estimation. For example, in the characteristic model, we may want to compare the average R squared, BIC, AIC of the cross-sectional regressions (Hastie et al., 2008). Similar metrics are available for the other estimation approaches. I shall skip the treatment of these, since they are not chosen based on a strong theoretical foundation, and do not impact directly the performance of the model. We can put them under the rubric “practical knowledge”. Second, there are *principled* metrics that quantify the quality of the *overall* asset covariance matrix. Sometimes, these are derived from loss functions that quantify the quality of univariate volatility forecasts. These can be adapted to estimating the performance of a covariance matrix, by applying them to portfolio returns, and then imposing either a distribution on portfolios, or identifying the worst-case portfolio performance.

7.2 Evaluating The Covariance Matrix

7.2.1 Robust Loss Functions for Volatility Estimation

A major application of a factor model is volatility estimation. The quality of volatility prediction is one that has been at the forefront of the early developments of risk models. If a model predicts asset volatilities $\hat{\sigma}_i$ at a certain time, then a natural measure of the quality of the volatility predictions is given by the bias statistic

$$\text{bias}^2(\hat{\sigma}, \tilde{\sigma}) := \frac{1}{n} \sum_{i=1}^n \left(\frac{\tilde{\sigma}_i^2}{\hat{\sigma}_i^2} - 1 \right)$$

where $\tilde{\sigma}_i$ is an empirical estimate of the observed volatility (e.g., the realized volatility metric defined previously). The simplest empirical estimate of realized volatility is the one based on a single observation of the return vector r . In this case the formula becomes

$$\text{bias}(\hat{\sigma}, r) := \frac{1}{n} \sum_{i=1}^n \left(\frac{r_i^2}{\hat{\sigma}_i^2} - 1 \right)$$

If the model predicts well, we should observe a bias close to 0, and it is straightforward to formulate test of hypothesis and asymptotic confidence intervals on this statistic.

In the bias statistic above, we use a volatility proxy $\tilde{\sigma}_i$ instead of the unknown true volatility σ_i . Hansen and Lunde (2006a) introduce a concept of *rank robustness* for losses: if we have two alternative volatility forecasts $\hat{\sigma}_i^{(j)}$, one is better than the other using an unbiased volatility proxy if and only if one is better than the other using the true volatility. I.e.,

$$\text{bias}(\hat{\sigma}^{(1)}, \sigma) \leq \text{bias}(\hat{\sigma}^{(2)}, \sigma) \Leftrightarrow \text{bias}(\hat{\sigma}^{(1)}, \tilde{\sigma}) \leq \text{bias}(\hat{\sigma}^{(2)}, \tilde{\sigma})$$

Patton and Sheppard (2009) and Patton (2011) completely characterize these loss functions and show that these two are robusts:

$$\begin{aligned} \text{QLIKE}(\hat{\sigma}, r) &:= \frac{1}{n} \sum_{i=1}^n \left(\frac{r_i^2}{\hat{\sigma}_i^2} - \log \left(\frac{r_i^2}{\hat{\sigma}_i^2} \right) - 1 \right) \\ \text{MSE}(\hat{\sigma}, r) &:= \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2 \left(\frac{r_i^2}{\hat{\sigma}_i^2} - 1 \right)^2 \end{aligned}$$

These two loss functions are increasingly being used in place of the bias statistic.

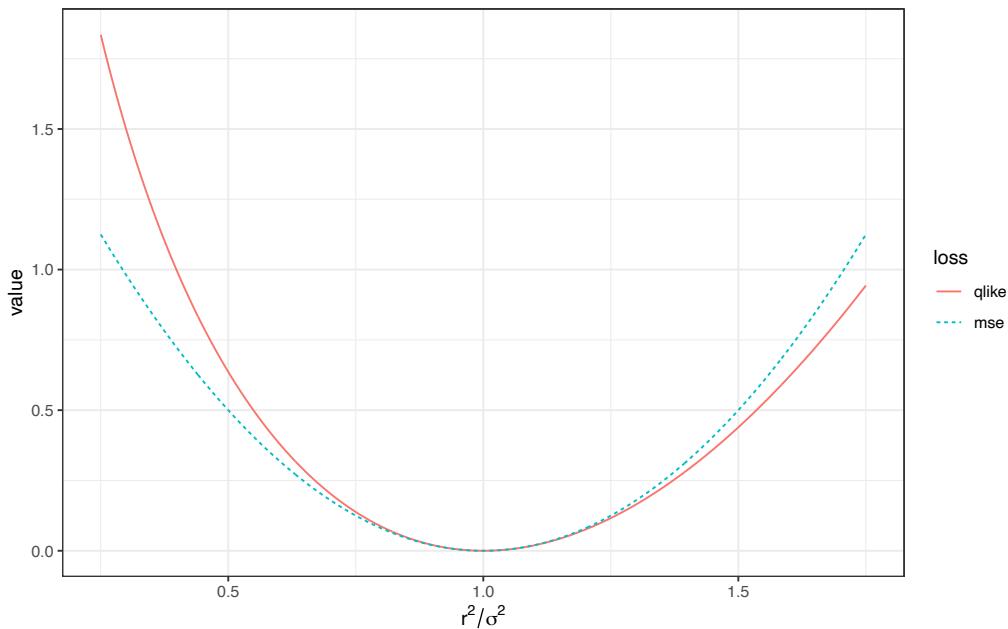


Figure 7.1: QLIKE and MSE comparison. Notice that QLIKE is skewed, with higher losses when the realized variance is greater than estimated variance.

7.2.2 Application to Multivariate Returns

The loss functions QLIKE and MSE apply to univariate returns, not to covariance matrices. Below are a few ways to adapt the univariate setting to a multivariate one.

Production Strategies. If strategies are already running, a straightforward and necessary test is to evaluate their simulated performance under different factor models. QLIKE and MSE are important and should be checked jointly with metrics that are important for the portfolio manager, like Sharpe Ratio or PnL. It is important that the covariance matrix be the input to the entire production process, i.e., portfolios should be generated on the basis of the factor model themselves. If a portfolio is generated using factor model A and then tested on model B, the test will be marred by this asymmetry.

Average-Case Analysis. An alternative approach is to estimate the expected loss, where the expectation is taken on a distribution of portfolios as well as of asset returns. For the distribution of returns, we use the empirical measure \hat{P} of historical returns; for the distribution of portfolios, we may choose a simple one, like uniform on a sphere. Then we estimate $E_{\mathbf{w} \sim D, \mathbf{r} \sim \hat{P}}[L((\mathbf{r}'\mathbf{w})^2, \mathbf{w}'\hat{\Omega}_r\mathbf{w})]$. There are a few drawbacks to this approach. First, there is a degree of arbitrariness in

choosing a portfolio distribution. The actual distribution of portfolio is almost certainly *not* uniform; and it is not even warranted that the distribution of alphas is uniform. Secondly, it is computationally expensive. We are simulating in highly dimensional spaces, with \mathbf{w} ranging in size from 1,000 assets to tens of thousands. Computational issues, such as simulation scheme and convergence criteria, become important. An approximation is to select a *portfolio basis*, i.e., n portfolios $\mathbf{w}_1, \dots, \mathbf{w}_{n,t}$, and then apply an “average”-case analysis to these n portfolios. A special case of this special case is that of *eigenportfolios*: decompose $\Omega_r = \mathbf{U}\mathbf{S}\mathbf{U}'$, and set the basis portfolios equal to the columns of \mathbf{U} . The approach is computationally tractable. One important drawback is that this average loss is not independent of the choice of the portfolios; in fact, it is quite sensitive to it. Even if we restrict our choice to an orthonormal basis, like eigenportfolios, the measured performance still depends on the basis. Since the choice of an appropriate basis cannot be easily justified based on principles, the outcome is arbitrary.

Procedure 7.1: Random Portfolios Average Variance Testing

1. **Inputs:** candidate covariance matrices $\hat{\Omega}_{r,t}$ and returns \mathbf{r}_t for $t = 1, \dots, T$, loss function L .
2. Set $L_{\text{tot}} = 0$, $n_{\text{iter}} = 0$.
3. Set $\mathbf{w} \sim N(0, \mathbf{I}_n)$, $\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$. Choose s uniformly at random in $1, \dots, T$.
4. Set $L_{\text{tot}} = L_{\text{tot}} + L[(\mathbf{r}'_s \mathbf{w})^2, \mathbf{w}' \hat{\Omega}_s \mathbf{w}]$. Set $n_{\text{iter}} \leftarrow n_{\text{iter}} + 1$
5. If $L[(\mathbf{r}'_s \mathbf{w})^2, \mathbf{w}' \hat{\Omega}_s \mathbf{w}]/L_{\text{tot}} \geq \epsilon_{\text{tol}}$, go to Step 3.
6. **Output:** $L := L_{\text{tot}}/n_{\text{iter}}$.

Worst-case under/over-prediction. Yet another approach is to estimate the worst-case loss function:

$$\begin{aligned} & \max E_{\mathbf{r} \sim \hat{P}}[L((\mathbf{r}' \mathbf{w})^2, \mathbf{w}' \hat{\Omega}_r \mathbf{w})] \\ & \text{s.t. } \|\mathbf{w}\| \leq 1 \end{aligned}$$

This problem with this approach is that the objective function (be it QLIKE or MSE) is not convex. When the number of assets is large, the problem is not

computationally tractable.

Procedure 7.2: Worst-Case Variance Testing

1. **Inputs:** candidate covariance matrices $\hat{\Omega}_{\mathbf{r},t}$ and returns \mathbf{r}_t for $t = 1, \dots, T$, loss function L .
2. Set $L_{\text{tot}} = 0$, $n_{\text{iter}} = 0$, $\mathbf{w} \sim N(0, \mathbf{I}_n)$, $\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$.
3. Set s uniformly at random in $1, \dots, T$.
4. Set $\mathbf{w} \leftarrow \mathbf{w} - n_{\text{iter}}^{-1} \nabla_{\mathbf{w}} L[(\mathbf{r}'_s \mathbf{w})^2, \mathbf{w}' \hat{\Omega}_s \mathbf{w})]$. Set $n_{\text{iter}} \leftarrow n_{\text{iter}} + 1$
5. If $L[(\mathbf{r}'_s \mathbf{w})^2, \mathbf{w}' \hat{\Omega}_s \mathbf{w})] / L_{\text{tot}} \geq \epsilon_{\text{tol}}$, go to Step 3.
6. **Output:** $L := L_{\text{tot}} / n_{\text{iter}}$.

Neither of the options above dominates the others. Whenever available, production strategies are always being tested against alternative approaches. Average and worst-case analyses are both computationally very demanding. Moreover, in the case of average-case analysis the result depends on the assumption on portfolio distribution.

Leading alpha MVO portfolios. Another option is to construct portfolios that are constructed on the realized leading returns of the securities. This scheme has the advantage to test the predictiveness of the strategy for “relevant” portfolios and is described in Procedure 7.3 Volatility prediction matters if we have alpha. If we don’t, then we have other problems to be worried about. An advantage of this approach is that it can be easily augmented. For example, we could test the performance on portfolios with added noise:

$$\mathbf{w} := \hat{\Omega}_{\mathbf{r},t}^{-1}(\hat{\alpha}_t + \boldsymbol{\eta}_t), \quad \boldsymbol{\eta}_t \sim N(0, \sigma^2 \mathbf{I}_n)$$

Distribution Likelihood. An alternative that does not depend on the portfolio choice is to use the log-likelihood for the zero-mean multivariate normal distribution, applied to the returns of the estimation universe. Modulo constant

Procedure 7.3: *Realized Alpha Variance Testing*

1. **Inputs:** candidate covariance matrices $\hat{\Omega}_{\mathbf{r},t}$ and returns \mathbf{r}_t for $t = 1, \dots, T$, loss function L , $\tau \in \mathbb{N}^+$.
2. Set $L_{\text{tot}} = 0$.
3. For each $t = 0, T - \tau$, let

$$\begin{aligned}\hat{\alpha}_t &:= \frac{1}{\tau} \sum_{s=t+1}^{t+\tau} \mathbf{r}_s \\ \mathbf{w} &:= \hat{\Omega}_{\mathbf{r},t}^{-1} \hat{\alpha}_t \\ L_{\text{tot}} &:= L_{\text{tot}} + L(\mathbf{r}'_t \mathbf{w}, \mathbf{w}' \hat{\Omega}_{\mathbf{r},t} \mathbf{w})\end{aligned}$$

4. **Output:** $L := L_{\text{tot}}/(T - \tau + 1)$.

terms, the negative log likelihood is proportional to

$$\text{QDIST} = \sum_t \left(\mathbf{r}'_t \hat{\Omega}_{\mathbf{r},t}^{-1} \mathbf{r}_t + \log |\hat{\Omega}_{\mathbf{r},t}| + n_t \log(2\pi) \right)$$

It is therefore

7.3 Evaluating the Precision Matrix

7.3.1 Minimum-Variance Portfolios

As we saw repeatedly throughout the book, the quality of a factor model is reflected in the accuracy of precision matrix. We propose two methods. The first one is using a well-known test: minimum-variance portfolios. Consider a very simple example: construct a portfolio \mathbf{w} of minimum variance and with unit net market value $\sum_i w_i = 1$. This is the *ex ante* minimum variance portfolio; the realized variance will differ. The intuition is that a “better” covariance matrix will result in a lower realized variance. We make this intuition rigorous, and generalize to the case where the portfolio has a given exposure to an *arbitrary* factor, i.e., $\sum_i b_i w_i = 1$.

Let $\hat{\Omega}_r \in \mathbb{R}^{n \times n}$ be a candidate covariance matrix and Ω_r be the true covariance matrix. Let $b \in \mathbb{R}^n$, and solve the risk minimization problem

$$\begin{aligned} & \min w' \hat{\Omega}_r w \\ & \text{s.t. } b' w = 1 \end{aligned} \tag{7.1}$$

and let $w(\hat{\Omega}_r)$ be its solution. Denote the realized variance of the portfolio $\text{var}(w(\hat{\Omega}_r), \Omega_r)$. Then the realized volatility of portfolio $w(\hat{\Omega}_r)$ is greater than the one of $w(\Omega_r)$, and the two are identical if and only if $\Omega_r \propto \hat{\Omega}_r$. This is Theorem 14.1 in Appendix (Section 14.1). A way to apply this result is as follows: Set $b = E[r]$, the alpha vector. Then the correct covariance matrix results in the best possible Sharpe Ratio, and a natural ranking of covariance models is by the realized Sharpe Ratio for a certain strategy.

We can use all the portfolio-dependent schemes introduced for volatility tests to evaluate the precision matrix. The realized variance acts as a loss function.

7.3.2 Mahalanobis Distance

There is another test that is portfolio-independent and that involves the precision matrix only. The *Mahalanobis distance* is defined for a multivariate zero-mean random vector r and an associated covariance matrix Ω_r as $d(r, \Omega_r) := \sqrt{r' \Omega_r^{-1} r}$. For gaussian returns and under the true covariance matrix, d^2 is distributed according to a Chi-squared distribution with n degrees of freedom¹. One test then is

$$\begin{aligned} \nu_t &:= \frac{1}{n_t} r_t' \hat{\Omega}_{r,t}^{-1} r_t \\ \text{MSTD} &:= \text{stdev}(\nu) \end{aligned}$$

The lower the value of $\text{MSTD}(r, \hat{\Omega}_r)$, the better the performance of the precision matrix. If the standard deviation is very low (say, of the order of $\sqrt{2/n}$), then the inverse of the covariance matrix is, save for a constant, predicting perfectly. We don't primarily care about the constant in this test, because volatility test should address that issue better. A different way to interpret the result is the following: if returns are Gaussian, then they are distributed as $\Omega_r \xi$, with ξ multivariate standard normal. Moreover $x_t' x_t$ has mean n_t and standard

¹To prove this, note that the vector r can be generated by $r := \Omega_r^{1/2} \xi$, where $\xi \sim N(\mathbf{0}, \mathbf{I}_n)$. Therefore $r' \Omega_r^{-1} r \sim \sum_i \xi_i^2 \sim \chi_n^2$.

deviation $\sqrt{2n_t}$; We rewrite MSTD as

$$\begin{aligned}\nu_t &:= \frac{1}{n_t} \mathbf{r}'_t \hat{\boldsymbol{\Omega}}_{\mathbf{r},t}^{-1} \mathbf{r}_t \\ \text{MSTD} &:= \frac{1}{T} \sum_t \left(\frac{1}{n_t} \boldsymbol{\xi}'_t \boldsymbol{\Omega}_{\mathbf{r},t} \hat{\boldsymbol{\Omega}}_{\mathbf{r},t}^{-1} \boldsymbol{\Omega}_{\mathbf{r},t} \boldsymbol{\xi}_t - \kappa \right)^2 \\ &\simeq \frac{1}{T} \sum_t \frac{1}{n_t^2} \left(\boldsymbol{\xi}'_t \boldsymbol{\Omega}_{\mathbf{r},t} \hat{\boldsymbol{\Omega}}_{\mathbf{r},t}^{-1} \boldsymbol{\Omega}_{\mathbf{r},t} \boldsymbol{\xi}_t - \kappa \boldsymbol{\xi}'_t \boldsymbol{\xi}_t \right)^2 \\ &= \frac{1}{T} \sum_t \frac{1}{n_t^2} \left[\boldsymbol{\xi}'_t (\boldsymbol{\Omega}_{\mathbf{r},t} \hat{\boldsymbol{\Omega}}_{\mathbf{r},t}^{-1} \boldsymbol{\Omega}_{\mathbf{r},t} - \kappa \mathbf{I}_n) \boldsymbol{\xi}_t \right]^2\end{aligned}$$

We are testing the closeness of $\boldsymbol{\Omega}_{\mathbf{r},t} \hat{\boldsymbol{\Omega}}_{\mathbf{r},t}^{-1} \boldsymbol{\Omega}_{\mathbf{r},t} - \kappa \mathbf{I}_n$ to zero by multiplying to the left and right by standard multivariate Gaussian random vectors. In Section 5.3.2 we saw that this matrix difference is responsible for bounding the Sharpe Ratio Efficiency.

7.4 Ancillary Tests

In addition to the performance

7.4.1 Beta vs realized beta

Compute the realized volatility of the test portfolio when they are hedged by the benchmark, using the predicted beta of the test portfolio to the benchmark, using a given model.

7.4.2 Model Turnover

For a model, compute the turnover of the minimum-variance portfolio, and of the portfolio with weights $[w]_i = 1/[\Omega_\epsilon]_{ii}$. Report the average turnovers.

7.5 Further Reading

For the relationship between IR and Sharpe, see Chincarini and Kim (2007, 2022).

Chapter 8

Fundamental Factor Models

Fundamental (or characteristic-based) factor models estimate Equation (3.1) using as inputs \mathbf{r}_t and \mathbf{B}_t . The outputs of the models are estimates of the factor and idiosyncratic returns $\mathbf{f}_t, \boldsymbol{\epsilon}_t$, as well as their covariance matrices $\Omega_f, \Omega_\epsilon$. Fundamental factors are perhaps the most popular model among practitioners. Reasons for their popularity are:

- *Good Performance.* Commercial models are the outcome of a long process of refinement. The first models date back to the mid 1970s. Consequently, some important factors have been identified.
- *Interpretability.* Firm characteristics provide summary description of individual firms, and exposures based on these characteristics give a summary of a portfolio;
- *Connections to Academic Research.* In the asset pricing literature, multi-factor models originate with the Arbitrage Pricing Theory of Ross, and the reference model used by academic researchers to identify pricing anomalies is the three-factor model by Fama and French (1993).
- *Alpha Research.* Fundamental models are the workhorse of alpha research, because they allow the portfolio manager to incorporate almost any data source, to analyze very large data sets, to interpret the outcome of the analysis, and to feed the outcome to a portfolio construction system.

8.1 The Inputs and the Process

There are five major steps needed to identify a factor model. Some of them require sound quantitative methods; others are more art than science. Before we even begin to describe the steps, we should focus on the inputs.

8.1.1 The Inputs

Fundamental model inputs are:

1. a set of returns per asset/date, i.e., the \mathbf{r}_t part;
2. a set of *raw characteristics* per asset/date/characteristic identifier; from these inputs we generate the \mathbf{B}_t .

Asset Returns. Returns are usually reported over intervals of equal duration. These intervals determine the periodicity of the model. Daily returns may be based on close-to-close prices. Intraday returns may be based on the last transaction price observed in the intraday interval. The interval can range from thirty minutes to sub-minute interval. It would seem that returns are unambiguously defined, but this is *not* the case. The answer to the question “what is the final price in a time period” is not easy and is not unique. Ultimately models of returns should help the portfolio manager develop a profitable real-life strategy. If prices are such that we could not have executed transactions reliably at their quoted values, then the factor model will not be reliable, and neither will be the strategy built on the model. Consider the closing price. Where does it come from? In many stock exchanges, at the end of the day the Limit Order Book (LOB) is replaced by a Closing Auction (CA). Without delving in the details of a CA, suffices to know that a CA is a very liquid event, in which about 10% of the daily volume of a stock is traded. Consequently *for a liquid stock* the closing price is meaningful, in the sense that it is exploitable by a portfolio manager, at a non-negligible size. Now, compare this scenario to one in which we are interested in modeling a small-cap stock that is a component of the Russell 2000 index. Such a stock would likely qualify to be a member of the risk model estimation universe. However, it could trade at very low volumes. In addition, if we model intraday returns, then we must pay additional care. What does it mean that the price at the end of a 10-minute interval was \$6.93? Maybe the stock did not even transact in that interval, and the period return is zero. Or maybe there were only a handful of trades. What is the correct price, then? The transaction price? That is not obvious. Maybe the transaction happened at the ask, but the transaction just before happened at the bid, just because of random circumstances. Or, should we use the mid price between the bid and the ask? Could our strategy reliably transact at either of these prices in real life? These are just some of the many questions one should ask when developing models based on intra-day return models, but not only. For many asset classes, determining good closing daily closing prices is a very challenging, important

and thankless task. The details are too asset- and data-vendor-specific to be covered in detail; moreover, this is an area where traders accumulate Intellectual Property that is very material to their success. A few heuristic rules should help. First, the shorter the interval, the harder the problem. We saw this already in Section 2.2.4. You have to model what price is appropriate (bid, ask, mid); whether to explicitly model the noise in price observations; whether time-of-day non-stationarity matters; whether to model close-to-open returns and intraday returns separately. The second recommendation is to think explicitly about the relationship between asset liquidity and model periodicity. Liquidity, for our purposes, could be proxied by trading volume in a fixed interval. Usually average daily trading volume is available. Liquidity is related to price discovery. The price of a very liquid stock is less prone to observation errors, and it is more transactable, than that of an illiquid stocks. Therefore, the choice of model universe and model periodicity are related. Unless you want to model market microstructure explicitly, and want to rely on closing per-period prices, then a shorter period will imply that your model universe will be smaller.

Raw characteristic data. This is the “art”, or better, “dark art” part of the modeling task. “Raw data” can mean almost anything. A possible classification of raw data is in structured and unstructured. The former include numerical data and categorical data that can be associated to a security and to an estimation period. *Categorical* data take only a finite set of values, which do not necessarily have an ordering relationship. Examples are countries and sectors of a stock. A slightly less common example is the credit rating of a company, which does admit an order. Unstructured data include any data that do not come directly in such a tabular form. Examples are the earnings transcripts of a company, or its regulatory filings (Forms 10-K, 10-Q, 8-Q); or the web scraping of a firm with information about its products; or the consumer credit card transactions with a firm; or, even more, location data of customers visiting the store of a firm. These few examples give just a taste of the immensity of possible inputs to a model. For asset return modeling purposes, we extract from these vast troves of alternative data some representative statistics that can be interpreted as structured data. For example, from transactional data, we can extract levels (dollars transacted in a quarter by a consumer firm’s web portal) and trends (quarterly changes in such level); or we can extract measures of geographical dispersion. Moreover, some of these operations can be automated, or require the use of machine learning tasks like classification and clustering. One important feature of all these operations, though, is that they entail some form of human

expertise. In fact, the task of extracting structured data from unstructured information is perhaps the one that requires the highest amount of human intelligence. A great amount of papers have been published on characteristic data. I summarize some of the essential results in Section xx. Maybe in the future we will be able to feed such disparate sources of information into a black box and directly predict prices, or even recommend trades. In that event, I'll gladly write a second edition to this book.

8.1.2 The Process

The estimation steps of a characteristic model are:

1. *Data Ingestion.* This step encompasses receiving data sets from vendors, checking their integrity and performing essential data checks. Among them:
 - Ensure that data are of the correct type and not corrupt. This happens with positive probability.
 - Ensure that the set of securities is not substantially different for previous period.
 - Ensure that the fraction of missing data per asset and characteristic is not substantially different.
 - Identify and report data outliers.
2. *Estimation Universe Selection.* I introduced issues related to this set earlier in the chapter. The criteria for inclusions are:
 - *Tradability.* The assets must be sufficiently liquid, because factor-mimicking portfolios include all and all these assets.
 - *Data quality.* This is closely related to liquidity of the assets, but for a different goal. We need securities for which prices are *discovered*, i.e., close to their economic fundamental value, since we are using those prices for return calculations and model estimation.
 - *Relevance to investments.* The estimation universe should be overlapping to some extent with the investment universe of the strategy. This is more of an art. There is not (to my knowledge) a rigorous treatment of this problem.

3. *Winsorization.* Identify outliers in returns of the estimation universe and winsorize them.
4. *Loadings Generation.* Generate characteristics by transformations and combinations.
5. *Cross-Sectional Regression.* For each $t = 1, \dots, T$, perform a cross-sectional regression of asset returns \mathbf{r}_t against the loadings \mathbf{B}_t . The outputs of this step are the vectors $\hat{\mathbf{f}}_t$ and $\hat{\boldsymbol{\epsilon}}_t$.
6. *Time-Series Estimation.* Using the time series from the first step, estimate:
 - a) the factor covariance matrix $\hat{\Omega}_{\mathbf{f},t}$;
 - b) the idiosyncratic covariance matrix $\hat{\Omega}_{\boldsymbol{\epsilon},t}$;
 - c) the risk-adjusted performance of factors returns.

Procedure 8.1: *Steps in Fundamental Factor Modeling*

1. Winsorize the returns
2. Transform the loadings
3. Select the estimation universe
4. Regress returns against loadings
5. Estimate volatilities and performances

Whereas step 3 is relatively straightforward, the last step is at the core of the research process.

The next two sections cover the essentials of factor models: regression and covariance estimation. Winsorization comes next. Finally, the process of identifying the characteristics receives its own final, long section.

8.2 Cross-Sectional Regression

The first step is cross-sectional regression

$$\mathbf{r}_t = \mathbf{B}_t \mathbf{f}_t + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{N} \quad (8.1)$$

where the parameters to be estimated are \mathbf{f}_t and $\boldsymbol{\epsilon}_t$. This is a case of *random design*: the tuple $(\mathbf{r}_t, \mathbf{B}_t, \mathbf{f}_t, \boldsymbol{\epsilon}_t)$ can be viewed as independent samples drawn from a common distribution. We observe $(\mathbf{r}_t, \mathbf{B}_t)$, and we estimate $(\mathbf{f}_t, \boldsymbol{\epsilon}_t)$.

Several regression approaches are possible. One may minimize the square loss $\|\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t\|^2$. The assumptions behind this step are:

1. The matrix $\mathbf{B}_t \in \mathbb{R}^{n \times m}$ has full rank. A necessary but not sufficient condition for this is that $m \leq n$.
2. Residual returns $\boldsymbol{\epsilon}_{t,i}$ have zero-mean, are homoskedastic (i.e., have the same variance) and are independent of each other.
3. Factor returns and residual returns are independent of each other.
4. Factor and residual returns are “well-behaved”, in the sense of having at least finite fourth moments.

These assumptions can be relaxed. If the matrix is rank-deficient, the solution to the minimum-norm problem exists but is not unique, and factor returns are not identified. Later in this section we will introduce ways to deal with rank-deficient matrices, in order to have a unique solution. Homoskedasticity is also not a necessary assumption. If residuals have different variances for different assets, we can weight the losses for each assets differently. The intuition is that, if the residual return of an asset has a large variance, we should weight the loss for that asset less, so that this single term does not dominate the sum of losses, and unduly affects the parameters’ estimation.

We estimate Model (8.1) by minimizing the sum of a loss function $L : \mathbb{R}^n \rightarrow \mathbb{R}^+$ and a penalty term $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^+$:

$$\min L(\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t)$$

In this section, we choose to minimize the weighted sum of the residuals. We know a diagonal, positive matrix \mathbf{W} , whose diagonal terms can be interpreted as weights assigned to observation (i.r., asset) i . We then find \mathbf{f} that minimizes

$$L(\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t) := (\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t)' \mathbf{W} (\mathbf{r}_t - \mathbf{B}_t \mathbf{f}_t) \quad (8.2)$$

There are good reasons for this choice. If we assume that Model (8.1) is the true model of returns, then Least Squares gives the lowest-variance *unbiased* estimate among all the linear models (Hansen, 2022). The lack of bias matters for performance attribution and alpha identification. Even a small bias in factor (and, consequently, in residual returns) would accumulate over the course of a multi-period performance attribution, thus distorting the results and the insights from the analysis. An additional benefit of Weighted Least Squares regression is that its estimates have a natural interpretation in terms of *factor-mimicking portfolios*. We will cover these in details later. For now, it should suffice to say these are investable portfolio whose returns track as well as possible the true—but unobservable—factor returns. Our recommendation therefore is to use this loss function at least as a starting point, and to run thorough diagnostics to identify its possible shortcomings. At the very least we

To fix ideas, we make the assumption that the model is $\mathbf{r}_t = \mathbf{B}\mathbf{f}_t + \boldsymbol{\epsilon}_t$, with $\mathbf{f}_t \sim N(0, \boldsymbol{\Omega}_f)$ and $\boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{\Omega}_\epsilon)$. The two crucial assumptions here are:

1. Residual returns are assumed to be heteroskedastic. We can address the issue of heteroscedasticity by premultiplying both sides of the equation $\mathbf{r}_t = \mathbf{B}_t \mathbf{f}_t + \boldsymbol{\epsilon}_t$ by matrix $\boldsymbol{\Omega}_\epsilon^{-1/2}$. Now idiosyncratic returns are homoskedastic. We use the Ordinary Least Squares loss function $\|\boldsymbol{\Omega}_\epsilon^{-1/2}(\mathbf{r}_t - \mathbf{B}\mathbf{f}_t + \boldsymbol{\epsilon}_t)\|$, which is equivalent to the loss function in Equation (8.2), with a weight matrix $\mathbf{W} = \boldsymbol{\Omega}_\epsilon^{-1}$.
2. Factor loadings are assumed to be constant over time. This simplifies the formulas below, but can be relaxed by simply regressing the returns on the time-varying loadings.

Given \mathbf{B} , $\boldsymbol{\Omega}_\epsilon$, the Gaussian likelihood is given by

$$\prod_{t=1}^T \frac{1}{(2\pi)v|\boldsymbol{\Omega}_\epsilon|} \exp\left(-\frac{1}{2}(\mathbf{r}_t - \mathbf{B}\mathbf{f}_t)' \boldsymbol{\Omega}_\epsilon^{-1} (\mathbf{r}_t - \mathbf{B}\mathbf{f}_t)\right)$$

If we denote the matrix of returns $\mathbf{R} \in \mathbb{R}^{n \times T}$, the log-likelihood is equivalent to $-\|\mathbf{R} - \mathbf{BF}\|_{\boldsymbol{\Omega}_\epsilon^{-1}}^2$. We write the optimization problem as

$$\begin{aligned} \min & \quad \left\| \boldsymbol{\Omega}_\epsilon^{-1/2}(\mathbf{R} - \mathbf{BF}) \right\|^2 \\ \text{s.t.} & \quad \mathbf{F} \in \mathbb{R}^{m \times T} \end{aligned}$$

whereas consider first the case of a single period. In this case \mathbf{R} and \mathbf{F} are column vectors. The solution is the ordinary least squares solution: $\mathbf{F} = (\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{R}$. In the case of multiple periods, the problem is the sum of the single-period problems:

$$\|\mathbf{R} - \mathbf{BF}\|_{\Omega_\epsilon^{-1}}^2 = \sum_t \left\| \Omega_\epsilon^{-1/2}(\mathbf{r}_t - \mathbf{B}\hat{\mathbf{f}}_t) \right\|^2$$

Each term can be minimized independently. Hence we have

$$\hat{\mathbf{f}}_t = (\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{r}_t \quad (8.3)$$

The problem of minimizing $\|\mathbf{A} - \mathbf{BX}\|$ has a closed-form solution:

$$\arg \min_{\mathbf{X}} \|\mathbf{A} - \mathbf{BX}\|_F^2 = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{A} \quad (8.4)$$

8.2.1 Rank-Deficient Loadings Matrices

In some cases the loadings matrix is rank-deficient: even if there are m factors, the number of independent columns is $p < m$. As concrete (and very common!) examples, consider the following two:

- There is a factor with loadings for each asset equal to 1. This is sometimes called a “country”, “region” or “universe” factor, since all assets are identically affected by changes in this factor. The interpretation is that this is an “intercept” term in the regression. However, the same loadings matrix contains at the very least industry loadings, which can be interpreted as non-negative weights summing to one. For simplicity, assume that the first factor is the country, and the next $m - 1$ are industries. Then the vector $\mathbf{v} = (1, -1, -1, \dots, -1)'$ is such that $\mathbf{B}\mathbf{v} = 0$. The matrix is rank-deficient.
- In most multi-country models there are industry as well as country loadings. Say the first m_{ind} are country loadings, followed by m_{ctry} country factors. The vector $\mathbf{v} = (1, 1, \dots, 1, -1, -1, \dots, -1)'$, where the first m_{ind} are ones and the remaining m_{ctry} are negative ones, also annihilates \mathbf{B} .

We generalize this to the case where are $m - p$ independent vectors \mathbf{v}_i such $\mathbf{B}\mathbf{v}_i = 0$. Because of this deficiency, $\mathbf{B}'\mathbf{B}$ is not invertible and it is not possible to estimate $\hat{\mathbf{f}}$ using Equation (3.1). There are at least three ways to address such an issue.

- The first one is to remove the redundancy. For example, remove one industry and/or a country and/or a country factor. The benefit is that we can reuse a familiar formula. The drawback is that the original loading matrix is easier to interpret. We would like to know a portfolio exposure to the country *and* to all industries. The country exposure is telling us whether the portfolio is long or short, an information which the individual industries exposures don't immediately convey. And, of course, all countries are useful. Just ask the portfolio manager whose main covered industry was removed from the model.
- The second one is to add a small quadratic penalty term to $\|r - Bf\|^2$, i.e., $\delta \|f\|^2$. This removes the degeneracy. The factor estimates are no longer unbiased in the linear model, so a careful analysis would be needed before using this method.
- Finally, we can add $m - p$ side constraint of the form $C'f = a$ ([Heston and Rouwenhorst, 1994, 1995](#)) and solve a constrained linear regression problem. This adds some (minor) complexity to the estimation process, but maintains or even enhances the interpretability of factor returns. For example, we may require that the market-weighted sum of industry factor returns be zero. This would be written as $w'B_{ind}f_{ind} = 0$, where B_{ind} is the column subset of industry factors, f_{ind} is the subset industry factor returns, w is a weight vector of asset market caps per asset. The constraint says “the sum across assets of market-weighted industry returns must be 0”. If w are chosen to be the weights of a benchmark portfolio, this can be read as “the benchmark portfolio must have no industry returns”.

8.2.2 Conditions for Constrained Identification*

Before we move on, we address two issues related to estimation. You won't encounter these issues except in pathological cases. First, what are the valid constraints $C'f = a$ that remove the degeneracy? Second, are different choices of the constraint resulting in different risk models?

Start with the problem $\min_f \|r - Bf\|^2$, whose First-Order Necessary Condition (FONC) is $(B'B)f = B'r$. When B is rank-deficient, the set of solution is given by $f^* + g$, where f^* is the minimum-norm f that solves the FONC, and g is in the null space of B , i.e., $Bg = 0$. We can represent all elements of this

space as a linear combination of the v_i defined above, so $\mathbf{g} = \mathbf{V}\mathbf{x}$, where

$$\begin{aligned}\mathbf{V} &\in \mathbb{R}^{m \times (m-p)} \\ \mathbf{x} &\in \mathbb{R}^{m-p} \\ \mathbf{BV} &= 0\end{aligned}$$

When we impose a constraint $\mathbf{C}'\mathbf{V}\mathbf{f} = \mathbf{a}$, we fix the values of \mathbf{g} :

$$\begin{aligned}\mathbf{C}'\mathbf{f} &= \mathbf{a} \\ \Rightarrow \mathbf{g}^* &= \mathbf{V}(\mathbf{C}'\mathbf{V})^{-1}(\mathbf{a} - \mathbf{C}'\mathbf{f}^*) \\ \Rightarrow \mathbf{f}^* + \mathbf{g}^* &= (\mathbf{I}_m - \mathbf{V}(\mathbf{C}'\mathbf{V})^{-1}\mathbf{C}')\mathbf{f}^* + \mathbf{V}(\mathbf{C}'\mathbf{V})^{-1}\mathbf{a}\end{aligned}$$

This has a unique solution as long as $\mathbf{C}'\mathbf{V} \in \mathbb{R}^{(m-p) \times (m-p)}$ is non-singular. This is a necessary condition for the constraint.

When $a \neq 0$, the expected value of the factor is changed by the offset $V(C'V)^{-1}a$. This offset is irrelevant for investment purposes because

$$B(V(C'V)^{-1}a) = 0$$

and is irrelevant for covariance estimation purposes because it is constant. Factor returns are also modified by the matrix $H := (I_m - V(C'V)^{-1}C')$. H has the property that $BH = B$. The empirical factor covariance dependent on the constraint C becomes

$$\hat{\boldsymbol{\Omega}}_{f^*+g^*} = H\hat{\boldsymbol{\Omega}}_{f^*}H'$$

This is a different covariance matrix. But the difference is not relevant because, from direct calculation, the factor volatility for any portfolio is unchanged:

$$w'B[\hat{\boldsymbol{\Omega}}]_{f^*+g^*}B'w = w'BH\hat{\boldsymbol{\Omega}}_{f^*}H'B'w = w'B\hat{\boldsymbol{\Omega}}_{f^*}B'w$$

because of the equation $BV = 0$.

8.3 Estimating The Factor Covariance Matrix

We have a random vector of factor returns $\hat{\mathbf{f}}_t$, from which we want to estimate $\boldsymbol{\Omega}_{\mathbf{f}}$. We assume that the $\mathbf{f}_{t,i}$ have fourth moments, but unlike in the chapter on statistical model estimation, we cannot assume that $\boldsymbol{\Omega}_{\mathbf{f}}$ has a special structure. By construction, we do not expect the matrix to be spiked. The number of samples over which we estimate the covariance matrix is can be larger than the

number of factor; for example, we could estimate a model with 10 factors and 500 days of estimation. The assumptions m constant, $T \rightarrow \infty$ seem appropriate.

Let $\boldsymbol{\Omega}_T^{\text{emp}} := T^{-1} \sum_{t=1}^T \hat{\mathbf{f}}_t \hat{\mathbf{f}}_t'$. By the Law of Large Numbers, $\boldsymbol{\Omega}_T^{\text{emp}} \rightarrow \boldsymbol{\Omega}_f$ almost surely. Both eigenvalues and eigenvectors converge to the covariance matrix. See Section 14.2 in the Appendix. Factor volatilities converge to the true (also denoted *population*) volatilities, and the relative standard error is $\sqrt{2/T}$. The principal components of the factor covariance matrix also converge to their population counterpart, so long as the volatilities of factors are all sufficiently separated. This seems to settle the issue of covariance estimation: just take the empirical covariance matrix. There are two problems, though:

- Oftentimes, though, the number of factors is not much smaller than the number of observations. In this case, shrinkage may improve the quality of the estimate.
- We will see that factor return estimates are inflated by the estimation process. This is another argument in favor of shrinkage.
- Factor returns are nonstationary, sometimes dramatically so at the onset of a crisis. We need to take this into account.
- Factor returns are mildly autocorrelated. We need to correct for that.

8.3.1 Factor Covariance Shrinkage

The first one lies in the fact that the factor return estimates $\hat{\mathbf{f}}_t$ are just that: estimates. They are the outcome of WLS linear regression estimates, Equation (3.16). The standard error of the estimates $\hat{\mathbf{f}}_t$ is $(\mathbf{B}'\boldsymbol{\Omega}_{\epsilon}^{-1}\mathbf{B})^{-1}$. This implies that

$$\text{var}(\hat{\mathbf{f}}_t) = \boldsymbol{\Omega}_f + (\mathbf{B}'\boldsymbol{\Omega}_{\epsilon}^{-1}\mathbf{B})^{-1} \quad (8.5)$$

How big is the correction? In the simpler but instructive case $\mathbf{B}'\mathbf{B} = \mathbf{I}_m$, $\boldsymbol{\Omega}_{\epsilon} = \mathbf{I}_n$, the estimated factor returns are $\hat{\mathbf{f}}_t = \mathbf{B}'\mathbf{r}_t$, and $\text{var}(\hat{\mathbf{f}}_t) = T^{-1} \sum_t \hat{\mathbf{f}}_t \hat{\mathbf{f}}_t' = \mathbf{B}'\hat{\boldsymbol{\Omega}}_r\mathbf{B}$, and therefore

$$\hat{\boldsymbol{\Omega}}_f = \mathbf{B}'\hat{\boldsymbol{\Omega}}_r\mathbf{B} - \mathbf{I}_m \quad (8.7)$$

In applications the number of factors ranges between 1 and 100 and the number of periods ranges between 126 (six months, for daily returns) and 500 (two years of daily returns); therefore we are not always in the regime $p \ll T$ and

Insight 8.1: Factor-Mimicking Portfolio Interpretation

An alternative lens to interpret this result is via factor-mimicking portfolios. The return of factor-mimicking portfolio \mathbf{w}_i is $\hat{f}_i = f_i + \epsilon' \mathbf{w}_i$. The covariance of the returns of FMP i and j , using Equation (4.11), is $\text{cov}(\hat{f}_i, \hat{f}_j) = \text{cov}(f_i, f_j) + [(\mathbf{B}' \Omega_\epsilon^{-1} \mathbf{B})^{-1}]_{i,j}$, which is Equation (8.5). This suggests that we should shrink the empirical covariance matrix in order to obtain an unbiased estimate:

$$\hat{\Omega}_f = \text{var}(\hat{\mathbf{f}}_t) - (\mathbf{B}' \Omega_\epsilon^{-1} \mathbf{B})^{-1} \quad (8.6)$$

the asymptotics of Section 14.2 do not apply; neither do results for spiked covariance matrix. A popular shrinkage applied to covariance matrix is Ledoit-Wold Shrinkage (Ledoit and Wolf, 2003a,b, 2004). It has the advantage of being simple to implement and to optimize, and with good performance. The shrinked covariance matrix is

$$\Omega_{\text{shrink}}(\rho) = (1 - \rho)\hat{\Omega}_f + \rho \frac{\text{trace}(\hat{\Omega}_f)}{m} \mathbf{I}_m$$

which we combine with Equation (8.6):

$$\Omega_{\text{shrink}}(\rho) = (1 - \rho)(\hat{\Omega}_f - (\mathbf{B}' \Omega_\epsilon^{-1} \mathbf{B})^{-1}) + \rho \frac{\text{trace}(\hat{\Omega}_f - (\mathbf{B}' \Omega_\epsilon^{-1} \mathbf{B})^{-1})}{m} \mathbf{I}_m$$

where $\rho \in (0, 1)$ is a tunable parameter.

8.3.2 Dynamic Conditional Correlation

An alternative and common approach to estimating the empirical covariance matrix $\text{var}(\hat{\mathbf{f}}_t)$ is to model the factor volatility and correlations separately. Namely, we decompose the population covariance matrix in the product of a correlation matrix C and a diagonal matrix V containing the factor volatilities:

$$\Omega_f = \mathbf{V} \mathbf{C} \mathbf{V}$$

Bollerslev (1990) modeled the volatilities as time-varying and the correlation matrix as constant. Practitioners estimate the empirical correlation matrix and

the volatility vector using exponential weighted averages with different half-lives.

$$\text{diag}(\mathbf{V}_t^2) = \kappa_V \sum_{s=0}^T e^{-s/\tau_V} \hat{\mathbf{f}}_{t-s} \circ \hat{\mathbf{f}}_{t-s}$$

$$\mathbf{C} := \kappa_C \sum_{s=0}^T e^{-s/\tau_C} \mathbf{V}_{t-s}^{-1} \hat{\mathbf{f}}_{t-s} \hat{\mathbf{f}}'_{t-s} \mathbf{V}_{t-s}^{-1}$$

with $\tau_C > \tau_V$ being normalizing constants. In many equity models estimated using daily returns, the half lives are set between three months (for exceptionally responsive variance estimations) and two years.

8.3.3 Short-Term Factor Updating

Estimated factor returns often exhibit large, unanticipated values. Anecdotally, their volatility does not vary smoothly but discontinuously, with regimes of high volatility, followed by a quick transition to low-volatility regimes. This poses two challenges for the modeler. First, a simple exponential weighted estimator will react too slowly to sudden increases in volatility. The very concrete effect of this is that the investors will severely underestimate systemic risk at the time when they need accurate estimates the most. Secondly, the estimates react too slowly to *reductions* in volatility. By the nature of the weighting scheme, volatilities decay no faster than exponentially, with half lives of several months. Several approaches have been proposed to address this issue. We mention one that performs well and is simple to implement: Short-Term Factor Updating (STFU). First, we model the multivariate factor returns so that they are modulated by a latent state variable:

$$\mathbf{f}_t = e^{x_t/2} \mathbf{V}_t \mathbf{C}_t^{1/2} \boldsymbol{\eta}_t \quad (8.8)$$

$$\boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{I}_n)$$

$$x_{t+1} = \phi x_t + \sigma \gamma_t$$

In the degenerate case where $x_t = 0$ for all t , the model reduces to one where the factor covariance matrix is $\boldsymbol{\Omega}_f$.

Define $\mathbf{u}_t := \mathbf{C}_t^{-1/2} \mathbf{V}_t^{-1} \mathbf{f}_t$. Then $x_t = \log \|\mathbf{u}_t\|^2 - \log \|\boldsymbol{\eta}_t\|^2$. This is a linear

state-space model. Define

$$\begin{aligned}\kappa &:= E(\log \|\boldsymbol{\eta}_t\|^2) \\ \epsilon_t &:= \kappa - \log \|\boldsymbol{\eta}_t\|^2 \\ x_t &:= \xi_t \\ y_t &:= \log \|\mathbf{u}_t\|^2 - \kappa\end{aligned}$$

The state-space equation is:

$$\begin{aligned}y_t &= x_t + \epsilon_t \\ x_{t+1} &= \phi x_t + \sigma \gamma_t\end{aligned}$$

and the estimate of the state takes the form

$$e^{\hat{x}_t/2} = \kappa_0 \exp \frac{1}{2} \left(\sum_{s=0}^{\infty} e^{-s/\tau_0} (\log \|\mathbf{u}_{t-s}\|^2 - \kappa) \right)$$

With $\kappa_0 = \sum_{s=0}^{\infty} e^{-s/\tau_0} = \exp(1/\tau_0)/(\exp(1/\tau_0) - 1)$. From the first equation in model (8.8), the factor covariance matrix is then adjusted by multiplying by the factor $e^{\hat{x}_t}$.

Some implementations use the linear approximation of this formula and the approximate equality $\kappa = \log(m) + E \log(\|\boldsymbol{\eta}_t\|^2/m) \simeq \log m$ since $\|\boldsymbol{\eta}_t\|^2/m \rightarrow 1$ a.s. for $m \rightarrow \infty$.

$$\begin{aligned}e^{\hat{x}_t/2} &\simeq \kappa_0 \exp \left(\frac{1}{2} \sum_{s=0}^{\infty} e^{-s/\tau_0} \left(\frac{\|\mathbf{u}_{t-s}\|}{\sqrt{m}} - 1 \right) \right) \\ &\simeq \kappa_0 \exp \left(\left(\sum_{s=0}^{\infty} e^{-s/\tau_0} \frac{\|\mathbf{u}_{t-s}\|}{\sqrt{m}} \right) - 1 \right) \\ &\simeq \kappa_0 \sum_{s=0}^{\infty} e^{-s/\tau_0} \frac{\|\mathbf{u}_{t-s}\|}{\sqrt{m}}\end{aligned}$$

The interpretation of the formula is clearest in the special case of uncorrelated factor returns. In this case, \mathbf{u}_t is the vector of z-scored returns. If $x_t = 1$, they have unit variance. If we view the random variables u_i as iid samples of a random variable, the term $\|\mathbf{u}_{t-s}\|/\sqrt{m}$ gives us an estimate of its standard deviation, and if this estimate exceeds one, then our original estimates for the standard deviations of the factor need to be revised upward. That is what the model does. The half-life τ_0 is typically between 10 and 20 days for daily risk models, in order to incorporate the rapid onset of a shock.

8.3.4 Correcting for Autocorrelation in Factor Returns

Daily factor and asset returns usually exhibit mild, but non-zero, short-term autocorrelation. When the factor covariance matrix is estimated on shorter time intervals, the autocorrelation may be more pronounced. In these cases, adjusting for autocorrelation improves the model's performance. Cohen et al. (1983) build on previous work by Scholes and Williams (1977) and assume that the observed returns follow an autoregressive process of order l_{\max} that is function of underlying uncorrelated returns. The coefficient in the $AR(l_{\max})$ equation are random, but sum to 1. Let the lagged covariance matrix C_l be defined as

$$(C_l)_{i,j} := \text{cov}(\mathbf{f}_{t,i}, \mathbf{f}_{t-l,j})$$

Then the autocorrelation-consistent estimator is given by

$$\Omega_f = \hat{\Omega}_f + \frac{1}{2} \sum_{l=1}^{l_{\max}} (C_l + C'_l)$$

An alternative approach, which is asymptotically consistent in the limit $T \rightarrow \infty$ is Newey and West's estimator (Newey and West, 1987):

$$\Omega_f = \hat{\Omega}_f + \frac{1}{2} \sum_{l=1}^{l_{\max}} \left(1 - \frac{l}{l_{\max}} \right) (C_l + C'_l)$$

8.4 Estimating the Idiosyncratic Covariance Matrix

Next, we need to estimate the covariance matrix Ω_ϵ based on the period estimated idiosyncratic returns $\hat{\epsilon}_t$.

8.4.1 Exponential Weighting

As in the case of factor volatility, we use exponential weighting for idiosyncratic volatility estimation. Let $\mathbf{E} \in \mathbb{R}^{n \times T}$ be the matrix of estimated idiosyncratic returns, with $[\mathbf{E}]_{i,t} := \epsilon_{i,t}$. The exponential weighting parameter is the half-life τ . The weighting matrix is $\mathbf{W} \in \mathbb{R}^{T \times T}$, with $\mathbf{W} := \text{diag}(\kappa e^{-1/\tau}, \dots, \kappa e^{-T/\tau})$, and $\kappa := (1 - e^{1/\tau}) / (1 - e^{-t/\tau})$ is a normalizing constant. The EWMA empirical idiosyncratic covariance matrix is then $\hat{\Omega}_\epsilon := \mathbf{W}\mathbf{E}'\mathbf{E}\mathbf{W}$.

8.4.2 Visual Inspection

This matrix should be diagonal, or at least sparse. The sample covariance matrix based on estimated returns does not satisfy these requirements. The sample covariance matrix $\hat{\Omega}_\epsilon$ is neither sparse nor positive definite, since $T < n$. We (the modelers) could set all the non-diagonal terms to zero, which effectively amounts to a radical shrinkage of the idiosyncratic correlation matrix. This step, however, is not warranted. As a sanity check it is always recommended to perform a visual inspection of the empirical covariance matrix. Oftentimes, there are striking patterns that can be interpreted as factors that should be added to the model. For example, some Chinese securities are listed both in Mainland China (A and B Shares) and in Hong Kong (H shares). These securities have highly correlated but not identical returns, and the correlations will show up in the idiosyncratic covariance matrix. In such a case, rather than assuming that A, B, H shares are identical (they are not), it is more appropriate to add a “share class” factor to the model.

8.4.3 Short-Term Idio Update

Idiosyncratic returns, like factor returns, are subject to sudden changes in volatility that are not captured well by exponential weighting with long half-lives τ . A very responsive daily return model has $\tau = 126$ trading days, and the shocks may occur over ten trading days. As a remedy, we reuse the SFTU machinery, but with one minor but important modification. We use as an example the case of equities, even though the technique is easily applicable to other asset classes. Stocks are likely to receive large shocks in proximity of earnings, either because new information is released before or on earnings date, or because such information becomes fully priced in the following days. We introduce tent-shaped variables $a_{i,t}$. Let $T_{\text{earn},i}$ be the earning date, and τ_{earn} be a time horizon during which earnings information is received. Define the function as

$$a_{i,t} = \begin{cases} 1 - |t - T_{\text{earn},i}|/\tau_{\text{earn}} & \text{if } |t - T_{\text{earn},i}| \leq \tau_{\text{earn}} \\ 0 & \text{otherwise} \end{cases}$$

$a_{i,t}$ ranges from zero to one, when t is within τ_{earn} number of days from the earnings date $t_{i,\text{earn}}$. We use Model (8.8), but restricting our updates to those stocks within the earnings announcement window. The SFTU model is

somewhat simplified by the fact that the correlation matrix is approximately diagonal. We restrict our attention to the linear approximation: the corrective term is

$$e^{\hat{x}_t/2} = \kappa_0 \sum_{s=0}^{\infty} e^{-s/\tau_0} \sqrt{\frac{\sum_i a_{i,t} (\epsilon_{i,t}/\hat{\sigma}_{i,t})^2}{\sum_i a_{i,t}}}$$

and applies only to the assets affected by earnings.

$$\hat{\sigma}_{i,t}^2 \leftarrow [(1 - a_{i,t}) + a_{i,t} e^{\hat{x}_t}] \hat{\sigma}_{i,t}^2$$

8.4.4 Off-Diagonal Clustering

Finally, we need to identify those assets whose idiosyncratic returns are highly correlated. Two instances are important. The first one is the case of different securities that refer to the same underlying asset. For example, some stocks are listed as different share classes; for example, Berkshire Hathaway trades under BRK.A and BRK.B, with different fractional values. The liquidity of the two securities differs; yet, their returns are highly correlated. Whether to include the two securities in a factor model or not depends on the nature of the trading strategy employing the model itself. If the strategy intends to exploit the temporary small mispricing between two assets, then we should include both assets. If instead we only intend to invest in the company based on fundamental information, then we should only include a security representative of the underlying asset; typically we choose the most liquid asset. The second instance instead has to do with stocks whose dependencies are not described by factors. In order to be identifiable, factors must be pervasive. A factor influencing only a handful of assets is not a factor, and cannot be identified in a large cross-section of assets. The dependency among these stocks is still detectable in the correlation between their idiosyncratic returns. To identify them, we resort to *correlation thresholding*. We transform the correlation elements by applying a simple clipping operator: $\text{thres}_\lambda(\rho_{i,j}) := \rho_{i,j} \mathbf{1}\{|\rho_{i,j}| > \lambda\}$. The optimal threshold λ is $K \sqrt{\log n/T}$, for some positive constant K . In practice, however, it is more instructive to explore the clusters emerging for different values of the threshold. For some value of λ , the clusters are a) stable, in the sense they do not change much for perturbed values of the threshold; b) interpretable, in the sense that they are comprised of “similar” stocks, in the sense that they belong to the same sector or industry, and sometimes they have similar style factor loadings

as well¹. It is important to check that, for every level of the threshold, the correlation matrix is positive definite (and well-conditioned). As an example, we use the residual returns from a commercial factor model (Axioma US V.4, Short Horizon, AXUS4SH). We use the residual returns for year 2010. We compute the equal-weighted correlations for continuous constituents of the Russell 3000 index, and threshold their absolute values at 0.55. Graph 8.1 shows the resulting clusters. The number of stocks is small: less than 90, out of a set of nearly 2,900 stocks; about 3%. Table 8.1 lists tickers and associated names. The pairs are quite intuitive: Visa and MasterCard; Wynn and Lass Vegas Sands; Peabody Energy and Alpha Natural Resources; and a mining cluster composed of HL, NEM, CDE, RGLD.

Table 8.1: Ticker and company names of cluster components in Figure 8.1.

Ticker	Name	Ticker	Name
AAI	AIRTRAN HOLDINGS	HOT	STRW HTL RES WRLWD
ACTI	ACTIVIDENTITY CORP	KG	KING PHARMCUTCL
ACV	ALBERTO CULVER CO NEW	LVS	LAS VEGAS SANDS
ALKS	ALKERMES INC	MA	MASTERCARD INC
ALTR	ALTERA CORP	MAR	MARRIOTT INTL INC NEW
ALY	ALLIS CHALMERS ENERGY INC	MDVN	MEDIVATION INC
AMLN	AMYLIN PHARMACEUTICALS INC	MFE	MCAFEE INC
ANR	ALPHA NATURAL RES	MNTA	MOMENTA PHARMACEUTICALS INC
ARQL	ARQUELE	MTG	MGIC INVT CORP WIS
ATSG	AIR TRANSPORT SERVICES GRP I	MWW	MONSTER WORLDWIDE INC
AVNR	AVANIR PHARMACEUTICALS INC	NAL	NEWALLIANCE BANCSHARES INC
BTH	BLYTH INC	NCI	NAVIGANT CONSULTING INC
BTU	PEABODY ENERGY	NEM	NEWMONT MINING CORP
CASY	CASEYS GEN STORES INC	NOVL	NOVELL INC
CCL	CARNIVAL CORP	PMI	PMI GROUP INC
CDE	COEUR D ALENE MINES CORP IDAHO	QLGC	QLOGIC CORP
RCL	ROYAL CARIBBEAN CRUISES LTD	RDN	RADIAN GROUP INC
CMTL	COMTECH TELECOMMUNICATIONS CP	RGLD	ROYAL GOLD INC
CYH	COMMUNITY HEALTH SYS INC NEWCO	RVI	RETAIL VENTURES INC
DSW	DSW-A	SUR	CNA SURETY CORP
DYN	DYNEGY INC DEL	SVR	SYNIVERSE HLDGS INC
FMR	FIRST MERCURY FIN	V	VISA INC
FTO	FRONTIER OIL CORP	WL	WILMINGTON TRUST CORP
GCA	GCA HLDGS	WYNN	WYNN RESORTS
HL	HECLA MNG CO	XCO	EXCO RESOURCES INC
HMA	HEALTH MGMT ASSOC INC NEW	XLNX	XILINX INC
HOC	HOLLY CORP		

8.4.5 Shrinking of Variances

¹We could characterize more rigorously this within-cluster similarity as a distance in among factor loadings, and given this similarity measure, propose a more systematic thresholding procedure, but it would fall beyond the scope of the book.

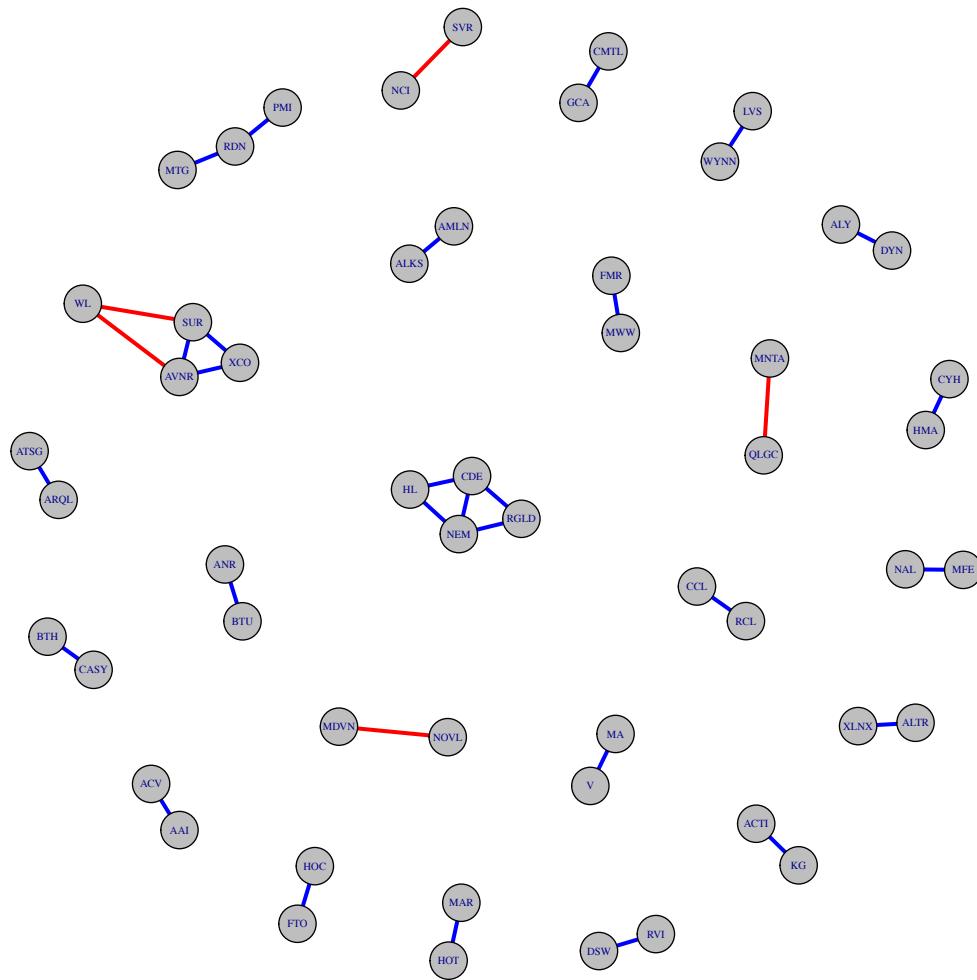


Figure 8.1: Clusters for idiosyncratic matrix. Blue links (darker) denote positive correlation; red links (lighter) negative correlation.

8.5 Winsorization of Returns

The issue of outlier detection is, if not central, at the very least very important both for risk modeling and alpha research. There are many instances of outliers in return data, each one of them responsible for ruining the career of a finance researcher. Before proposing some remedial measures to improve the research process and save a few careers, let us discuss where they come from, and

the impact that they may have. The sources of outliers are primarily three. First, the data provider may be providing generally low-quality data. This is, unfortunately, quite common, and good researchers spend a large proportion of their time evaluating and comparing data and questioning providers on the data collection methodology and their applicability. The sources of error are too many to list. In the worst case, prices are not correctly adjusted for stock splits or reverse split. Data collection may not be synchronous; the ultimate source of the returns may be a broker located in a farming village in New Zealand, and even more unlikely instances. Bad providers are the perfect breeding ground for outliers. Secondly: authentic outliers do exist. Instances:

- There are rare stray transaction for a liquid stock².
- Very illiquid stocks exhibit higher volatility, and occasionally large returns, even intraday.
- Stocks in the process of being delisted, or entering/exiting bankruptcy, usually trading over the counter (OTC) also exhibit very large returns.
- There are genuinely large jumps reflecting new information in the market: surprises in earnings and forward guidance; announcement of market entry by a competitor; merger announcements and news of merger break; accidents or likely liability; macro-economic drivers.

Of these instances, the first one can be dealt with by inspecting price data carefully; i.e., by not only taking the last price of a five-minute interval, but by inspecting the entire price trace. The second and third can be avoided by choosing the estimation and investment universe carefully. Microcap stocks that are very illiquid should be excluded. The last class of outliers, however large they may be, should *not* be excluded from the estimation process. The main rationale is that this exclusion will make the output of the analysis much less reliable. If we winsorize a large absolute return, we affect the estimated factor returns from the cross-sectional regression. A factor's return is the return of its mimicking portfolio, and by winsorizing returns in the cross-sectional regression, it is effectively the portfolio return using winsorized data. The true portfolio return unfortunately is based on historical returns. This affects the evaluation of

²If I can self-indulge in a personal recollection: I was working for Enron in the summer of 2000. One day, at the closing auction, Enron's stock price jumped from \$90 after vibrating around \$86 all day. And people *on Enron's trading floor* were openly wondering "should we short Enron now?". The stock fell back to mid-80s the day after. Lesson: if traders want to short themselves, then it's a likely outlier of some kind.

the factor returns, as well as of the idiosyncratic returns. Two easy qualitative recommendations that follow from all of this is are:

1. whatever winsorization method you use, make sure to report all the instances of winsorized data in a backtest or in production, and examine them one by one.
2. make sure that your investment universe comprises only liquid, tradable assets.

For the remaining assets, use a robust outlier detection strategy. There is no ideal and completely justified method. A method that works well enough is to compute at the security level, the robust z-score of return:

$$d_{i,t} = \frac{|\log(1 + r_{i,t})|}{\text{median}(|\log(1 + r_{i,t-1})|, \dots, |\log(1 + r_{i,t-T})|)}$$

and to winsorize returns exceeding a threshold d_{\max} . The threshold depends on asset class, region and other attributes, and is set by trial-and-error between 5 and 10.

8.6 Selecting Factors: the Large Number of Predictor Case

Status: This is too speculative/advanced to make it in the book.

Oftentimes, we encounter the case of large number of predictors.

The primitive is a set of matrices $\mathbf{A}_t \in \mathbb{R}^{n \times p}$, with $p \gg n$, with columns standardize to have unit norm and zero mean, from which we want to derive loadings matrices $\mathbf{B}_t \in \mathbb{R}^{n \times m}$. The matrices \mathbf{A}_t won't do as loadings, since the cross-sectional regressions are undetermined, and, even in the case $p \lesssim n$, would result in poor performance. One natural approach is to determine factors as a linear combination of predictors. As a special case of their approach, we set $\mathbf{B}_t = \mathbf{A}_t \mathbf{C}$, where $\mathbf{C} \in \mathbb{R}^{p \times m}$ is time-invariant. \mathbf{C} needs to be identified. We solve the estimation problem,

$$\begin{aligned} & \min \sum_t \|\mathbf{r}_t - \mathbf{A}_t \mathbf{C} \mathbf{f}_t\|^2 \\ \text{s.t. } & \mathbf{C} \in \mathbb{R}^{p \times m} \\ & \mathbf{f}_t \in \mathbb{R}^m \end{aligned}$$

We solve this problem by considering a simpler instance, and then generalizing it. We solve a single-factor estimation problem:

$$\min \sum_t \|\mathbf{r}_t - \mathbf{A}_t \mathbf{c} f_t\|^2 \tag{8.9}$$

$$\text{s.t. } \mathbf{c} \in \mathbb{R}^p \tag{8.10}$$

$$f_t \in \mathbb{R} \tag{8.11}$$

This can be solved alternating by alternating Least Squares, minimizing over $(f_1, \dots, f_T) \in \mathbb{R}^T$ and $\mathbf{c} \in \mathbb{R}^p$.

Note that we use the Moore-Penrose pseudoinverse, because there is no guarantee that the matrix $\sum_t \mathbf{A}'_t \mathbf{A}_t \in \mathbb{R}^{p \times p}$ is invertible. Each matrix $\mathbf{A}'_t \mathbf{A}_t$ has rank n , and the sum of T matrices has rank at most $n + T$, which is in general smaller than p . Once the solution has been found, we orthogonalize the matrices \mathbf{A}_t :

$$[\tilde{\mathbf{A}}_t]_{\cdot, i} := [\mathbf{A}_t]_{\cdot, i} - \frac{[\mathbf{A}_t]_{\cdot, i}' \mathbf{A}_t \mathbf{c}}{\|[\mathbf{A}_t]_{\cdot, i}\|^2} \mathbf{A}_t \mathbf{c} \tag{8.12}$$

And then we iterate Procedure

Procedure 8.2: *as*

1. Set $i = 0$ and initialize $\mathbf{c}^{(0)}$ such that $\|\mathbf{c}^{(0)}\| = 1$, $\text{tol} > 0$.

2. Estimate

$$f_t^{(i)} = \frac{\mathbf{r}'_t \mathbf{A}_t \mathbf{c}^{(i)}}{((\mathbf{c}^{(i)})' \mathbf{A}'_t \mathbf{A}_t \mathbf{c}^{(i)})}$$

3. Set $\mathbf{x} := \left(\sum_t (f_t^{(i)})^2 \mathbf{A}'_t \mathbf{A}_t \right)^+ \left(\sum_t f_t^{(i)} \mathbf{A}'_t \mathbf{r}_t \right)$.

4. Set $\mathbf{c}^{(i)} = \mathbf{x} / \|\mathbf{x}\|$.

5. If $\|\mathbf{c}^{(i)} - \mathbf{c}^{(i-1)}\| > \text{tol}$, set $i \leftarrow i + 1$ and go to 2.

6. Return $\mathbf{c} = \mathbf{c}^{(i)}$ and $f_t = f_t^{(i)}$.

8.7 Multi-Country Models

8.7.1 Model Linkage

8.7.2 ★Currency Rebasing

In a multi-country model risk factor model the return of an asset is usually expressed in a different currency than the one in which the asset is traded. Consider the case of US-based manager trading a security denominated in Euro. The Dollar/Euro pair is a *currency pair*. In order to purchase a Euro-denominated security, we purchase the currency in which the security is traded (the Euro), also called the *base* currency, and sell dollars, the *quote* currency. The direct exchange rate is the dollar amount needed to buy one Euro. More details on usage: when referring to a currency pair, the ordering is base-quoted. In this case: EURUSD³. Let us denote the *direct exchange rate* in period t by $p_{EURUSD}(t)$. The *indirect exchange rate* is the exchange rate of the reversed pair, and it is equal to the reciprocal of the direct exchange rate. The exchange rate return is defined as the return received by holding the base currency in one

³The currency codes are identified by three letters. The most common currency are USD, EUR, GBP (UK pounds), AUD (Australian dollars), CAD (Canadian dollars), CNY (Yuan Renmibi), JPY (Yen).

period, and is equal to $r_{\text{EURUSD}}^c(t) = [p_{\text{EURUSD}}(t) - p_{\text{EURUSD}}(t-1)]/p_{\text{EURUSD}}(t-1)$. We define the currency return $r_{i,i}^c$ when the base and quote are the same to be zero.

Let us analyze first the realized return of holding EUR in a simple transaction, in which we buy and sell the currency on consecutive days. We denote by r_{USD}^f , r_{EUR}^f the risk-free return in the interval between the two transaction epochs for the two currencies.

- On day 0, we borrow \$1 and purchase $1/p_{\text{EURUSD}}(0)$ EUR.
- On day 1 the EUR holding is worth $(1 + r_{\text{EUR}}^f)/p_{\text{EURUSD}}(0)$. We buy back dollars at the price $p_{\text{EURUSD}}(1)$. The dollar amount we are left with is

$$\begin{aligned} & (1 + r_{\text{EUR}}^f)p_{\text{USDEUR}}(1)/p_{\text{EURUSD}}(0) \\ &= (1 + r_{\text{EUR}}^f)(1 + r_{\text{EURUSD}}) \end{aligned}$$

we then pay our USD loan for an amount $-1 - r_{\text{USD}}^f$. We are left with

$$\begin{aligned} & (1 + r_{\text{EUR}}^f)(1 + r_{\text{EURUSD}}) - 1 - r_{\text{USD}}^f \\ &\simeq r_{\text{EURUSD}}^c + r_{\text{EUR}}^f - r_{\text{USD}}^f \end{aligned}$$

Let us extend this result. Instead of holding the EUR in a cash account, we invest it in a security with *local* return (in EUR) equal to r . Following the same calculations, the return is

$$\begin{aligned} r &:= r^l + r_{\text{EURUSD}}^c - r_{\text{USD}}^f \\ &= r^e + r_{\text{EURUSD}}^{cf} \\ r^e &:= r - r_{\text{EUR}}^f \\ g_{\text{EURUSD}} &:= r_{\text{EURUSD}}^c + r_{\text{EUR}}^f - r_{\text{USD}}^f \end{aligned}$$

The return is the sum of two components: first, the local excess return $r^e := r - r_{\text{EUR}}^f$. Second, the currency return g_{EURUSD} , adjusted by the difference in risk-free rates.

There is yet another identity of interest. This links the currency returns of three (or more) currencies. Let us denote the currencies by an index $i \in \{1, 2, 3\}$. A no-arbitrage argument implies that $p_{k,i}(t) = p_{k,j}(t)p_{j,i}(t)$, and from this relationship and the linear approximation we obtain $r_{k,i}^c = r_{k,j}^c + r_{j,i}^c$, from which the identity holds:

$$g_{k,j} := g_{k,i} - g_{j,i}$$

Now we consider the problem of changing numeraire. For example, we want to change the numeraire from USD to GBP.

$$r^e + g_{EURGBP} = r^e + g_{EURUSD} - g_{GBPUSD} \quad (8.13)$$

Let us say that our factor model contains securities traded in q currencies. The assets total return can be decomposed into the sum of a local return and an exchange rate return:

$$\mathbf{r} = \underbrace{\mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}}_{(local\ factor\ structure)} + \underbrace{\mathbf{C}\mathbf{g}^{k_0}}_{(currency\ factor\ structure)}$$

The matrix $\mathbf{C} \in \mathbb{R}^{n \times q}$ is binary, with $[\mathbf{C}]_{i,j} = 1$ if asset i has reference currency j and 0 otherwise. The return $g_i^{k_0}$ is the return of the base currency i , with quoted currency k_0 . We rebase from currency k_0 to currency k_1 by way of transforming the currency returns. Let $\mathbf{A} \in \mathbb{R}^{q \times q}$ a matrix whose elements are all zeros with the exception of $[\mathbf{A}]_{\cdot, k_1} = 1$.

$$\begin{aligned} \mathbf{g}^{k_1} &= (\mathbf{I}_q - \mathbf{A})\mathbf{g}^{k_0} \\ \Rightarrow \mathbf{r} &= \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon} + \mathbf{C}(\mathbf{I}_q - \mathbf{A})^{-1}\mathbf{g}^{k_1} \end{aligned}$$

We close this section with several comments related to modeling extensions and practical implementation:

- We have ignored the question of modeling the joint distribution of the spot currency returns, \mathbf{g}^{k_0} . One natural avenue is to model those using a factor model, either fundamental or statistical, so that we can express $\mathbf{g}^{k_0} = \mathbf{H}\boldsymbol{\xi} + \boldsymbol{\eta}$. We need to model the relationship only with respect to one quote currency.
- We have also ignored the correlations between \mathbf{f} and \mathbf{g} . They are not to include. The model is

$$\mathbf{r} = \begin{pmatrix} \mathbf{B} & \mathbf{C}_{\mathbf{f}, \mathbf{g}} \\ \mathbf{C}_{\mathbf{g}, \mathbf{g}} & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g}^{k_0} \end{pmatrix} + \boldsymbol{\epsilon}$$

- Currency risk depends heavily on institutional arrangement. For example, an investment firm may have a fixed capital housed in a different country, and trade using only this capital as collateral. The net exposure is fixed. In this case the foreign currency exposure is given by the capital level, which is usually hedged by currency forward contracts.

8.8 A Tour of Factors

This chapter would not be complete without at least a cursory description of fundamental factors. Because factors should explain cross-sectional returns, they feature prominently in the financial literature exploring return anomalies and extensions to the CAPM or the standard Fama-French three-factor model. The literature on these factors is very large; see [Harvey et al. \(2016\)](#); [Green et al. \(2013\)](#); [Ilmanen \(2011\)](#) for surveys.

- *Market.* By far the most pervasive factor in the model, it is usually either the vector e or a vector β of regression coefficients of the asset total return to a “market” factor return (e.g., SPX or RUA in the US). In the first case, the interpretation is that every return is identically affected by the market, and it is left to other factors to capture the dependence on β .
- *Countries and Industries.* Countries and Industries are represented as 0/1 variables summing to 1 for each asset. We consider these factors as homogeneous not only because the information is coded in the same way in the factor loadings, but also because the relative importance of the two has been a subject of intense study both for financial economist and macroeconomists. Aside from the papers by Heston and Rouwenhorst cited above, see also [Brooks and Del Negro \(2005\)](#); [Cavaglia et al. \(2000\)](#); [Berben and Jansen \(2005\)](#); [Puchkov et al. \(2005\)](#); [Miralles Marcelo et al. \(2013\)](#).
- *Momentum.* Stocks that have outperformed (underperformed) their peer over the three to twelve month previous to a given date outperform (underperform) their peers in the future. Jegadeesh and Titman document this anomaly in the academic literature [Jegadeesh and Titman \(1993\)](#). They review 20 years of literature in [Jegadeesh and Titman \(2011\)](#).
 - *value*
 - *low beta/low vol*

8.9 Further Reading

Factor zoo: [Bender et al. \(2013\)](#); [Bryzgalova et al. \(2022\)](#); [Jacobs \(2015\)](#); [Harvey and Liu \(2020\)](#); [Freyberger et al. \(2020\)](#)

Alternative Data: Kolanovic and Krishnamachari (2017)
What Factors Matters? Harvey et al. (2016), Kagan and Tian (2017), Feng et al. (2020), Chen and Velikov (2019), Jacobs (2015) Covariance Thresholding: Bickel and Levina (2008); El Karoui (2008); Cai et al. (2016)

Chapter 9

Statistical Factor Models

In the statistical model framework we assume that we don't know neither the factor returns nor the exposures; we estimate both. The estimation relies of Principal Component Analysis. Starting with [Chamberlain \(1983\)](#), this approach has been motivated using an asymptotic argument: if the number of factors is finite, say m , and if the specific risk stays bounded over bounded portfolios, then when the number of assets is large, there is a clear separation between the m largest eigenvalues and the remaining eigenvalues. The PCA solution constitutes then a good approximation and, in the limit, the PCA solution converges to the true model. In applications, one may question the merit of an approach that, unlike the fundamental and macroeconomic ones, ignores additional information about the firm characteristics or the macroeconomic environment. Developing a statistical model is useful for several reasons:

- *Complementarity.* Using several models helps understand the shortcomings of each individual model. We can project an existing model on the statistical model, or augment it with statistical factors.
- *Optimization.* In a portfolio optimization problem, it is often beneficial to compare solutions in which we have bounded the total factor variance using different models; or, we could include *both* constraints;
- *Data.* In certain asset classes, firm characteristics or relevant macroeconomic factors may not be available. In the case of joint return estimation When only returns are available, statistical models are the only option;
- *Availability at Short Time Scales.* At certain time scales, such as one-or-five-minute intervals, fundamental factors may not be as relevant;
- *Performance.* A statistical model may just outperform the alternatives.

The main disadvantage of statistical models is that its loadings are less interpretable than in the case of alternatives estimation methods. The first factor is usually easy to interpret as the market. The second and third ones *can* find an interpretation. For example, [Litterman and Scheinkman \(1991\)](#) interpret three statistical factors as level, steepness and curvature of the bond yield curve. The situation is not helpless; in Section 9.4 I describe approaches to interpret statistical models. In the words of [Johnson and Wichern \(2007\)](#), “Analyses of principal components are more of a means to an end rather than an end in themselves because they frequently serve as intermediate steps in much larger investigations”. This is perhaps true of all factor models, but is certainly more true with regards to statistical models, because of the possible challenges in interpretation.

This chapter starts with a minimal description of the approach. Then, we take a detour in the real world.

9.1 Statistical Models: The Basics

9.1.1 Best Low-Rank Approximation and PCA

Let $\mathbf{R} \in \mathbb{R}^{n \times T}$ the matrix of observed returns, whose t th column is the vector of returns in period t ; the matrices $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{F} \in \mathbb{R}^{m \times T}$ denote a matrix of loadings and of factor returns respectively. If we wanted to find the loadings and factor returns that minimized the total “unexplained” variation of returns, summed across periods and assets, then we would solve the problem

$$\min_{\mathbf{B}, \mathbf{F}} \|\mathbf{R} - \mathbf{BF}\|_F \tag{9.1}$$

where $\|\cdot\|_F$ is the Frobenius norm. A matrix of the form \mathbf{BF} above has rank less than or equal to m . Conversely, every matrix with rank less or equal than m can be decomposed as \mathbf{BF} (Exercise 9.1). The problem can be restated as

$$\min_{\text{rank}(\hat{\mathbf{R}}) \leq m} \|\mathbf{R} - \hat{\mathbf{R}}\|^2 \tag{9.2}$$

Here, we have not specified whether the norm is Frobenius. It could be Frobenius, but it could be also any unitarily invariant norm¹.

¹These are matrix norms that are invariant for left- and right-multiplication by orthonormal matrices: $\|\mathbf{A}\| = \|\mathbf{UAV}'\|$. Spectral, Frobenius and nuclear norms are unitarily invariant.

This minimization problem was formulated and solved by [Eckart and Young \(1936\)](#) and generalized by [Mirsky \(1960\)](#). We use the *Singular Value Decomposition*² of $\mathbf{R} = \mathbf{USV}'$, with \mathbf{U} , \mathbf{V} square orthonormal matrices of size n and T respectively, and \mathbf{S} a matrix of size $n \times T$, which has positive values (called *Singular Values*) on the main diagonal (i.e., $[\mathbf{S}]_{i,i}$) and zero values elsewhere. The solution to Problem (9.2) is given by $\hat{\mathbf{R}} = \mathbf{US}_m\mathbf{V}'$ where \mathbf{S}_m has the singular values after the m th one set to zero. The solution can also be written ([Golub and Van Loan, 2012](#)) in compact form as $\hat{\mathbf{R}} = \mathbf{U}_m\mathbf{S}_m\mathbf{V}'_m$, where \mathbf{U}_m and \mathbf{V}_m are the matrices obtained taking the first m columns of \mathbf{U} and \mathbf{V} , and \mathbf{S}_m is the square matrix obtained taking the first m columns and m rows of \mathbf{S} . Then, the original Problem (9.1) is solved by setting

$$\mathbf{B} = \mathbf{U}_m \tag{9.3}$$

$$\mathbf{F} = \mathbf{S}_m\mathbf{V}'_m \tag{9.4}$$

As noted in earlier chapters, there are equivalent “rotated” solutions, of the form $\tilde{\mathbf{B}} = \mathbf{BC}$, $\tilde{\mathbf{F}} = \mathbf{C}^{-1}\mathbf{F}$, for some non-singular $\mathbf{C} \in \mathbb{R}^{m \times m}$. For example, this is also a solution:

$$\mathbf{B} = \mathbf{U}_m\mathbf{S}_m \tag{9.5}$$

$$\mathbf{F} = \mathbf{V}'_m \tag{9.6}$$

A related problem, with which many readers are acquainted, is *Principal Component Analysis*. In this setting, we start with a covariance matrix $\hat{\Sigma} \in \mathbb{R}^{n \times n}$. Our goal is to generate a linear combination of the original vectors $\mathbf{r}^1, \dots, \mathbf{r}^T$, i.e., $\mathbf{w}'\mathbf{r}^1, \dots, \mathbf{w}'\mathbf{r}^T$; the vector $\mathbf{w} \in \mathbb{R}^n$ is a vector of weights, normalized to have unit Euclidean norm. We want these random observations $\mathbf{w}'\mathbf{r}^i$ to have the greatest possible variance. With a little work (which we did in previous chapters; or do Exercise 9.4), you can show that this variance is equal to $\mathbf{w}'\hat{\Sigma}\mathbf{w}$. The problem than can be stated as

$$\begin{aligned} & \max \mathbf{w}'\hat{\Sigma}\mathbf{w} \\ & \text{s.t. } \|\mathbf{w}\| \leq 1 \end{aligned} \tag{9.7}$$

The vector \mathbf{w} is called the *first principal component* of $\hat{\Sigma}$. You can interpret Problem (9.7) as a financial problem too: find a maximum-variance portfolio, where the sum of the squared net notional positions is less or equal than 1.

²This is referred to as the Full SVD, as opposed to the reduced SVD; see [Trefethen and Bau \(1997\)](#).

The connection between PCA and eigenvalue problems is well known, but it still useful to highlight it. The Lagrangian of Problem (9.7) is $\nabla_{\mathbf{w}}(\mathbf{w}'\hat{\Sigma}\mathbf{w}) + \lambda\nabla_{\mathbf{w}}(1 - \|\mathbf{w}\|^2) = 2\hat{\Sigma}\mathbf{w} - 2\lambda\mathbf{w}$; a necessary condition for the maximum is that the Lagrangian be zero. This is equal to the eigenvalue equation $\hat{\Sigma}\mathbf{v} = \lambda\mathbf{v}$. From this equation it follows that $\lambda = \mathbf{v}'\hat{\Sigma}\mathbf{v}$. Therefore, the solution is the eigenvector with the highest associated eigenvalue.

Once this maximum-variance portfolio $\mathbf{w}^{(1)}$ has been found, we repeat the process and find another maximum-variance portfolio that is orthogonal to $\mathbf{w}^{(1)}$:

$$\begin{aligned} & \max \mathbf{w}'\hat{\Sigma}\mathbf{w} \\ \text{s.t. } & \|\mathbf{w}\| \leq 1 \\ & \mathbf{w}'\mathbf{w}^{(1)} = 0 \end{aligned} \tag{9.8}$$

To see the relationship between PCA and SVD, we write the uncentered covariance matrix using the SVD decomposition:

$$\hat{\Sigma} = \frac{1}{T}\mathbf{R}\mathbf{R}' = \frac{1}{T}\mathbf{U}\mathbf{S}^2\mathbf{U}' \tag{9.9}$$

Replace this decomposition of $\hat{\Sigma}$ in the optimization problem, Equation (9.7), and notice that $\|\mathbf{U}\mathbf{w}\| = \|\mathbf{w}\|$ because the matrix \mathbf{U} is orthonormal. We are left to solve

$$\begin{aligned} & \max \mathbf{v}'\mathbf{S}^2\mathbf{v} \\ \text{s.t. } & \|\mathbf{w}\| \leq 1 \\ & \mathbf{w} = \mathbf{U}\mathbf{v} \\ & \mathbf{v} \in \mathbb{R}^n \end{aligned} \tag{9.10}$$

The solution is straightforward: $\mathbf{v} = (1, 0, \dots, 0)'$, and \mathbf{w} equal to the first column of \mathbf{U} . If we were to find the first m principal components, we would find that the columns of \mathbf{U}_m solve our problem. These columns, however, are not uniquely identified when two or more eigenvalues are equal. For example should verify for yourself that, if $\lambda_1 = \lambda_2$, then any vector $\mathbf{v} = (v_1, v_2, 0, \dots, 0)$, with $v_1^2 + v_2^2 = 1$, is indeed a solution. Figure 9.1 gives a geometrical interpretation of this fact.

We call these vectors interchangeably *Principal Components*, *Eigenvectors*, and *Eigenfactors*. The variances of the components are the squared singular values of the SVD.

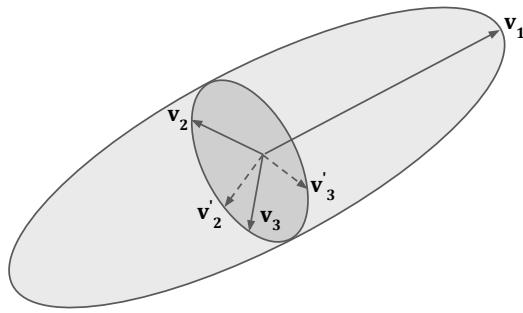


Figure 9.1: The eigenvectors associated to identical eigenvalues are not uniquely identified.

Finally, we note that the optimization problem (9.7) can be extended to the case of m eigenvectors:

$$\begin{aligned} & \max \text{trace} (\mathbf{W}' \hat{\Sigma} \mathbf{W}) \\ \text{s.t. } & \mathbf{W}' \mathbf{W} = \mathbf{I}_m \\ & \mathbf{W} \in \mathbb{R}^{n \times m} \end{aligned} \tag{9.11}$$

9.1.2 Maximum Likelihood Estimation and PCA

The statistical model was introduced as a norm-minimization problem, but is not directly related to a factor model formulation

$$\mathbf{r} = \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon} \tag{9.12}$$

In fact, if we approximated the covariance matrix with a principal component approximation using the top m eigenvalues, we would obtain a singular covariance matrix, which is highly undesirable.

The goal of this section is to establish a first connection between spectral methods and the standard factor model. We consider the model above as a starting point. We assume for simplicity that $\sigma_1, \dots, \sigma_n$, the asset idiosyncratic volatilities, are all equal to σ . Furthermore we assume, without loss of generality, that $\Sigma_f = \mathbf{I}_m$. This is allowed, because rotational invariance affords us this choice of covariance matrix.

Under the assumptions $\mathbf{f} \sim N(0, \mathbf{I}_m)$ and $\boldsymbol{\epsilon} \sim N(0, \Sigma^2 \mathbf{I}_n)$ the return covariance matrix is $\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n$. The first m eigenvalues of the covariance matrix are greater than σ^2 (Exercise 9.5). Let Σ_r be the empirical covariance matrix.

The log-likelihood function for a zero-mean multivariate normal distribution is (Johnson and Wichern, 2007; Bishop, 2006)

$$\mathcal{L}(\hat{\Sigma}_r) = -\frac{T}{2} \left[\log |\hat{\Sigma}_r| + \langle \hat{\Sigma}_r^{-1}, \Sigma_r \rangle + n \log(2\pi) \right] \quad (9.13)$$

where we denote the scalar product of two matrices $\langle \mathbf{A}, \mathbf{B} \rangle := \text{trace}(\mathbf{A}'\mathbf{B})$. The parameters \mathbf{B}, σ can be estimated via maximum likelihood:

$$\max -\log |\hat{\Sigma}_r| - \langle \hat{\Sigma}_r^{-1}, \Sigma_r \rangle \quad (9.14)$$

$$\text{s.t. } \hat{\Sigma}_r = \hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\sigma}^2 \mathbf{I}_n \quad (9.15)$$

The solution to this problem is especially simple and intuitive (Tipping and Bishop, 1999). Decompose $\Sigma_r = \mathbf{U}\mathbf{S}\mathbf{U}'$. Then

$$\begin{aligned} \hat{\mathbf{B}} &= \mathbf{U}_m (\mathbf{S}_m^2 - \hat{\sigma}^2 \mathbf{I}_n)^{1/2} \\ \hat{\sigma}^2 &= \bar{\lambda} \end{aligned} \quad (9.16)$$

where $\bar{\lambda}$ is the average of the last $n - m$ eigenvalues of Σ_r . An alternative rotation of this risk model is:

$$\mathbf{B} = \mathbf{U}_m \quad (9.17)$$

$$\Sigma_f = (\mathbf{S}_m^2 - \bar{\lambda} \mathbf{I}_n) \quad (9.18)$$

$$\Sigma_\epsilon = \bar{\lambda} \mathbf{I}_n \quad (9.19)$$

The model offers several insights. First, it links a probabilistic model of returns to the PCA of the empirical covariance matrix. Second, in the model rotation above, the factor covariance matrix is diagonal and the factor variances are equal to the shrunk empirical variances obtained by PCA. Indeed, the PCA solution can be obtained as an asymptotic result. consider the limit $\hat{\sigma} \downarrow 0$. In this scenario, the idiosyncratic risks are much smaller than the factor risk. In the limit, the formula then simplifies to

$$\mathbf{B} = \mathbf{U}_m \quad (9.20)$$

$$\Sigma_f = \mathbf{S}_m^2 \quad (9.21)$$

$$\Sigma_\epsilon = 0 \quad (9.22)$$

which is the Principal Component Analysis solution.

We show how PPCA works in a simulated instance. We choose $\sigma = 1$, $T = 250$, n equal to 1000 and 3000 assets, $m = 10$ and factor volatilities equal

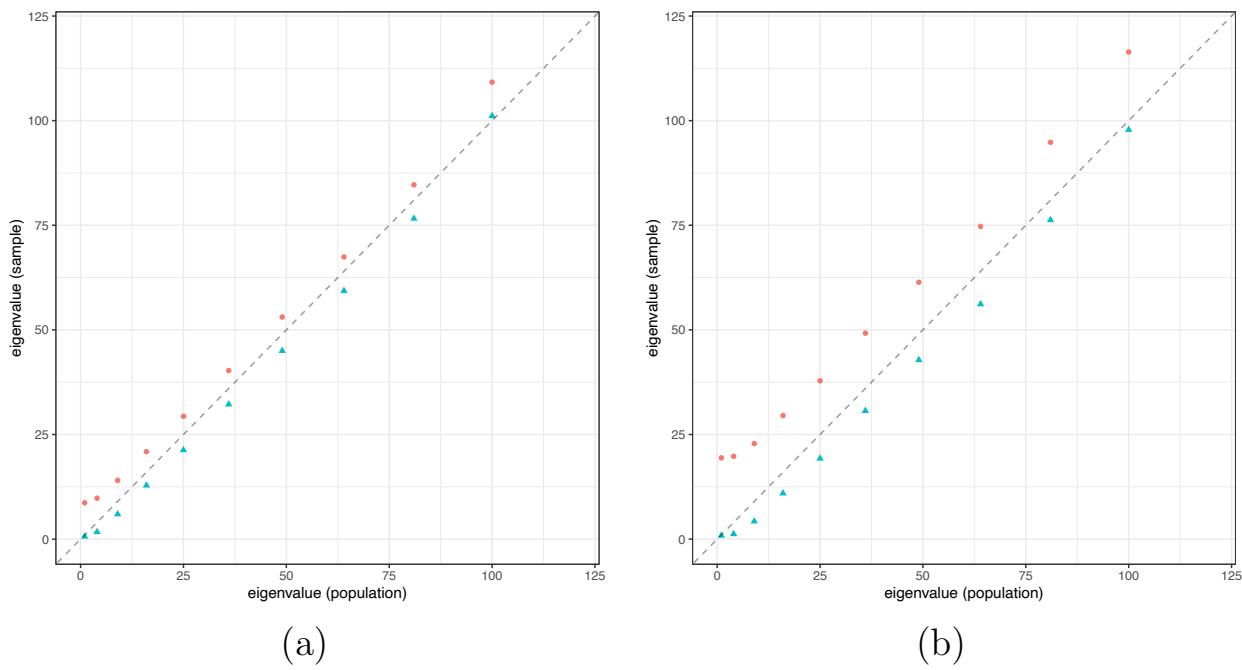


Figure 9.2: (a): Probabilistic PCA for a universe of 1000 assets, with 10 factors with volatilities $1, 2, \dots, 10$. Circle-shaped points are the sample factor variances against the true variances; triangle-shaped are the shrunk factor variances against the true variances. (b): All parameters are unchanged, with the exception of the number of assets, here equal to 3000.

to $1, 2, \dots, 10$. For each set of parameters, we run 50 simulations. Figures 9.2 (a), (b) show the true (population) factor variances on the x axes. On the y-axes we plot the sample factor variances (circles) and the shrunk factor variances (triangles). You can see that, when the ratio between number of assets and number of period is greater, the upward bias of the sample eigenvalues—i.e., of the sample factor variances—is higher. Shrinkage eliminates such bias. However, the shrunk eigenvalues are now biased downwards and, in addition, this downward bias is not constant, which suggests that the optimal shrinkage should not be a constant offset. There are three take-aways from these simulations, which could be confirmed empirically for other choices of the parameters:

- Sample factor eigenvalues are higher than their population counterparts;
- Shrinkage helps, but optimal shrinkage may be more complex than a constant offset;
- Maximum likelihood estimation, which we could solve analytically in this special case, will give in general bias estimates on the factor volatilities.

9.1.3 Cross-Sectional and Time-Series Regressions via SVD

A popular approach to PCA is to take the first m principal components of the PCA as factors loadings, and then estimate the factor returns via cross-sectional regression. What are these factor returns? Start by setting $\mathbf{B} = \mathbf{U}_m$. The estimated factor returns are the result of T cross-sectional regressions. We can write the relation as follows:

$$\mathbf{R} = \mathbf{U}_m \hat{\mathbf{F}} + \mathbf{E} \quad (9.23)$$

The least-squares estimate is

$$\hat{\mathbf{F}} = (\mathbf{U}'_m \mathbf{U}_m)^{-1} \mathbf{U}'_m \mathbf{R} \quad (9.24)$$

or, since \mathbf{U} is orthonormal,

$$\hat{\mathbf{F}} = \mathbf{U}'_m \mathbf{R} = \mathbf{U}'_m \mathbf{U} \mathbf{S} \mathbf{V}' = \mathbf{S}_m \mathbf{V}'_m \quad (9.25)$$

Behold, these are the same factor estimates we computed from the SVD in Equation (9.4). If we throw away the factor returns of an SVD, the loadings of the SVD itself allow us to recover them from cross-sectional regressions. Similarly, you can easily prove (Exercise 9.9) that, if we only know the estimated factor returns $\hat{\mathbf{F}}$ from Equation (9.25), then we can estimate the loadings using time-series regression of asset returns against these factor returns, and obtain as a result $\hat{\mathbf{B}} = \mathbf{U}_m$. Indeed, the SVD decomposition is the *only* factorization of the returns matrix such that the loadings are the time-series betas of the asset returns to the factor returns, and the factor returns are the cross-sectional betas of the asset returns to the loadings³. This is a computational simplification, but also has several applications. It is a useful relationship when we estimate factor loadings for assets with incomplete return data; it helps explain discrepancies in time-series and cross-sectional performance attribution in fundamental factor models; and establishes a connection between statistical and fundamental factor models.

9.2 Beyond the Basics

It is important to understand the behavior of PCA in finite samples, and in settings that are relevant to practitioners. There are a few parameters that

³You can prove this! Solve Exercise 9.8. You will also find additional reading sources in Section 9.6.

intuitively should matter to the portfolio manager. The first two are trivial: the number of assets n and the number of factors m . In addition, we will perform SVD on a rolling window of observations of width T (Figure 9.3). This width is chosen so that the data can be considered broadly homogeneous (i.e., the cross-section of returns are drawn from the same distribution), but also so that the data has a sufficiently high number of observations to estimate the parameters. Finally, another important quantity is the gap between the m th

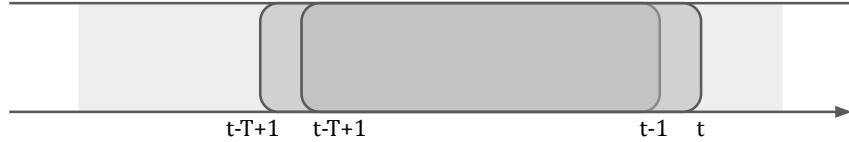


Figure 9.3: We estimate the risk model parameters using data in an interval of width T .

and the $(m + 1)$ the eigenvalues, corresponding to the separation between the smallest variance of a factor and the largest idiosyncratic variance. How do these quantities interact? This question has been at the center of intense research in the past twenty-five years. PCA, a 120-year old technique, has witnessed a theory renaissance, which is still far from being concluded. This chapter attempts to give some intuition about the analytical approach; to summarize the state-of-the-art results; to compare them to simulated scenarios; and finally to administer some practical advice in using PCA.

9.2.1 The Spiked Covariance Model

Let $\lambda_{T,i}$, with $i = 1, \dots, n$, be the sorted eigenvalues of the empirical covariance matrix

$$\tilde{\Omega}_r := T^{-1} \sum_{t=1}^T \mathbf{r}_t \mathbf{r}'_t \quad (9.26)$$

The spiked covariance model posits the following: there is $0 < m < n$ and a positive constant C such that as $T \rightarrow \infty$

$$\lambda_i := \lim_{T \rightarrow \infty} \lambda_{T,i} \begin{cases} = 1 & \text{for all } i > m \\ \geq Cn & \text{for all } i \leq m \end{cases} \quad (9.27)$$

There is a spectral gap between the largest m eigenvalues and the remaining ones. How does this relate to factor models? Consider the original model specified by Equation (9.4) and choose, like we did in Section (9.1.2), and with $\sigma = 1$, a formulation

$$\mathbf{B}\mathbf{B}' + \mathbf{I}_n \quad (9.28)$$

Why should the eigenvalues λ_i grow at least linearly in n ? The first m eigenvalues of $\mathbf{B}\mathbf{B}'$ are the same as those of $\mathbf{B}'\mathbf{B}$. To see this, write the SVD decomposition of $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}'$ and consider the two matrix products $\mathbf{B}\mathbf{B}' = \mathbf{U}\mathbf{S}^2\mathbf{U}'$ and $\mathbf{B}'\mathbf{B} = \mathbf{V}\mathbf{S}^2\mathbf{V}'$. The two products have the same first m eigenvalues and different eigenvectors. Instead of analyzing the properties of $\mathbf{B}\mathbf{B}'$, we will work on $\mathbf{B}'\mathbf{B}$.

A reasonable assumption for \mathbf{B} is that its rows \mathbf{b}_i , representing the loadings of a single stock to the factors, are iid samples from a probability distribution D , so that $\mathbf{b}_i \sim P(\mathbb{R}^m)$. We can then write $\mathbf{B}'\mathbf{B} = \sum_{i=1}^n \mathbf{b}_i' \mathbf{b}_i = n(n^{-1} \sum_{i=1}^n \mathbf{b}_i' \mathbf{b}_i)$. For large values of n the terms in parentheses converges to an expectation $E_D(\mathbf{b}'\mathbf{b})$. We denote μ_i the eigenvalues of this matrix. The eigenvalues of $\mathbf{B}'\mathbf{B}$ are then in the limit $n \rightarrow \infty$ equal to $n\mu_i$, and the eigenvalues of $\mathbf{B}\mathbf{B}' + \mathbf{I}_n$ are $n\mu_i + 1$. This heuristic argument justifies the scaling assumption for the largest eigenvalues eigenvalues: for large stock universes, the pervasive (or *spike*) eigenvalues separate for the rest (or *bulk*), and the gap grows linearly in the size of the stock universe.

Let ν_i be the eigenvalues of $\mathbf{B}\mathbf{B}'$. The spectrum of the covariance matrix is then given by $\nu_1 + 1, \dots, \nu_m + 1, 1, \dots, 1$, so a factor model, after rescaling (so that $\Omega_\epsilon = \mathbf{I}_n$) and rotation (so that $\Omega_f = \mathbf{I}_m$), has an associated spiked covariance matrix. We can see how these condition translate into practice. Recall from Section 4.3 That the i th factor-mimicking portfolio i is $\mathbf{w}_i = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{e}_i$. Consider the risk decomposition:

- The factor variance is $\mathbf{w}_i'(\mathbf{B}\mathbf{B}')\mathbf{w}_i = \mathbf{e}_i' \mathbf{e}_i = 1$.
- The idiosyncratic variance is $\mathbf{w}_i' \mathbf{w}_i = \mathbf{e}_i' (\mathbf{B}'\mathbf{B})^{-1} \mathbf{e}_i = \mathbf{e}_i' \mathbf{V} \mathbf{S}^{-2} \mathbf{V}' \mathbf{e}_i \leq \nu_m^{-1} \|\mathbf{V}' \mathbf{e}_i\|^2 \leq \nu_m^{-1} \|\mathbf{V}'\|^2 \|\mathbf{e}_i\|^2 \leq 1/(Cn)$, since the norm of an orthonormal matrix \mathbf{V} is one.

Therefore for large asset universes, i.e., $n \rightarrow \infty$, factor-mimicking portfolios have a vanishing small percentage idiosyncratic variance. They “mimick” the true factor returns well. A different way to state the approximation property is that the idiosyncratic risk “diversifies away” as the number of assets becomes

larger; and that there are “factor portfolios” with factor risk that is well above their idiosyncratic risk.

9.2.2 Spectral Limit Behavior of the Spiked Covariance Model

The first asymptotic limits for PCA were concerned with large samples: $T \rightarrow \infty$ and n constant. In this case, [Anderson \(1963\)](#) showed that the sample eigenvalues and eigenvectors converge to their population counterparts (see Appendix 14.2). For modern application, the case where both T and n go to infinity is more relevant, with $\gamma := n/T \in [0, \infty)$. This limit is interesting in applications, because the number of observations is often of the same order of magnitude as the number of variables.

Here, \mathbf{r}_t is a sequence of iid rv taking values in \mathbb{R}^n . Assume that:

1. The elements of \mathbf{r}_t has finite fourth moments;
2. There are m constants c_i , with $0 < c_1 < c_2 < \dots < c_m$, such that as $n, T \rightarrow \infty$

$$\frac{\gamma}{\lambda_i} \rightarrow c_i, \quad i = 1, \dots, m \tag{9.29}$$

3. The remaining $n - m$ eigenvalues are equal to one.

Then the following holds ([Shen et al., 2016; Johnstone and Paul, 2018](#)):

1. When $\lambda_i > 1 + \sqrt{\gamma}$:
 - Let $\hat{\lambda}_i$ be the sample eigenvector. Then

$$\hat{\lambda}_i \rightarrow \mu_i := \lambda_i (1 + c_i) \quad \text{a.s.} \tag{9.30}$$

Because of Equation (9.29), in the limit $n, p \rightarrow \infty$, this is the same as

$$\hat{\lambda}_i \rightarrow \lambda_i \left(1 + \frac{\gamma}{\lambda_i}\right), \quad i = 1, \dots, m \tag{9.31}$$

The empirical eigenvalues are asymptotically unbiased for large values of λ ; see Figure 9.4.

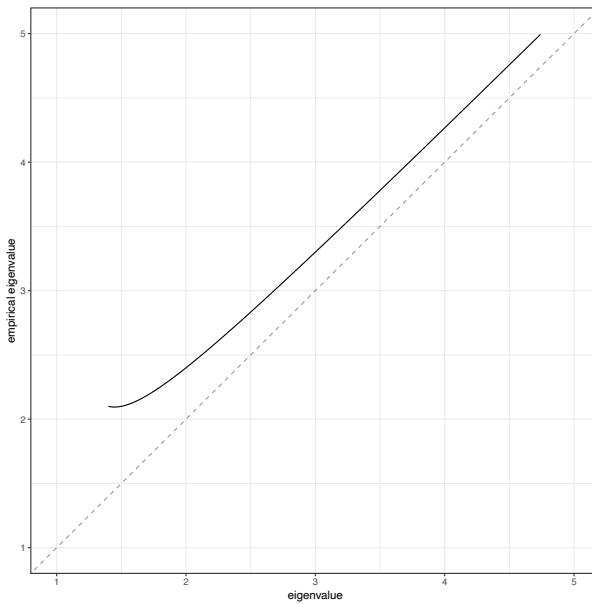


Figure 9.4: Inflation of sample eigenvalues, Equation (9.30), for $\gamma = 0.2$.

- let \mathbf{u}_i the population eigenvector and $\hat{\mathbf{u}}_i$ the sample eigenvector. Then, almost surely,

$$|\langle \mathbf{u}_i, \hat{\mathbf{u}}_i \rangle| \rightarrow \begin{cases} \frac{1}{\sqrt{1+c_i}} & i \leq m \\ O(1/\sqrt{\gamma}) & i > m \end{cases} \quad (9.32)$$

2. When $\lambda_i \leq 1 + \sqrt{\gamma}$:

- $\hat{\lambda}_i \rightarrow (1 + \sqrt{\gamma})^2$ in probability;
- $|\langle u_i, \hat{u}_i \rangle| \rightarrow 0$ a.s.

Even if this strong result only applies to a single spiked model, it offers a few insights that can be verified experimentally. In addition, there are similar results that extend to the multiple spiked eigenvalue case, albeit with more assumptions. First, let us review the insights:

- Under the spiked model assumptions, the spiked empirical eigenvalues are asymptotically upwardly biased. The bias is higher if λ_1 is closer to the ground eigenvalues; it becomes smaller when λ_1 gets bigger. This makes intuitive sense. When λ_1 is close to one, then the probability that the largest empirical eigenvalue is a “noisy” ground eigenvalue becomes non-negligible. This brings us to the second insight.

- There is a critical threshold at $1 + \sqrt{\gamma}$. For eigenvalues larger than $1 + \sqrt{\gamma}$, it is possible to separate the largest eigenvalue from the spectrum. Indeed, the largest sample eigenvalue is further biased upward. The sample eigenvector is collinear with the population eigenvector. The larger the first eigenvalue, the better the eigenvector's collinearity.
- Below the threshold, the largest eigenvalue, even if it is larger than 1, cannot be easily be identified from data. The associated eigenvector contains no information about the population eigenvector.

In practice, for many applications, the number of asset in a model is in the interval $(10^3, 10^4)$, and the number of observations ranges between 250 and 1,000, so that $1 + \sqrt{\gamma}$ ranges between 2 and 7. This is a useful starting point to reason about thresholding eigenvalues, and their associated eigenvector.

9.2.3 Optimal Shrinkage of Eigenvalues

We know that the empirical eigenvalues are biased. This means that, should we evaluate portfolios in the subspace spanned⁴ by the spike eigenvectors, the predicted volatility of the portfolios will be biased upward by γ . Let $\mathbf{a} \in \mathbb{R}^m$ be a unit-norm vector, and let the portfolio be⁵ $\mathbf{w} = \mathbf{U}_m \mathbf{a}$. Then $\hat{\sigma}_{\mathbf{w}}^2 = \mathbf{a}' \hat{\Lambda} \mathbf{a} = \sum_{i=1}^m \hat{\lambda}_i a_i^2 = \sum_{i=1}^m \lambda_i a_i^2 + \gamma = \sigma_{\mathbf{w}}^2 + \gamma$. If the portfolio is in the subspace orthogonal to the eigenfactors, then the portfolio variance estimate is also biased. We can repeat the calculations and use the BBP theorem to obtain $\hat{\sigma}_{\mathbf{w}}^2 = \sigma_{\mathbf{w}}^2 + 2\sqrt{\gamma} + \gamma$. A possible solution to the problem of eigenvalue estimation error is to apply a function to the sample eigenvalues. For example, from Equation (9.30), one could invert λ_i from $\hat{\lambda}_i$ by applying the function

$$\ell(\lambda) = \lambda - \gamma, \quad \lambda \geq 1 + \sqrt{\gamma} \tag{9.33}$$

For large values of λ , this shrinkage function is an offset of the empirical eigenvalues, like the one we first saw in PPCA, Equation (9.18). When we apply this to a diagonal matrix $VECS$ filled with eigenvalues, we use the notation $\ell(\mathbf{S})$, which returns a diagonal matrix with the corresponding diagonal terms shrunk using Equation (9.33). However, this is not necessarily the best choice.

⁴The subspace spanned by vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ is the set of vectors that can be expressed as a linear combination of \mathbf{v}_i . The column subspace of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the subspace spanned by its column vectors, i.e., the set $\{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^m\}$.

⁵As in Subsection 9.1.1, \mathbf{U}_m is the submatrix of \mathbf{U} obtained by taking the first m columns.

The choice of the loss function matters. [Donoho et al. \(2018\)](#) characterize the optimal *shrinking* of eigenvalues for a large number of loss types. Based on what we learned in Chapter 8, we focus only on a few: the operator norm $\|\mathbf{A} - \mathbf{B}\|$ and the operator norm on precision matrix $\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|$. For these two losses, the shrinkage formula Eq. (9.33) is optimal. For large values of λ , this formula simplifies to $\ell(\lambda) \simeq \lambda + 1 - \gamma$. We subtract a constant offset from each eigenvalue. Large eigenvalues are shrunk proportionally less than the small ones. This result is connected to what is perhaps the best-known covariance shrinkage method among practitioners: the Ledoit-Wolf shrinkage. This method starts with finding a matrix of the form $\hat{\Sigma}_r$ of the form

$$\hat{\Sigma}_r = \rho_1 \tilde{\Sigma}_r + \rho_2 \mathbf{I}_n \quad (9.34)$$

and they identify ρ_1 and ρ_2 so that $\hat{\Sigma}_r$ minimizes the distance induced by the Frobenius norm from Σ_r :

$$\begin{aligned} & \min \left\| \hat{\Sigma}_r - \Sigma_r \right\|_F \\ & \text{s.t. } \hat{\Sigma}_r = \rho_1 \tilde{\Sigma}_r + \rho_2 \mathbf{I}_n \end{aligned} \quad (9.35)$$

The space of $n \times n$ matrices is a Hilbert space with scalar product $\langle \mathbf{A}, \mathbf{B} \rangle := \text{trace}(\mathbf{AB}')$. The induced norm $\sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$ is the Frobenius norm. This is then just a special case of the well-known problem of minimum distance of a subspace from a point in a Hilbert space ([Luenberger \(1969\)](#), Sec 3.3). They assume iid returns, finite fourth moments, and an asymptotic regime in which n is constant and $T \rightarrow \infty$. They find that the optimal solution is of the form

$$\hat{\Sigma}_r = \left(1 - \frac{\kappa}{T}\right) \tilde{\Sigma}_r + \frac{\kappa}{T} \mathbf{I}_n \quad (9.36)$$

This solution has many interpretations, aside from the geometric one that follows from the solution to Problem (9.35). While these interpretations may be of independent interest, I will devote some time to justify why this approach is *not* recommended to estimate returns with a spiked covariance. A first issue is using the Frobenius-induced distance is generally not helping to identify the structure of the model, as shown in the previous chapter. Secondly, because the regime n fixed, T diverging is not relevant to applications in which $n > T$ or $n \asymp T$. Thirdly, because the condition that the estimate lie in the subspace spanned by $\tilde{\Sigma}_r$ and I_n may be overly restrictive. Lastly, because the eigenvalues of the target matrix are of the form $\lambda_i - \kappa T^{-1}(\lambda_i - 1)$. For the leading eigenvalues, this shrinkage does not match the optimal asymptotic shrinkage of the spiked covariance model.

9.2.4 Eigenvalues: Experiments Vs. Theory

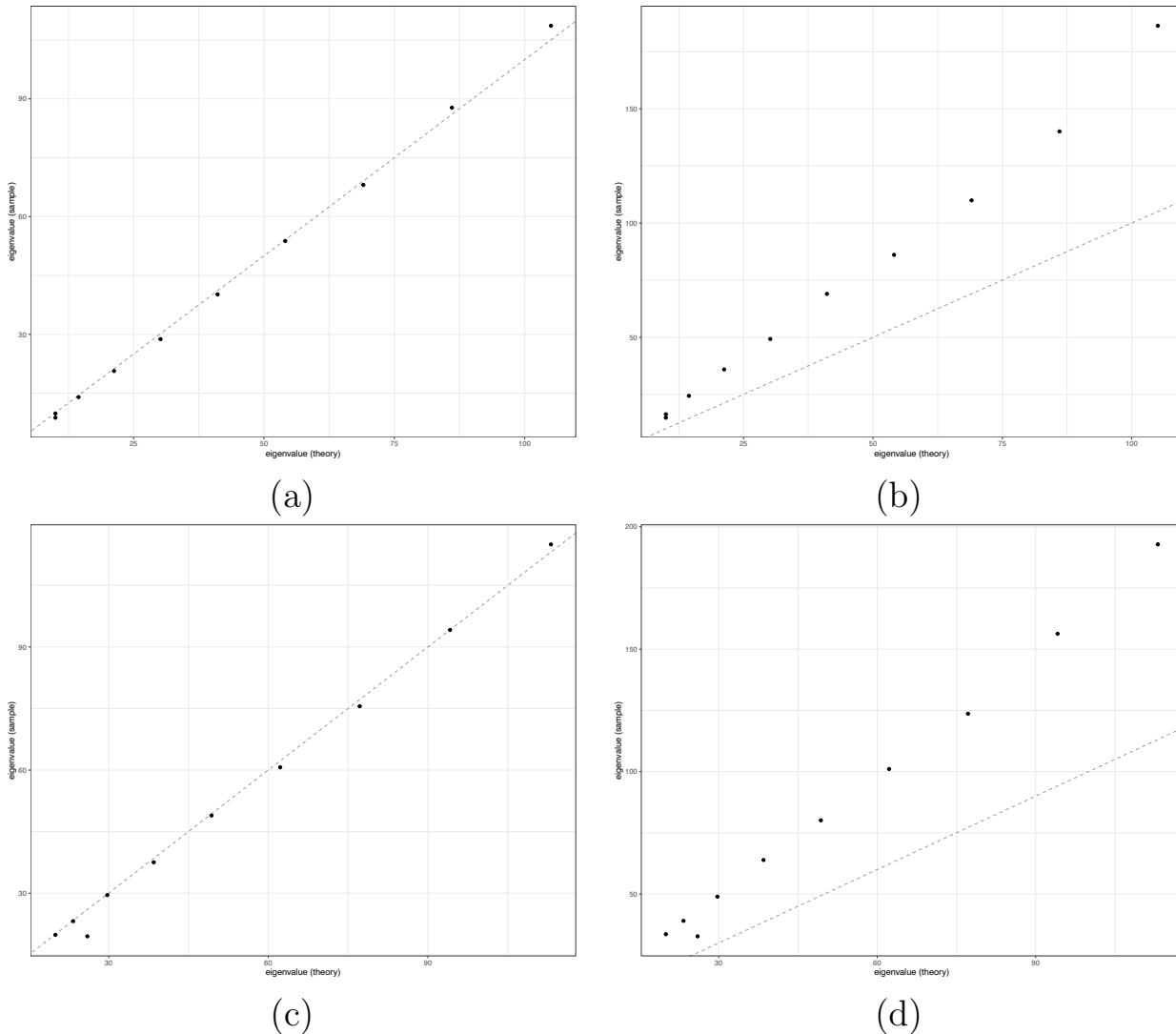


Figure 9.5: (a): 1000 assets, normally distributed returns; (b) 1000 assets, t-distributed returns; (c): 3000 assets, normally distributed returns; (d) 3000 assets, t-distributed returns. The x -axis denotes the population eigenvalues, while the y -axis denotes the shrunk empirical eigenvalues. The dashed line is the line $y = x$.

We now compare these theoretical results to simulations. We use the same parameters we used for the Probabilistic PCA in Section 9.1.2: 10 factors with standard deviations ranging between 1 and 10, uniformly spaced; unit idiosyncratic standard deviations; 250 periods, and either 1000 or 3000 assets. In addition to the case of normal returns, I also consider the case of heavy-tailed returns. Specifically both factor returns and idiosyncratic returns are t-distributed with five degrees of freedom. This choice is meant to simulate

returns that have four finite moments, which is a reasonable assumption for daily asset returns.

We simulate 50 instances of each factor model. For each model, we compute the first ten empirical top eigenvalues, and we shrink them using Formula (9.33) for $\ell(\hat{\lambda})$. The simulation shows that the shrinkage function ℓ works well for normally distributed returns, but not for heavy-tailed returns. In this case, it appears that a better shrinkage approach is to scale the eigenvalues by a common factor. This is a different shrinkage than the one of Equations (9.18) and (9.33), which consistent in a constant offsetting term. Combining the empirical observations from simulated data, and theoretical results, it seems at least reasonable to consider a linear shrinkage

$$\ell(\lambda) = \kappa_1 \lambda - \kappa_2 \quad (9.37)$$

$$\kappa_2 \geq \lambda_{\min} \quad (9.38)$$

$$\kappa_1 \in (0, 1) \quad (9.39)$$

when identifying a model.

9.2.5 Choosing the Number of Factors

In the example above, we assumed that the number of factors was known in advance. This is not the case in applications. An important component of the model definition procedure is the determination of the number of factors. There are some criteria motivated by theoretical models, and others that are the outcome of experiments and trial-and-error by generations of practitioners. The theory-based models themselves prescribe different numbers of factors, so we should premise this section with a few considerations. First, finding the *right* factors matters more than finding their exact right number. By “right”, I mean of course the factor loadings with the best “performance”, and by performance, I mean one of many metrics introduced in chapter 8. Because there are many metrics, many of which not even considered in the theoretical treatments on the number of factors, there is no one-size-fits-all criterion. Second consideration: telling the exact number of factor in practice is either very easy or hopelessly hard. Under the assumptions of pervasive factors, you won’t need complex criteria: there is a wide gap between the smallest factor eigenvalue and the largest idiosyncratic one. When the assumption does not hold, eigenvalues will decrease gradually, and a hard rule is unlikely to choose the exact threshold. A final consideration, which is both grounded in theory and in practice is that, one should err on the side of selecting more factors, rather then fewer.

The cost of selecting too few factors is that, in portfolio optimization, we will choose portfolios that underestimate their true risk, which can result in steep degradation of the Sharpe Ratio. This is covered in depth in Chapter 5. The cost of choosing too many factors is a slight decrease in the Sharpe Ratio.

After these qualifications, let us review the most common methods.

- **Threshold-based methods.** For matrices with ground eigenvalues equal to 1, the results of Section 9.2.2 suggest that we should select as factor eigenvalues those that exceed the threshold $1 + \sqrt{\gamma}$, i.e.

$$m = \max\{k | \hat{\lambda}_k \geq 1 + \sqrt{\gamma}\} \quad (9.40)$$

An older method is the *scree plot*. This is the best-known method. It consists of plotting the eigenvalues against their rank. The largest eigenvalues dominate and decrease rapidly, to a value where the eigenvalue are small and decrease gradually, usually almost linearly. The method consists of choosing the last eigenvalue preceding this group. A variant of this method plots the logarithm of the eigenvalues.

- **Maximum Change Points.** Associated to these two methods, are two additional ones that select the number of factors based on the largest gap between consecutive factor eigenvalues, or consecutive $\log(\text{eigenvalues})$:

$$\begin{aligned} m &= \max_{2 \leq k \leq k_{\max}} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \\ m &= \max_{2 \leq k \leq k_{\max}} (\log \hat{\lambda}_k - \log \hat{\lambda}_{k-1}) \end{aligned} \quad (9.41)$$

Where k_{\max} is a threshold chosen iteratively (Onatski, 2010).

- **Penalty-Based Methods.** We began the chapter with the problem of minimizing the square residual error, Equation (9.2). We can select the number of factors by adding a penalty term, and by making m a decision variable

$$\min_{k, \text{rank}(\hat{\mathbf{R}}) \leq k} \|\mathbf{R} - \hat{\mathbf{R}}\|^2 + kf(n, T) \quad (9.42)$$

$$f(n, T) = \frac{n+T}{nT} \log \left(\frac{nT}{n+T} \right) \quad (9.43)$$

9.3 Real-Life Stylized Behavior of PCA

We now explore a real-life data set with the goal of comparing the observed behavior of principal components and eigenvalues to the ideal spiked covariance model. We employ daily stock total returns belonging to the Russell 3000 index for the period 2007-2017. Assets that are included in this index must satisfy some essential requirements. As of 2022, on a designated day in May (“rank day”), Russell evaluates eligibility for inclusion in its indices based on several criteria. Among them, the company must be U.S.-based (no ADR/GDRs allowed⁶); the stock price must exceed \$1; the market capitalization must exceed \$30mm; and the percentage of float (shares traded on exchange) must exceed 5% of the total shares issued. In addition, some governance requirements and corporate structure must be met; for example ETFs, trusts, closed-end funds investment companies and REITs are excluded. Out of this eligible set, Russell assigns to R3000 the first 3,000 assets by market cap, and effectively changes the composition of the index on the fourth Friday of June. These criteria ensure that the asset characteristics are sufficiently homogeneous (based on geography, revenue source and corporate governance) and that the returns can be reliably computed based on daily closing prices (based on stock price and market capitalization⁷).

9.3.1 Concentration of Eigenvalues

For our exploration we consider Principal Components based on three types of returns. First, stock total returns. This is the simplest approach. Secondly, we normalize returns by dividing them by their predicted idiosyncratic volatilities. The benefit of this approach is that it should make the spectrum closer to the assumptions we made in the previous sections: the idiosyncratic volatilities of the normalized empirical covariance matrix are all equal to one, and the spike volatilities should be greater than one. Lastly, we normalize returns by their predicted total volatilities. The rationale for this choice is that we study the properties of the empirical *correlation* matrix. It is at least reasonable to hypothesize that the correlation matrix has different properties than the

⁶An American Depository Receipt (ADR) is a foreign company that is listed on a foreign stock exchange, which also offers shares in U.S. exchanges. A Global Depository Receipt (GDR) is similar to an ADR, but is offered on exchanges in more than one country outside of the primary market.

⁷Note, however, that Russell does not screen stocks based on trading volume, and that the smaller-capitalization companies in R3000 and R2000 may not be sufficiently liquid to be traded in large sizes.

covariance matrix. Correlations may be more stable than covariances; for example, this is the modeling assumption made in [Bollerslev \(1990\)](#), and in Barra's and Axioma's U.S. statistical models. Our procedure is relatively simple. We use one full year of return data, for eight non-overlapping years. When we normalize by idiosyncratic volatilities, we use the data provided by Axioma's US model AXUS Short Horizon. I take this shortcut for one simple reason: I introduce a self-contained idio vol estimation process later in the chapter, but did not wait any further to show some empirical data. I will show later that the statistical model idio vols are indeed quite close to those of a commercial model, so that this illustrative example is in fact quite close to a self-contained analysis. The raw returns are winsorized at daily returns of $(-90\%, +100\%)$, and the z-scored returns at returns of $(-10, +10)$, i.e., plus or minus ten standard deviations. Figure 9.6 shows the variances of the first forty factors, normalized by the variance of the first factor (to make them comparable). Two features are conspicuous. The first one is that there is no obvious gap between variances. The second is that there is a consistent ranking between the spectra of the three covariance matrices. The plot shows the ratio of the variance of lower-order factors to that of the first factor, and this value is smallest for the return/idio vol covariance matrix, followed by the return/total vol covariance matrix, and lastly by the covariance matrix of total returns. This suggests that the first

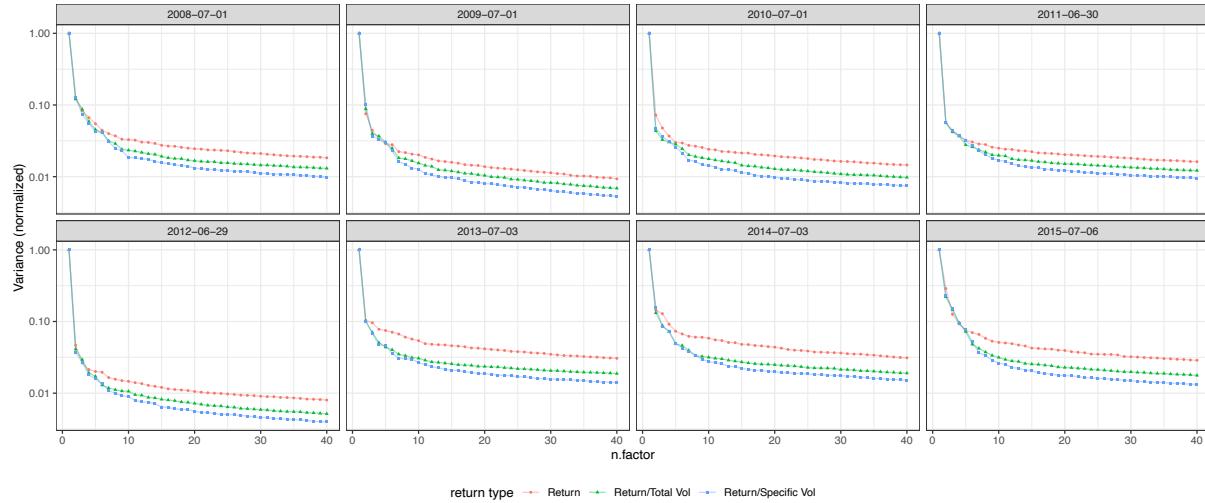


Figure 9.6: Variances of the eigenfactors (normalized to the variance of the first eigenfactor) for the first forty factors. Note that the scale of the y axis is logarithmic.

few eigenfactors explain a larger percentage of total variance of the associated covariance matrix. This is confirmed by Figure 9.7. For example, say that we

would like to have a number of factors sufficient to capture 50% of the variance of asset returns. For the period ending on July 1, 2008, we need 30 factors for the raw covariance matrix, 20 factors for the z-scored returns, using total volatility, and only 10 factors for the z-scored returns, using idiosyncratic volatility. This in itself does *not* mean that this choice is preferable, because the performance of a risk model has no direct relationship with this metric. Nonetheless, it suggests that, in this specific instance, a model built on a transformed sequence of returns is more parsimonious.

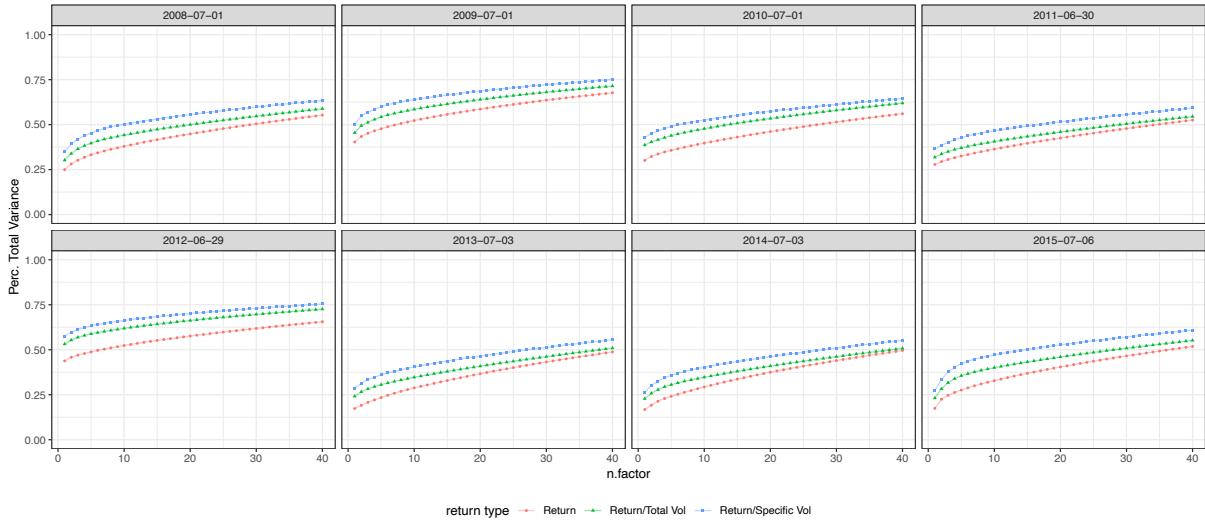


Figure 9.7: Cumulative percentage of variance described by the first n factors, for difference covariance matrices.

9.3.2 Turnover of Eigenvectors

So far, we have focused on the properties of the eigenvalues. Eigenfactors exhibit a distinctive behaviour as well. One important property of eigenfactors is their turnover. The turnover for two consecutive portfolios $\mathbf{v}(t), \mathbf{v}(t+1)$ is usually measured as the gross market value traded, as a percentage of the gross market value of the portfolio: $\text{turnover}_1(\mathbf{v}(t)) := (\sum_i |v_i(t) - v_i(t-1)|) / \sum_j |v_j(t-1)|$. An alternative is to use the definition uses the square of the gross notional:

$$\text{turnover}_2(\mathbf{v}(t)) := \frac{\sum_i (v_i(t) - v_i(t-1))^2}{\sum_j v_j^2(t-1)} \quad (9.44)$$

There are good reasons for this. The first one is that the squared GMV is a fairly good approximation to the transaction costs associated to trading the factor portfolio. A second one is analytical tractability and a associated

geometric intuition. For eigenportfolios, recall that $\|\mathbf{v}(t)\| = 1$, and that the numerator $\|\mathbf{v}(t) - \mathbf{v}(t-1)\|^2$ can be rewritten as $2(1 - \mathbf{v}'(t)\mathbf{v}(t-1))$. The quadratic turnover is therefore related to the cosine similarity which we defined earlier in this chapter. Low-turnover eigenportfolios have high cosine similarity.

$$\text{turnover}_2(\mathbf{v}(t)) = 2[1 - |S_C(\mathbf{v}(t), \mathbf{v}(t-1))|] \quad (9.45)$$

In the equation above we use the absolute value of S_C because the eigenfactors are identified modulo the sign of the vector. In other terms, if $S_C(\mathbf{v}(t), \mathbf{v}(t-1)) < 0$, we can always flip the sign of $\mathbf{v}(t)$ in order to have a lower-turnover pair of eigenfactors. In Figure 9.9 we show the absolute values of the cosine distances over time for the first eight eigenfactors of our three sequences of covariance matrices computed on raw total returns (panel (a)), raw returns normalized by total volatilities (panel (b)), and raw returns normalized by idio volatilities (panel (c)). The covariance matrix on a given date is computed using the trailing 252 trading days of returns. The number of assets from one day to the next can change slightly as well, because the universe is not fixed. The charts have qualitatively similar behavior. The first eigenfactor, is associated to an eigenvalue that has a large gap from the second largest eigenvalue (see Figure 9.6). As a result, the PCA procedure has no issue in identifying it and its weights are very stable throughout the estimation period. This is essentially a “market” portfolio. The turnover has a more interesting structure for higher-order eigenportfolio. Consider the second eigenportfolio of the (non-normalized) total returns. There are occasional spikes; for example there are large spikes occurring on October 9, 2009 and November 20, 2009. The second one is so big that the eigenfactors on consecutive dates have a turnover of almost 200%. What is even more puzzling is that immediately before and immediately after the portfolio doesn’t turn over at all: it changes dramatically on a single day, to stabilize shortly afterwards. And this behavior qualitatively repeats across covariance matrices and eigenfactors: higher-order eigenfactors transition more often and with larger spikes, but transitions are still relatively rare: Even eigenfactor 8 has a cosine similarity below 1/2 only in 6% of the cases. Another qualitative phenomenon is that, as for the case of eigenvalues, standardizing returns seems to reduce turnover incidence and severity; more so for idio vol normalization. For example, in the latter case, eigenfactor 8 has a cosine similarity below 1/2 only in 1.5% of the cases. How to explain this phenomenon? The cause of the jumps is a direct consequence of the lack of eigenvalue separation. When eigenvalues are close, the addition and removal of an observation of cross-sectional returns, as well as the addition or removal of one or two assets in the estimation universe,

is sufficient to affect the numerical solution of the PCA. The distance between the eigenvalues (i.e., variances of the eigenfactors) is within the change of these same eigenvalues from one period to the next due to data updates. Even if the eigenfactors change, the subspace spanned by these eigenfactors may be in fact stable. Below, we show the subspace distance for the three cases above⁸ between the column subspaces in consecutive periods. The distances are very small (the largest being just 1E-5, for total return factors), and are smaller for idio vol z-scored returns. This does confirm yet again that statistical models built on normalized returns sequence are more stable, suggesting that the eigenvalues of such models are better separated from each other.

Aside from the quality of the PCA for different choices of covariance matrices, we are faced with an inescapable issue in statistical models. Except for a few high-order factors and variances, most factors in statistical models suffer from a kind of indeterminacy. In consecutive periods, PCA may give us very different loadings, even though the subspaces spanned by these factors are very close to each other. Is this a matter of concern? For most applications, it is not. The reason is that, even if loadings can change a lot from one period to the next, the covariance matrix does not change⁹. This means that a portfolio's volatility prediction does not depend on the orientation of the factor loadings, and therefore that any portfolio optimization problem is also not affected by the choice of loadings, so long as its formulation includes constraints or objective-function penalty terms on the portfolio volatility, or combined factor volatility of the degenerate factors (i.e., factors with identical volatilities). In integrated fundamental/statistical models, a topic we will cover in a later chapter, the indeterminacy of loadings is not affecting the final result, namely the volatility predictions and the performance characteristics of the fundamental factors. In Table 9.1 I summarize the relevance to specific applications.

Use	Impact of High Factor Turnover
Volatility Estimation	Not important
Portfolio Optimization/Hedging	Not important
Integrated Statistical/Fundamental Models	Not Important
Performance Attribution	Very High

Table 9.1: Summary of Impact of High Factor Turnover.

⁸For a definition of subspace similarity, see Exercise 9.11.

⁹If you are not convinced, or this statement does not seem obvious, this is a good time to solve Exercise 9.13.

Essentially only single-factor Performance Attribution is made irrelevant by eigenvalue quasi-degeneracy. However, single-factor attribution depends on factor turnover. Since model prediction is unaffected by rotations, we can always perform a rotation that minimize distance between loadings in consecutive periods; this is a zero-cost operation. In other words, if we have a sequence of loading matrices \mathbf{B}_t , we aim for new “rotated” loadings $\tilde{\mathbf{B}}_t$ that have low turnover¹⁰:

$$\mathbf{B}_0 := \mathbf{B}_0 \quad (9.46)$$

$$\begin{aligned} \tilde{\mathbf{B}}_{t+1} &= \arg \min \|\mathbf{B}_t - \mathbf{Y}\|_F^2 \\ \text{s.t. } \mathbf{Y} &= \mathbf{B}_{t+1}\mathbf{X} \\ \mathbf{X}'\mathbf{X} &= \mathbf{I}_m \\ \mathbf{X} &\in \mathbb{R}^{m \times m} \end{aligned} \quad (9.47)$$

First, we prove that the objective is equivalent to maximizing $\langle \mathbf{A}, \mathbf{X}' \rangle$, with $\mathbf{A} := \mathbf{B}'_t \mathbf{B}_{t+1}$. This follows from the sequence of identities

$$\begin{aligned} \|\mathbf{B}_t - \mathbf{B}_{t+1}\mathbf{X}\|_F^2 &= \|\mathbf{B}'_t - (\mathbf{B}_{t+1}\mathbf{X})'\|_F^2 \\ &= \text{trace}((\mathbf{B}'_t - (\mathbf{B}_{t+1}\mathbf{X})')(\mathbf{B}_t - \mathbf{B}_{t+1}\mathbf{X})) \\ &= \text{trace}(\mathbf{B}'_t \mathbf{B}_t) + \text{trace}((\mathbf{B}_{t+1}\mathbf{X})'(\mathbf{B}_{t+1}\mathbf{X})) \\ &\quad - \text{trace}(\mathbf{B}'_t \mathbf{B}_{t+1}\mathbf{X}) - \text{trace}(\mathbf{X}' \mathbf{B}'_{t+1} \mathbf{B}_t) \\ &= 2(m - \text{trace}(\mathbf{AX})) \end{aligned}$$

The last equality follows from the orthonormality of $\mathbf{B}_t, \mathbf{B}_{t+1}$. Let the SVD of \mathbf{A} be $\mathbf{A} = \mathbf{USV}'$. We prove that a solution is given by $\mathbf{X}^\star = \mathbf{VU}'$. let $\mathbf{A} = \mathbf{USV}'$ and $\mathbf{X} = \mathbf{VYU}'$ for some \mathbf{Y} . From orthonormality of \mathbf{X} follows directly $\mathbf{Y}'\mathbf{Y} = \mathbf{I}$. We replace these expressions in the objective function: $\max \text{trace}(\mathbf{AX}) = \max \text{trace}(\mathbf{SY})$. Now for the last step: unitary matrices have all eigenvalues equal to ones and orthogonal eigenvectors \mathbf{a}_i . The eigendecomposition of \mathbf{Y} is $\mathbf{Y} = \sum_i \mathbf{a}_i \mathbf{a}'_i$ and the objective function is $\text{trace}(\mathbf{SY}) = \sum_i s_i [\mathbf{a}_i \mathbf{a}'_i]_{i,i}$, but this is maximized when $\mathbf{a}_i = \mathbf{e}_i$, and $\mathbf{Y} = \mathbf{I}$, so that the solution is $\mathbf{X} = \mathbf{VU}'$.

$$\begin{aligned} \|\mathbf{B}_t - \mathbf{B}_{t+1}\mathbf{X}\|_F^2 &= \|\mathbf{B}'_t - (\mathbf{B}_{t+1}\mathbf{X})'\|_F^2 \\ &= \text{trace}((\mathbf{B}'_t - (\mathbf{B}_{t+1}\mathbf{X})')(\mathbf{B}_t - \mathbf{B}_{t+1}\mathbf{X})) \\ &= \text{trace}(\mathbf{B}'_t \mathbf{B}_t) + \text{trace}((\mathbf{B}_{t+1}\mathbf{X})'(\mathbf{B}_{t+1}\mathbf{X})) \\ &\quad - \text{trace}(\mathbf{B}'_t \mathbf{B}_{t+1}\mathbf{X}) - \text{trace}(\mathbf{X}' \mathbf{B}'_{t+1} \mathbf{B}_t) \\ &= 2(m - \text{trace}(\mathbf{AX})) \end{aligned}$$

¹⁰Historical note: this problem is closely related to *Wahba’s Problem* (Wahba, 1965).

The last equality follows from the orthonormality of $\mathbf{B}_t, \mathbf{B}_{t+1}$. Now, let $\mathbf{A} = \mathbf{USV}'$ and $\mathbf{X} = \mathbf{VYU}'$ for some \mathbf{Y} . From orthonormality of \mathbf{X} follows directly $\mathbf{Y}'\mathbf{Y} = \mathbf{I}$. We replace these expressions in the objective function: $\max \text{trace}(\mathbf{AX}) = \max \text{trace}(\mathbf{SY})$. Now for the last step: unitary matrices have all eigenvalues equal to ones and orthogonal eigenvectors \mathbf{a}_i . The eigendecomposition of \mathbf{Y} is $\mathbf{Y} = \sum_i \mathbf{a}_i \mathbf{a}_i'$ and the objective function is $\text{trace}(\mathbf{SY}) = \sum_i s_i [\mathbf{a}_i \mathbf{a}_i']_{i,i}$, but this is maximized when $\mathbf{a}_i = \mathbf{e}_i$, and $\mathbf{Y} = \mathbf{I}$, so that the solution is $\mathbf{X} = \mathbf{VU}'$.

9.4 Interpreting Principal Components

One criticism that is often leveled against Principal Component Analysis is that its loadings are hard to interpret. The goal of this chapter is to partially dispel this myth. The output of a PCA can be interpreted, and in fact sometimes it provides additional non-trivial perspectives for the user.

9.4.1 The Clustering View

The first avenue to interpretation is to do no transformation at all. The principal components are uniquely determined up to a change of sign: if \mathbf{u} is an eigenvector associated to eigenvalue λ , so is $-\mathbf{u}$. We show that their loadings can be interpreted as a *clustering membership index* (Ding and He, 2004). In order to make the connection between clustering and PCA, we first introduce the K -means approach. We partition our n assets into K clusters, each characterized by a set membership C_k and centroids $\mathbf{m}_k := \sum_{i \in C_k} \mathbf{r}^i / |C_k|$. The number of cluster is set in advance. The cluster membership is found by minimizing the sum of squared distances from the centroids:

$$\min \sum_{k=1}^K \sum_{i \in C_k} (\mathbf{r}^i - \mathbf{m}_k)^2 \quad (9.48)$$

$$\text{s.t. } \mathbf{m}_k := \sum_{i \in C_k} \mathbf{r}^i / |C_k| \quad (9.49)$$

$$C_i \cap C_j = \emptyset, i \neq j \quad (9.50)$$

$$\bigcup_i C_i = \{1, \dots, N\} \quad (9.51)$$

The objective function can be rewritten as

$$\sum_i \mathbf{r}_i - \sum_{k=1}^K |C_k|^{-1} \sum_{j, \ell \in C_k} (\mathbf{r}^j)' \mathbf{r}^\ell \quad (9.52)$$

The first sum is a constant and does not affect the optimization problem. We could represent cluster membership algebraically. Let $\mathbf{h}_k \in \mathbb{R}^n$ and define $[\mathbf{h}_k]_i = 1/\sqrt{|C_k|}$ if asset i is in cluster C_k , zero otherwise. Because an asset needs to belong to exactly one cluster, there is a constraint on the vectors: $\sum_k \sqrt{|C_k|} \mathbf{h}_k = \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones. Define $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_K) \in \mathbb{R}^{n \times K}$. Let $\mathbf{g} = (\sqrt{|C_1|}, \dots, \sqrt{|C_K|})$. The condition that each asset belongs

to precisely one cluster can be expressed as $\mathbf{H}\mathbf{g} = \mathbf{1}$. Therefore to solve a K -clustering problem, we need to solve

$$\max \text{trace}(\mathbf{H}'\mathbf{R}'\mathbf{R}'\mathbf{H}) \quad (9.53)$$

$$\text{s.t. } [\mathbf{H}]_{i,k} \in \left\{0, |C_k|^{-1/2}\right\} \quad (9.54)$$

Notice that the columns of \mathbf{H} have unit norm and are orthogonal. Then it is natural to relax the discrete requirements on \mathbf{H} and to solve

$$\max \text{trace}(\mathbf{H}'\mathbf{R}\mathbf{R}'\mathbf{H}) \quad (9.55)$$

$$\mathbf{H}'\mathbf{H} = \mathbf{I}_K \quad (9.56)$$

This is the same formulation as the optimization version of the uncentered PCA (9.11). The interpretation of the loadings then can be one of approximate cluster membership. The simplest case is when we cluster on the first principal component. We can separate the two clusters based on some clustering method on the loadings; oftentimes, a simple inspection of the loadings distribution will suggest an appropriate cut-off point. When inspecting multiple eigenvectors, a multivariate clustering algorithm will help identify groups.

9.4.2 The Regression View

Another way to interpret the loadings of a statistical model is to represent them as sums of vectors, whose weights are intuitive. Qualitatively, we proceed as follows. First, assemble meaningful stock characteristics for a given date. We denote the matrix of characteristics $\mathbf{G} \in \mathbb{R}^{n \times p}$, where each characteristic is a column \mathbf{g}^j of the matrix \mathbf{G} . We denote \mathbf{B} the matrix of loadings from the statistical model. Now, regress \mathbf{B}^i on the columns of \mathbf{G} , and denote the regression coefficients $\boldsymbol{\beta}^i \in \mathbb{R}^p$. In formulas, $\mathbf{B}^i = \mathbf{G}\boldsymbol{\beta}^i + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^n$ is a vector orthogonal to the column subspace of \mathbf{G} . If we are not using a very wide set of characteristics, then the regression weights help interpret the statistical loadings. The approach is, of course, not restricted to statistical models: we could apply this regression approach to any pair of risk models, to interpret one based on information contained in the other.

As a (very simplified) example, we consider a model built on US asset returns normalized by idio vols, for the date of July 6, 2017. In order to gain intuition about the eigenfactors, we regress them against style loadings only; we use Axioma AXUS4 as source of these loadings. In Tables 9.2 and 9.3, we report only the most significant loadings.

term	estimate	std.error	t-statistic	p value
Market Intercept	1.7E-02	1.6E-04	1.1E+02	0.0E+00
Volatility	-2.6E-03	1.9E-04	-1.4E+01	8.1E-43
Short-Term Momentum	1.2E-03	1.6E-04	7.6E+00	2.8E-14
Earnings Yield	7.7E-04	1.9E-04	4.1E+00	3.4E-05

Table 9.2: Regression coefficients for the first principal component.

term	estimate	std.error	t-statistic	p value
Dividend Yield	4.2E-03	3.3E-04	1.3E+01	8.7E-36
Short-Term Momentum	-3.4E-03	3.3E-04	-1.0E+01	1.1E-24
Size	3.3E-03	3.3E-04	1.0E+01	1.2E-23

Table 9.3: Regression coefficients for the second principal component.

The first principal component is overwhelmingly explained by the market factor, i.e., the factor of identical loadings all equal to ones. This is usually the case in statistical models. Regarding the second factor, the most important explanatory variables are a value factor (Dividend Yield), Size, and (with negative coefficient) Short-Term Momentum. The opposite signs for value and momentum are consistent with experience, since the returns of these factors are usually negatively correlated. Size and Dividend Yield loadings are usually *positively* correlated, the reason being that large caps are likely to pay higher dividends—or dividends at all—than large caps. For this specific date, the correlation is 0.32. The first factor can be interpreted as a “risk-on” factor, whereas the second factor can be interpreted as a defensive, or “risk-off” factor.

9.5 Statistical Model Estimation in Practice

So far we have only presented the theory of statistical factor models. The next two sections discuss the issues related to its implementation. Principal Component Analysis is usually applied to matrices (or *panels*) that do not have a time dimension. In contrast, we deal with temporal data; and we cannot assume that these data are drawn in each period from the same probability distribution. We will employ the PCA and SVD *locally*, i.e., on intervals in which the data can be presumed to be approximately stationary. We present two approaches that are used by practitioners. Without any aspiration to establish a winner, we compare their performance on historical US equity data.

9.5.1 Weighted and Two-Stage PCA

A recurring theme in factor estimation is that weighting observations differently helps. Observations in the distant past are less informative than recent ones; observed returns of stocks with high idiosyncratic risk should be downweighted, compared to those of low-idio stocks. There are therefore two basic transformations that we can apply to the raw return matrix. The first one is in the time dimension. We replace the empirical covariance matrix in Equation (9.9) with a weighted one. Let $\mathbf{W}_\tau \in \mathbb{R}^{T \times T}$ a diagonal matrix with positive diagonal elements. The diagonal terms could be, for example, exponential weights $[\mathbf{W}_\tau]_{t,t} = \kappa \exp(-t/\tau)$; the positive constant κ is such that the diagonal terms sum to T . Then the time-weighted empirical uncentered covariance matrix is

$$\hat{\Sigma} = \frac{1}{T} \mathbf{R} \mathbf{W}_\tau^2 \mathbf{R}' \quad (9.57)$$

This is the same as first transforming the returns $\tilde{\mathbf{R}} = \mathbf{R} \mathbf{W}$, and then computing the empirical covariance matrix, Eq. (9.9), on the transformed returns. In practice, we would not compute the covariance matrix and then perform the PCA, but rather perform the SVD on $\tilde{\mathbf{R}}$, which would be computationally less expensive and give us the same results.

A different type of transformation is cross-sectional reweighting. In the fundamental factor models chapter we saw that it is optimal to scale returns by the idiosyncratic volatility, or at least a proxy. Similarly to Boivin and Ng (2006), I propose a two-step procedure. First, perform an SVD (possibly, time-weighted) on the returns; $\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}'$. Take the first p components (say, $p = 5$) and compute the idiosyncratic returns $\mathbf{E} = \mathbf{R} - \mathbf{U}_p \mathbf{S}_p \mathbf{V}'_p$; a case we also consider is $p = 0$, in which case R . Define the proxy idiosyncratic volatilities: $\sigma_i^2 = \sum_i [\mathbf{E}]_{i,i}^2$, and $\mathbf{W}_\sigma := \text{diag}(\sigma_1^{-1}, \dots, \sigma_n^{-1})$. The asset-level reweighted covariance matrix is

$$\hat{\Sigma} = \mathbf{W}_\sigma \mathbf{R} \mathbf{R}' \mathbf{W}_\sigma \quad (9.58)$$

One can perform then a second-stage PCA and a factor model on the reweighted covariance matrix: $\hat{\Sigma} \sim \mathbf{U}_m \mathbf{S}_m^2 \mathbf{U}'_m + \mathbf{I}$. Finally, pre- and post-multiply by the idiosyncratic weighting matrices \mathbf{W}_σ .

We employ the steps above in the following process. We use two time-series reweightings: one with half-life τ_f (f is for “fast”) and τ_s (s is for “slow”). An empirical insight in asset return data is that volatilities and correlations change over different timescales. Volatilities change rapidly; in fact they may change dramatically over the course of a few days. The ratio between the

volatility of a stock during a crisis can four times as large as the volatility of the same asset during a quiet period. On the other side, pairwise correlations are quite stable. Even in the presence of major market stresses, these correlations marginally increase in absolute value. This suggests that we separate volatilities and correlations. Therefore, in the first stage, we use a short half-life to capture adequately changes in volatility. In the second stage, we use a longer-half life to estimate the factor structure of correlations.

This procedure is flexible enough to include several PCA-related procedures as special cases, and to serve as a basis for further experimentation. Some examples:

- When $p = 0$, then idio reweighting becomes a z-scoring, so that the second-stage PCA is effectively applied to the correlation matrix.
- The special case of equal-weighted observations in time is obtained in the limit $\tau \rightarrow \infty$.
- It is straightforward to use different shrinkage methods in the second-stage factor model step.
- In the second-stage factor model step, we use the Probabilistic PCA results of Section 9.1.2. The idio reweighting steps approximately “whitens” the idiosyncratic returns, i.e., it makes them unit-variance, so that PPCA applies. However, we could replace this a different estimation procedure, like Maximum Likelihood.

9.5.2 Implementing Statistical Models in Production

It is not sufficient to have a procedure that estimates the loadings and the covariance matrix at a point in time. In our applications, factor models are *dynamic*. At time t , we have an estimation universe of stocks, and we use return data up to T_{\max} periods in the past. We apply the two-stage PCA using returns data between $t - T_{\max} + 1$ and t , to obtain:

- loadings \mathbf{B}_t . This is the output loadings matrix.
- factor returns and idio returns estimate at time t :

$$\hat{\mathbf{f}}_t = (\mathbf{B}'_{t-1} \mathbf{W}_{\sigma,t-1}^2 \mathbf{B}_{t-1})^{-1} \mathbf{B}'_{t-1} \mathbf{W}_{\sigma,t-1}^2 \mathbf{r}_t \quad (9.59)$$

$$= [\hat{\mathbf{U}}_m]'_t \mathbf{W}_\sigma \mathbf{r}_t \quad (9.60)$$

$$\hat{\boldsymbol{\epsilon}}_t = \mathbf{r}_t - \mathbf{B}_t \hat{\mathbf{f}}_t \quad (9.61)$$

We need to address some outstanding problems:

1. Sign indeterminacy of eigenvectors;
2. Time-changing estimation universe;

Procedure 9.1: *Statistical Model Estimation*

1. **Inputs:** $\mathbf{R} \in \mathbb{R}^{n \times T}$, $\tau_s \geq \tau_f > 0$, $p \in \mathbb{N}$, $m > 0$.

2. **Time-Series Reweighting:**

$$\begin{aligned}\mathbf{W}_{\tau_f} &:= \kappa \operatorname{diag}(\exp(-T/\tau_f), \dots, \exp(-1/\tau_f)) \\ \tilde{\mathbf{R}} &= \mathbf{R} \mathbf{W}_{\tau_f}\end{aligned}$$

3. **First stage PCA:** $\tilde{\mathbf{R}} := \tilde{\mathbf{U}} \tilde{\mathbf{S}} \tilde{\mathbf{V}}'$

4. **Idio Proxy Estimation:**

$$\begin{aligned}\mathbf{E} &= \tilde{\mathbf{R}} - \tilde{\mathbf{U}}_p \tilde{\mathbf{S}}_p \tilde{\mathbf{V}}'_p && (\text{truncated SVD}) \\ \sigma_i^2 &= \sum_t [\mathbf{E}]_{i,t}^2 && (\text{idio vol proxies}) \\ \mathbf{W}_\sigma &:= \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_n^{-1})\end{aligned}$$

5. **Idio Reweighting:**

$$\begin{aligned}\mathbf{W}_{\tau_s} &:= \kappa \operatorname{diag}(\exp(-T/\tau_s), \dots, \exp(-1/\tau_s)) \\ \hat{\mathbf{R}} &:= \mathbf{W}_\sigma \mathbf{R} \mathbf{W}_{\tau_s}\end{aligned}$$

6. **Second Stage PCA:** $\hat{\mathbf{R}} := \hat{\mathbf{U}} \hat{\mathbf{S}} \hat{\mathbf{V}}$

7. **Second-Stage Factor Model:** $\hat{\mathbf{r}} = \hat{\mathbf{U}}_m \mathbf{f} + \hat{\boldsymbol{\epsilon}}$

$$\begin{aligned}\text{where: } \mathbf{f} &\sim N(0, \operatorname{diag}(\ell(s_1^2), \dots, \ell(s_m^2))) \\ \boldsymbol{\epsilon} &\sim N(0, \bar{\lambda} \mathbf{I}_n)\end{aligned}$$

$$\bar{\lambda} = \frac{1}{n-m} \sum_{i=m+1}^n s_i^2$$

8. **Output: Final Factor Model:** $\mathbf{r} := \mathbf{B} \mathbf{f} + \boldsymbol{\epsilon}$

$$\begin{aligned}\text{where: } \mathbf{B} &= \mathbf{W}_\sigma^{-1} \hat{\mathbf{U}}_m \\ \mathbf{f} &\sim N(0, \operatorname{diag}(\ell(s_1^2), \dots, \ell(s_m^2))) \\ \boldsymbol{\epsilon} &\sim N(0, \bar{\lambda} \hat{\mathbf{W}}_\sigma^{-2})\end{aligned}$$

3. Imputation of loadings for non-estimation universe assets;
4. Imputation of missing values for new or temporarily non-traded assets.
5. Adjustment for corporate events.

We tackle them in this order.

Sign indeterminacy of eigenvectors. Let us begin with a simple observation. In a statistical model, eigenvectors are identified modulo a sign change, i.e., if \mathbf{u} is an eigenvector of matrix Σ and associated eigenvalue λ , then so is $-\mathbf{u}$. When we compute the SVD for adjacent periods, we add and remove observations, which may lead to a sign flip in the loadings. It is important therefore that loadings be collinear, in the sense that the cosine angle¹¹ as between eigenvectors in adjacent periods be positive. Aside from the straightforward realignment exercise, the turnover of eigenvectors is important in two respects. First, because, if we observe that $S_C(\mathbf{u}^i(t), \mathbf{u}^i(t+1)) \simeq 0$, then it is difficult to determine the sign of consecutive eigenvalues. As a result, it is difficult to determine the sign of the factor return $f_i(t)$ over time. Aside from any statistical considerations, high-turnover statistical factors cannot be employed for performance attribution. The second consideration is that a very high-turnover factor results in factor-mimicking portfolios with very high turnover as well, and is therefore a factor that is very difficult to trade, either for hedging or speculation purposes.

Time-changing estimation universe. Similarly to fundamental models, statistical models are estimated on a predetermined set of assets. The rationale for the choice of such universe is the same as for fundamental models. Estimation universe assets may be representative of the investment universe of the trading strategy; may be sufficiently liquid to be considered tradeable; and, relatedly, may be sufficiently traded to ensure good price discovery and therefore reliable return calculations. Assets enter and leave the universe over time. We have a dilemma. We cannot use the latest universe composition, because the past returns of recent additions to the index may be unreliable because the asset was illiquid, or be missing altogether. We can still opt to keep these recent additions, provided that their returns are well defined; or alternatively we can use the assets at the intersection of all the universes over the time interval used for estimation. If the time interval used for model estimation is not too long,

¹¹The cosine angle (or cosine similarity) between two vectors \mathbf{u}, \mathbf{v} is defined as $S_C(\mathbf{u}, \mathbf{v}) := (\mathbf{u}'\mathbf{v})/(\|\mathbf{u}\| \|\mathbf{v}\|)$.

and if the universe turnover is not too high, we will still have a sufficiently broad panel of assets. It is preferable to employ an estimation universe that has the lowest possible turnover, and it is important to use a consistent procedure to select the assets to include in \mathbf{R} .

Imputing loadings for non-estimation universe assets. There are assets that are not in the estimation universe, but that have complete returns. They do not have loadings. We can impute loadings by performing a time-series regression of asset returns against the factor returns. This approach is justified by the results in Subsection 9.1.3: we can recover loadings from time series regression, provided that the factor returns we obtained using the estimation universe are close to the true factor returns. Below is a simple numerical example. We use one year of returns, and a universe corresponding approximately to the Russell 3000. We first perform a two-stage PCA on the entire universe, and estimate \mathbf{B} and \mathbf{F} . To test that estimated asset loadings can be recovered, we perform the same analysis, but this time on a universe of 2000 assets chosen at random. We now take the estimated factor returns, and estimated the loadings using a time-series regression for the assets we held out. Finally, we compare the loadings from the first PCA, the “population” loadings, to those from time series regression, the “imputed” loadings.

[*** TABLE AND CHARTS HERE***]

Imputation of missing values for new or temporarily non-traded assets. Some assets do not have sufficient return history to regress their loadings; examples are newly-listed assets (IPOs, ADRs), or assets that were either delisted for a long period of time, or had trading volumes considered too low to result in reliable returns. A possible solution is to use additional characteristics of the asset to impute its loadings. The approach is similar to the one we presented in Section 9.4.2 on the interpretation of loadings using regression. In this case, however, we usually are not afforded the luxury to know many of the asset’s style characteristics like momentum, beta, liquidity, or profitability. All we have is knowledge of the industry and country of the asset. We regress observed loadings against these two characteristics, and predict the missing loadings. It is common practice to shrink predicted loadings toward zero. We will cover a rationale for this practice in later sections devoted to hedging.

9.6 Further Reading

Standard references for PCA are Jolliffe (2010), Jolliffe and Cadima (2016), Johnson and Wichern (2007), Pourahmadi (2013), Yao et al. (2015); PCA is also covered in any popular graduate-level textbook on Statistical Learning, e.g. Hastie et al. (2008), Murphy (2012), Bishop (2006). Skillicorn (2007) is devoted to the interpretation of SVD, PCA, Non-negative Matrix Factorization and its applications.

In the statistical literature, the analysis of this model begins with Johnstone (2001). In a seminal paper, Chamberlain and Rothschild (1983) impose similar conditions, but in an asymptotic setting, by considering an increasing sequence of asset universes (with $n \rightarrow \infty$) and risk models in which the diversifiable risk goes to zero.

We have only touched briefly on the asymptotic limit of the spiked model in Section 9.2.2, to give a taste of what happens and give a basis for heuristics. Several papers have characterized the behavior the model. The first and seminal result is by Baik, Ben Arous and Péché (Baik et al., 2005), and the theorem is named BBP theorem after their initials. Several authors have generalized these results: Baik and Silverstein (2006); Paul (2017); Bai and Yao (2008); Mestre (2008); El Karoui (2008); Benaych-Georges and Nadakuditi (2011); Shen et al. (2016); Wang and Fan (2017); a survey is Johnstone and Paul (2018). General surveys on Random Matrix Theory, with a eye toward finance are Bun et al. (2017) and Bouchaud and Potters (2020). The line of research concerned with properties of the spectrum in the regime “ $p/n \gg 1$ ” begins¹² perhaps with Johnstone (2001). From the very first result on biased asymptotic estimators, a reader may wonder about shrinkage methods. There is an extensive literature on factor model shrinkage. Standard references are Ledoit and Wolf (2003a,b, 2004) on linear shrinkage, and more recent work on nonlinear shrinkage by the same authors (Ledoit and Wolf, 2012, 2015, 2020). The paper by Donoho et al. (2018) covers optimal shrinkage functions for a large set of loss functions.

For a relatively old survey on methods to select the number of factors, see Ferré (1995); a more recent survey is in Fan et al. (2020). The scree plot method is due to Cattell (1966), and its logarithmic version by Farmer (1971). The scree is the debris that form at the base of a cliff.

In the econometric literature, there are at least two strands of research: static factor models and dynamic factor models. For the former, see the survey

¹²The academic literature denotes the number of variables with p and the number of observations with n .

by Bai and Ng (2008); for the latter, Stock and Watson (2016).

The two-step procedure for reweighting the PCA is relatively common; Boivin and Ng (2006) reweights using idio volatilities, and Bollerslev (1990) using total volatilities.

The connection between PCA and clustering was made in the seminal paper by Ding and He (2004).

9.7 Exercises

N.B.: the solution sketches will be grouped at the end of the book.

Exercise 9.1 (Low-rank factorization). *Prove that a matrix $\mathbf{A} \in \mathbb{R}^{n \times T}$ is of rank $m \leq \min\{n, T\}$ if and only if it can be decomposed in the product of two matrices $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{C} \in \mathbb{R}^{m \times T}$.*

Exercise 9.2 (PCA solution). *Prove that the solution \mathbf{w}^* in Problems (9.7) and (9.10) is unique and that constraint $\|\mathbf{w}\|^2 \leq 1$ is always binding, i.e., $\|\mathbf{w}^*\| = 1$.*

Exercise 9.3 (Alternative PCA formulation). *Prove that the optimization (9.11) gives the same solution as finding the first m eigenvectors of $\hat{\Sigma}$, and as finding iteratively k unit-norm vectors $\mathbf{w}_1, \dots, \mathbf{w}_k$, with \mathbf{w}_k orthogonal to the first $k - 1$ vectors \mathbf{w}_k , that maximize $\mathbf{w}_k' \hat{\Sigma} \mathbf{w}_k$.*

Exercise 9.4 (Covariance Matrix of a Linear Transformation). *Prove that if the random vector \mathbf{r} taking values in \mathbb{R}^n has covariance matrix Σ , and if $\mathbf{A} \in \mathbb{R}^{m \times n}$, then the random vectors $\mathbf{x} = \mathbf{A}\mathbf{r}$ has covariance matrix $\mathbf{A}\Sigma\mathbf{A}'$.*

Exercise 9.5 (A Simple Spiked Matrix). *Let $\mathbf{B} \in \mathbb{R}^{n \times m}$ be an m -rank matrix. Prove that the first m eigenvalues of $\mathbf{B}\mathbf{B}' + \sigma^2 \mathbf{I}_n$ are greater than σ^2 .*

Exercise 9.6. *Solve the optimization problem, Eq. (9.16).*

Exercise 9.7 (The Power Method). *A simple (the simplest?) algorithm for computing the largest eigenvalue of a symmetric p.d. matrix Σ is the following:*

1. start with a unit-norm \mathbf{x}_0 chosen at random (say, sample the coordinates from a standard normal distribution, then normalize it);
2. iterate: $\mathbf{x}_{i+1} = \Sigma \mathbf{x}_i / \|\mathbf{x}_i\|$;
3. after the vector converges (say $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|$ is smaller than some tolerance), \mathbf{x}_{i+1} approximates the top eigenvector, and $\mathbf{x}_i' \Sigma \mathbf{x}_i$ the top eigenvalue.

1. *Prove the convergence and correctness of the power method. (Hint: $\mathbf{x}_i = \Sigma^i \mathbf{x}_0 / \|\Sigma^i \mathbf{x}_0\|$.)*

2. Let $i \geq \Omega(\log(1/\delta)/\delta)$. Prove that $\mathbf{x}_i' \Sigma \mathbf{x}_i \geq (1 - \delta)\lambda_1$.
3. How would you extend it to find all the eigenvalues of Σ ?

Exercise 9.8 (Iterated Projections for the SVD). A simple (the simplest?) algorithm for computing the largest singular value of a matrix $\mathbf{R} \in \mathbb{R}^{n \times T}$ is the following: 1. start with $\mathbf{x}_0 \in \mathbb{R}^T$ chosen at random (say, sample the coordinates from a standard normal distribution), 2. iterate:

$$\mathbf{y}_{i+1} = \mathbf{R}\mathbf{x}_i \quad (9.62)$$

$$\mathbf{x}_{i+1} = \mathbf{R}'\mathbf{y}_{i+1} \quad (9.63)$$

3. after the vectors converge \mathbf{x}_{i+1} approximates the highest left eigenvector, \mathbf{y}_{i+1} the higher right eigenvector, and $\mathbf{y}_{i+1}' \mathbf{R} \mathbf{x}_{i+1}$ the top singular value.

1. Prove the convergence and correctness of the algorithm (hint: power method);
2. How would you extend it to find the SVD of \mathbf{R} ? (Hint: not the same may of the power method).

Exercise 9.9 (Time Series Regression from the SVD). Let $\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}'$, and set $\hat{\mathbf{F}} := \mathbf{S}_m \mathbf{V}'_m$. The vector $\hat{\mathbf{f}}_i$, the i th row of $\hat{\mathbf{F}}$, is the time series of the i th factor return. Prove that the least-squares regression coefficient of the time series $\mathbf{r}_i = \beta_{i,j} \hat{\mathbf{f}}_j + \epsilon$, is $\beta_{i,j} = [\mathbf{U}]_{i,j}$.

Exercise 9.10 (Oja's Iterative Algorithm). Let $\mathbf{r}_t, t = 1, \dots, T$ be a time series of returns drawn from a common distribution on \mathbb{R}^n with covariance matrix Ω . Prove that the following algorithm converges to the first eigenvector of Ω :

1. Set $i = 0$ and choose a unit-norm $\mathbf{v}_1 \in B^n$ uniformly at random.
2. Choose column $\pi(i)$ uniformly at random between 1 and T .
3. Update the direction

$$\mathbf{v}_{i+1} = \mathbf{v}_n + i^{-1}(1 - \mathbf{v}_i' \mathbf{e})(\mathbf{r}'_{\pi(i)} \mathbf{v}_i) \mathbf{r}_{\pi(i)} \quad (9.64)$$

$$\mathbf{v}_{i+1} \leftarrow \frac{\mathbf{v}_{i+1}}{\|\mathbf{v}_{i+1}\|} \quad (9.65)$$

4. Set $i \leftarrow i + 1$. If $\|\mathbf{v}_{i+1} - \mathbf{v}_i\|$ then stop. Otherwise go to Step 2.

Solution (sketch): Let $\mathbf{R} \in \mathbb{R}^{n \times T}$, and \mathbf{X} a random matrix taking values in $\mathbb{R}^{n \times n}$. \mathbf{X} takes one of T values: $\mathbf{r}_i \mathbf{r}'_i$ with equal probability $1/T$. One can interpret the product $T^{-1} \mathbf{v}' \mathbf{R} \mathbf{R}' \mathbf{v}$ as the expectation $E(\mathbf{v}' \mathbf{X} \mathbf{v})$. The first eigenvalue of $T^{-1} \mathbf{R} \mathbf{R}'$ is

$$\max_{\|\mathbf{v}\|=1} E \left(\frac{\mathbf{v}' \mathbf{X} \mathbf{v}}{\|\mathbf{v}\|^2} \right) \quad (9.66)$$

We can apply the stochastic gradient algorithm to the maximum search. Let $f(\mathbf{X}, \mathbf{v}) := (\mathbf{v}' \mathbf{X} \mathbf{v}) / \|\mathbf{v}\|^2 - \lambda \|\mathbf{v}\|^2$. The derivative $\nabla_{\mathbf{v}} f$ for a unit-norm vector $\|\mathbf{v}\|$ is

$$2(1 - \mathbf{v}' \mathbf{e}) \mathbf{X} \mathbf{v} \quad (9.67)$$

Exercise 9.11 (Distance Between Subspaces). Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$, be orthonormal matrices. If the two column subspaces are “similar”, then any unit-norm vector in the column subspace of \mathbf{A} is well-approximated by some unit-norm vector in the column subspace of \mathbf{B} . Define similarity between the two subspaces

$$S(\mathbf{A}, \mathbf{B}) := \frac{1}{2} \max_{\|\mathbf{x}\| \leq 1} \min_{\|\mathbf{y}\| \leq 1} \|\mathbf{Ax} - \mathbf{By}\|^2 \quad (9.68)$$

1. Prove that $S(\mathbf{A}, \mathbf{B})$ is $1 - \sigma_1(\mathbf{A}' \mathbf{B})$, where $\sigma_1(\mathbf{A}' \mathbf{B})$ is the first singular value of $\mathbf{A}' \mathbf{B}$.
2. Prove that $S(\mathbf{A}, \mathbf{B})$ is not a distance because it does not satisfy the triangle inequality.

Exercise 9.12 (Angle Between Subspaces). Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$, be orthonormal matrices. Let the least cosine distance between subspaces be the cosine of the smallest achievable angle between two vectors, one belonging to the column subspace of \mathbf{A} , the other belonging to the column subspace of \mathbf{B} .

Prove that $S_C(\mathbf{A}, \mathbf{B}) = \sigma_n(\mathbf{A}' \mathbf{B})$, where $\sigma_m(\mathbf{A}' \mathbf{B})$ is the last singular value of $\mathbf{A}' \mathbf{B}$.

Exercise 9.13 (Covariance Matrix Invariance for Degenerate Eigenvalues). Consider a risk model with the following structure: its loading matrix \mathbf{B} has the form $\mathbf{B} = \mathbf{D} \mathbf{U}$, where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is diagonal positive definite, and $\mathbf{U} \in \mathbb{R}^{n \times m}$, $\mathbf{U}' \mathbf{U} = \mathbf{I}_m$; and its factor covariance matrix is proportional to the identity: $\Sigma_f = \bar{\lambda} \mathbf{I}_m$.

1. Prove that if we replace \mathbf{U} with an “equivalent” $\tilde{\mathbf{U}} \in \mathbb{R}^{m \times n}$ spanning the same subspace, the covariance matrix does not change.

2. Extend the result to the case where Σ_f is still diagonal, but with the first p variances being greater than the rest: $\lambda_1 > \lambda_2 > \dots > \lambda_p > \lambda_{p+1} = \dots = \lambda_m$, and with

$$\tilde{\mathbf{U}}[:, 1:p] = \mathbf{U}[:, 1:p]$$
$$\tilde{\mathbf{U}}[:, (p+1):m]' \tilde{\mathbf{U}}[:, (p+1):m] = \mathbf{I}_{m-p}$$

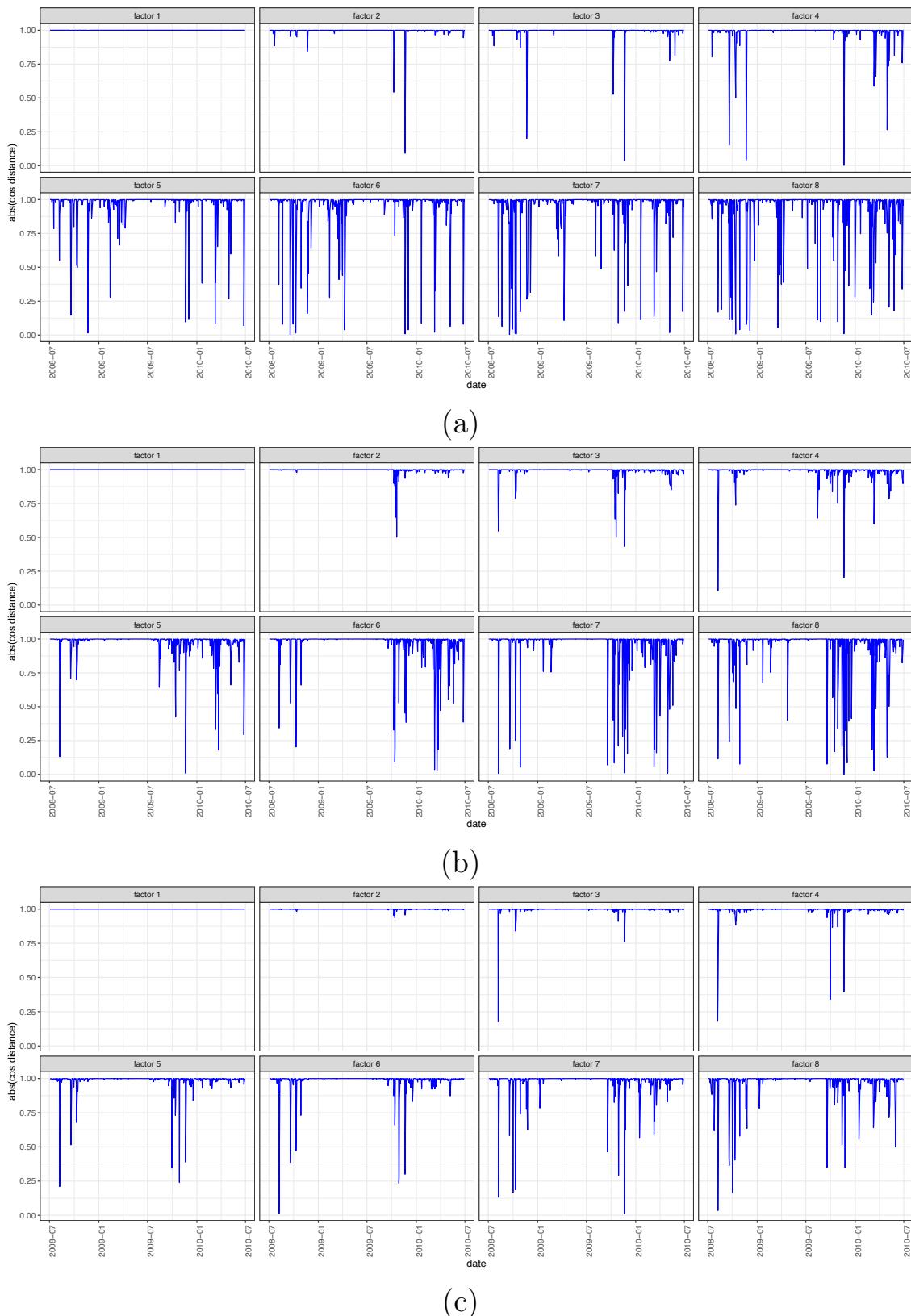


Figure 9.8: Eigenfactor turnover for different covariance matrices. (a): total returns; (b) total returns/total vol; (c) total returns/idio vol.

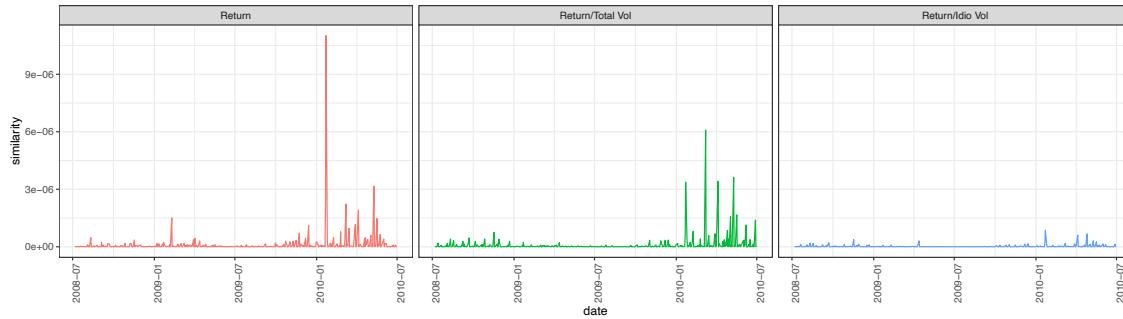


Figure 9.9: Distance between column subspaces of the first eight eigenfactors in consecutive periods. The eigenfactors are generated by PCAs on total returns, total returns/total vol, and total return/idio vol.

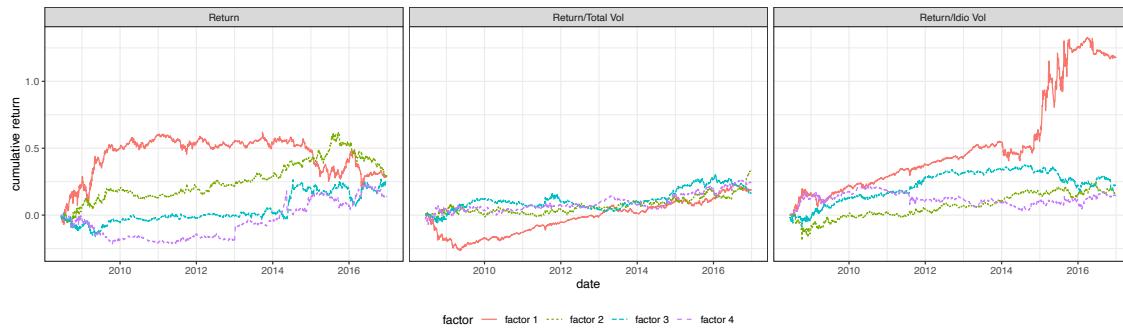


Figure 9.10: Factor returns for the first four eigenvectors. The eigenfactors are generated by PCAs on total returns, total returns/total vol, and total return/idio vol.

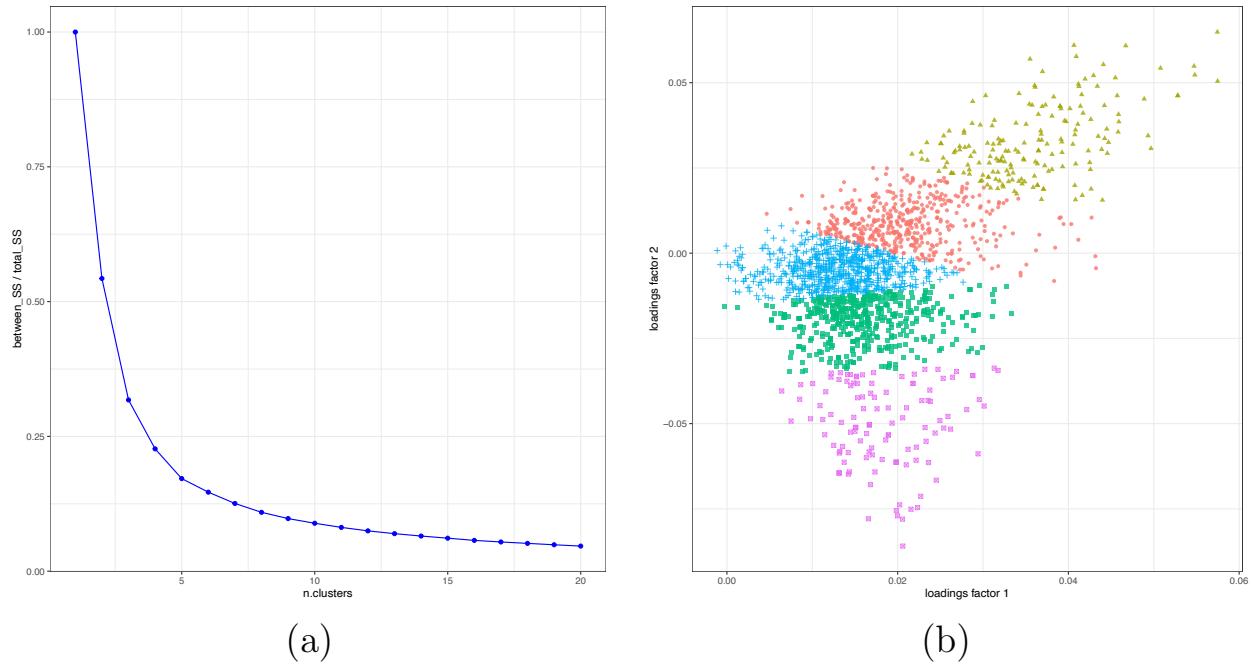


Figure 9.11: (a) Performance of clustering; b) Five clusters of the loadings of the first two principal components.

Part II

During The Trade

Chapter 11

Hedging

Hedging is the process of reducing the risk of a portfolio by means of augmenting the portfolio with additional investments, whose returns are negatively correlated to the existing portfolio. The most common forms of hedging are market hedging and currency hedging, but there are at least three additional cases of practical relevance to portfolio managers. The first is hedging to factor-mimicking portfolios obtained from a fundamental factor model. The second one is hedging to a future or liquid asset capturing non-equity risk. This includes energy and interest rate futures, and liquid ETFs and ETN describing sector or style risk. The last application of interest is the production of thematic tradable baskets by banks. One can buy these baskets to hedge or speculate on political risk (e.g., elections) or thematic risks (e.g., Citi has a global thematic engine with 80+ industry trends).

The chapter is broadly organized in three parts. The first one is covers vanilla hedging. There are no villains in this story—no transaction costs, no parameter uncertainty, and a single period. Yet, such a simple model is still widely used in a wide range of applications. In the second section we explore the impact of parameter error and how it affects optimal hedging. Lastly, we look at multi-period hedging in the presence of execution costs.

11.1 Toy Story

In its simplest form, we have the following ingredients:

- We have two decision epochs t_0, t_1 , and one realized return between them. We make investment decisions at t_0 , and observe realized returns at t_1 .
- We have two assets, which we denote *core* and *hedge* with expected returns $\mu_c, \mu_h = 0$, volatilities σ_c, σ_h , and return correlation between the two is

ρ . The first asset is the core portfolio , the second one is the hedging portfolio.

We decide the size of the hedging instrument in order to maximize the Sharpe Ratio of the combined portfolio. You already see how similar this problem is to the two-asset Mean-Variance Optimization instance we saw in Section 5.1. In that problem, we decided the optimal positions of both assets; not a major difference. The MVO optimization problem

$$\max_{x_h \in \mathbb{R}} \mu_c x_c - \frac{\lambda}{2} (\sigma_c^2 x_c^2 + \sigma_h^2 x_h^2 + 2\rho\sigma_h\sigma_c x_c x_h)$$

has solution

$$x_h^* = -\frac{\rho\sigma_c x_c}{\sigma_h} \quad (11.1)$$

$$\frac{x_h^*}{x_c} = -\frac{\rho\sigma_c}{\sigma_h} = -\beta(r_h, r_c) \quad (11.2)$$

The ratio $|x_h^*/x_c|$ is the *optimal hedge ratio* and is equal to the beta of the core portfolio's return to the hedging portfolio's return.

The unhedged variance is $\sigma_c^2 x_c^2$; after the hedge it is $(1 - \rho^2)\sigma_c^2 x_c^2$. The improvement in Sharpe Ratio is equal to the improvement in volatility:

$$\frac{\text{SR(hedged)}}{\text{SR(native)}} = \frac{1}{\sqrt{1 - \rho^2}} \quad (11.3)$$

The parameter beta is estimated either via time-series regression or by using a return covariance matrix, such as one supplied by a factor model. Define $\mathbf{w}_c, \mathbf{w}_h$ the core and hedge portfolios; the model beta is

$$\beta(r_c, r_h) = \frac{\mathbf{w}'_c \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w}_h}{\mathbf{w}'_h \boldsymbol{\Omega}_{\mathbf{r}} \mathbf{w}_h} \quad (11.4)$$

From that, formula (11.2) gives the relative size of the hedge, and formula (11.3) the improvement in Sharpe Ratio from hedging.

In their simplicity, Equations (11.1, 11.2, 11.3) are applied widely. A typical application involves the use of a single hedging instrument that is very liquid and inexpensive to trade, and whose expected return is negligible compared to that of the native portfolio. We perform intraday or end-of-day hedging in order to remove the associated risk.

Hedging in this specific instance rests on several implicit and explicit assumptions:

Procedure 11.1: Simple Single-Asset Hedging

1. **Inputs:** core portfolio NMV x_c with returns r_c . Hedging asset with return r_h . Parameter $\beta(r_h, r_c)$, obtained by means of time-series regression, or of an asset covariance matrix.
2. **Output:** Hedge NMV $x_h^* = -\beta(r_h, r_c)x_c$.

- We assume that the beta of the core portfolio to the hedging instrument can be estimated accurately;
- We assume that there is a single trading instrument;
- We assume that the trading costs are negligible;
- We assume that the hedging instrument has negligible expected return.

In the remainder of this chapter we reexamine these hypotheses and relax them.

11.2 Factor Hedging

11.2.1 The General Case

Factor models have made their appearance repeatedly in this book, and unsurprisingly they matter for hedging as well. In principle, portfolio construction should take into account predicted risk arising from factor exposures and from idiosyncratic bets, and generate a portfolio that meets our investment goals. In practice, there are situations in which this is not possible. An example is the one in which the core portfolio is the outcome of a portfolio construction process outside of our control. For example, we may have several groups of independent discretionary portfolio managers trading stocks based on their fundamental outlook. The sum of their individual portfolios constitutes a core portfolio that is not optimized, and that exhibits undesired systematic risk. In this case, the hedging process takes w_c as an input, and seeks to reduce the unwanted risk from factor exposures¹. We defined factor-mimicking portfolios in Chapter 4,

¹We are assuming, again, that the returns of the factors we want to hedge are zero, or negligible

Equation (4.11): they are the columns of matrix \mathbf{P} , and have unit exposure to factor i . One way to do so would be Procedure 11.2.

We have achieved zero factor exposure. The solution is simple, elegant, and

Procedure 11.2: A Simple Factor Hedging Procedure

1. Compute the core portfolio factor exposure $\mathbf{b}_c = \mathbf{B}'\mathbf{w}_c$.
2. “Trade out” the core exposure by buying an amount of factor exposure $-\mathbf{b}_c$. We do this by buying a hedge portfolio $-\mathbf{P}\mathbf{b}_c$.

unfortunately unrealistic. We have ignored two essential aspects of the hedging problem. First, factors have non-zero expected returns. Second, trading factor is expensive. However, we can change the formulation to include these modeling concerns. Let us begin with accounting the non-zero expected return of the hedging portfolio. To this end, we need to go back to Section 3.3, which introduced the definition of alpha orthogonal and alpha spanned. In formulas: the expected return of a portfolio \mathbf{w} is equal to $(\boldsymbol{\alpha}'_{\perp} + \boldsymbol{\mu}'\mathbf{B}')\mathbf{w}$. Regarding the execution costs, we can include them in the optimization formulation, using a square root impact model, or a quadratic model, as seen in Chapter 10. We denote the expected trading cost of a portfolio \mathbf{w} by $f(\mathbf{w})$. In a single-period setting, we can then write the problem as

$$\begin{aligned} & \max \boldsymbol{\alpha}'_{\perp}(\mathbf{w}_c + \mathbf{w}_h) + \boldsymbol{\mu}'\mathbf{b} - \frac{1}{2}\rho(\sigma_f^2 + \sigma_i^2) - f(\mathbf{w}_h - \mathbf{w}_{h,0}) \\ & \text{s.t. } \mathbf{b} = \mathbf{B}'(\mathbf{w}_c + \mathbf{w}_h) \\ & \quad \sigma_f^2 = \mathbf{b}'\boldsymbol{\Omega}_f\mathbf{b} \\ & \quad \sigma_i^2 = (\mathbf{w}_c + \mathbf{w}_h)' \boldsymbol{\Omega}_{\epsilon} (\mathbf{w}_c + \mathbf{w}_h) \\ & \quad \mathbf{w}_h \in \mathbb{R}^n \end{aligned} \tag{11.5}$$

I leave it as an exercise to prove that if execution costs are zero, orthogonal and spanned alphas are zero, and factor portfolio have zero idiosyncratic variance, then *of course* we would hedge out exposure. Not a single one of these assumption holds, and it’s worth spending some time commenting on them.

- Some of the factors do have zero expected returns², some don’t. Hedging

²Sometimes these are referred to as *unpriced factors*, because we receive no reward for holding their associated risk.

them can in fact be counterproductive because the gain in Sharpe Ratio are countered by expected PnL losses.

- The hedging portfolio may also have non-zero alpha orthogonal exposure. This must be taken into account, especially when alpha orthogonal is indeed what the profitability of the strategy depends on it, more than on alpha spanned.
- Even if we traded the pure factor-mimicking portfolios of Procedure 11.2, we would add idiosyncratic risk to our core portfolio. This additional idiosyncratic risk reduces the benefits of factor risk reduction. The optimization formulation takes this into account. In fact, the following Exercise asks you to work out the details and to show that the optimal hedging is not equal to $-\mathbf{b}_c$.

Exercise 11.1. (35) Assume that:

1. factor portfolios have zero expected returns;
2. we hedge only using factor portfolios;
3. we have no transaction costs.

Prove that the optimal hedging policy is

$$\begin{aligned}\mathbf{x}^* &= -(\mathbf{\Omega}_f + (\mathbf{B}'\mathbf{\Omega}_\epsilon^{-1}\mathbf{B})^{-1})^{-1}(\mathbf{B}'\mathbf{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{b}_c \\ &= -[\mathbf{I} + (\mathbf{B}'\mathbf{\Omega}_\epsilon^{-1}\mathbf{B})\mathbf{\Omega}_f]^{-1}\mathbf{b}_c\end{aligned}$$

Under what condition is the optimal hedging smaller than perfect factor neutralization of Procedure 11.2?

The solution is the Appendix. Meanwhile, here is a much easier problem to get you started:

Exercise 11.2. (10) For simplicity, consider the case where asset returns are described by one-factor model, and that there is a hedging portfolio that has exposure to that factor. Starting with Equation (11.4), show that it is optimal not to hedge entirely the exposure of the core portfolio to that factor.

- In the simplistic hedging procedures of the first part of this chapter, we could ignore our investment objective, because volatility reduction

was a zero-cost improvement: no execution concerns, no expected factor returns, no idiosyncratic volatility increase. But reality is complicated. The parameter ρ quantifies our risk tolerance and determines where do we want to be on the curve trading off volatility for expected costs. *This is a good thing.* In practice, we should explore this trade-off and determine the optimal operating point.

- In the special case of quadratic costs, Optimization Problem (11.5) can be rewritten as a multi-period optimization problem and solved using the techniques presented in Chapter 10 and specifically in Procedure 10.1. This is the subject of the following...

Exercise 11.3. (35) Extend Problem Optimization Problem (11.5) to the multiperiod setting. Discuss implementation complexity and propose some simplifying assumptions.

11.3 Hedging Tradable Factors with Time-Series Betas

A relatively common use case for hedging is the following. We have some non-equity tradable and liquid instrument that is associated to macroeconomic movements; for example, energy or metal commodity futures; or fixed income futures. Because of their ability to capture broad macroeconomic themes and of their liquidity, we would like to use these instruments for hedging. To fix our ideas further, consider the case of a portfolio composed of energy stocks, and of gas and crude future contracts, which are among the most liquid in the world. It stands to reason that the energy portfolio is correlated to energy prices, and at the same time that the portfolio manager or the trading algorithm does not have a view on the future price energy movements. A possible approach is to estimate time-series betas $\hat{\beta}_i := \beta(r_i, r_h)$, and then hedge the exposure $\hat{\beta}' \mathbf{w}_c$ using Procedure 11.1. Would this approach work? Surprisingly, in more than one instance, the realized risk of the *hedged* portfolio is worse than the realized risk of the core portfolio? This is somewhat counterintuitive. In this section, we try to shed some light on hedging for this particular scenario.

As in the previous sections, we denote the return of the tradable instrument r_h , with variance σ_h^2 . We model the estimated betas as $\hat{\beta}_i = \beta_i + \eta_i$. We denote with $\boldsymbol{\eta}$ the random vector of estimation error with covariance matrix $\Omega_{\boldsymbol{\eta}}$, and with $\boldsymbol{\beta}$ the vector of true betas. In order to see what could go wrong, let us hedge

with the “optimal” hedge ratio $x_h^* = -\hat{\beta}' \mathbf{w}$. The covariance matrix, augmented with the hedging instrument, is:

$$\begin{pmatrix} \Omega_r & \sigma_h^2 \beta \\ \sigma_h^2 \beta' & \sigma_h^2 \end{pmatrix}$$

$$\begin{aligned} \text{var}(\mathbf{r}' \mathbf{w} + r_h x_h)^2 &= E_{\eta} [E(\mathbf{r}' \mathbf{w} + r_h x_h^*)^2 | \eta] \\ &= E_{\eta} [\mathbf{w}' \Omega_r \mathbf{w} - 2\sigma_h^2 \beta \mathbf{w} (\beta - \eta)' \mathbf{w} + (\sigma_h (\beta - \eta)' \mathbf{w})^2] \\ &= \mathbf{w}' \Omega_r \mathbf{w} - (\sigma_h \beta' \mathbf{w})^2 + \sigma_h^2 \mathbf{w}' \Omega_{\eta} \mathbf{w} \end{aligned}$$

The variance of the hedged portfolio exceeds the unhedged variance when $\mathbf{w}' \Omega_{\eta} \mathbf{w} > (\beta' \mathbf{w})^2$. The left-hand side of the inequality is the squared estimation error of the portfolio’s beta. The right-hand side is the portfolio beta-related variance.

Between the non-hedged and the fully-hedged portfolio, maybe there is a hedging level that improves on both. We consider the case where we apply a positive *hedging shrinkage factor* y_h to the optimal hedging: $x_h = -y_h \hat{\beta}' \mathbf{w}$. We estimate the variance of $(\mathbf{r}' \mathbf{w} + x_h r_h)^2$, and then we minimize it with respect to y . The calculation is similar to the one we performed above:

$$E(\mathbf{r}' \mathbf{w} + r_h x_h)^2 = E_{\eta} [E(\mathbf{r}' \mathbf{w} + r_h x_h^*)^2 | \eta] = \mathbf{w}' \Omega_r \mathbf{w} - y_h (\beta' \mathbf{w})^2 \sigma_h^2 + y_h^2 \sigma_h^2 \mathbf{w}' \Omega_{\eta} \mathbf{w}$$

From which

$$y_h^* = 1 - \frac{\mathbf{w}' \Omega_{\eta} \mathbf{w}}{(\mathbf{w}' \hat{\beta})^2} \quad (11.6)$$

$$x_h^* = -\hat{\beta}' \mathbf{w} + \frac{\mathbf{w}' \Omega_{\eta} \mathbf{w}}{\mathbf{w}' \hat{\beta}} \quad (11.7)$$

Let’s sense-check this formula:

- The shrinkage factor y_h^* is independent of the units of the portfolio. If we measure the portfolio in cents or in dollar, we get the same value of y_h^* . Otherwise stated: if we hedge a portfolio 10 times the size of our current one, the fraction is unchanged, and the best hedge is 10 times the hedge of the original portfolio.
- If there are no estimation errors in the betas, then $\Omega_{\eta} = 0$ and $y_h^* = 1$: we use the optimal hedge ratio.

- The numerator is a weighted sum of the estimation errors. The larger the error, the smaller the shrinkage factor.
- The ratio in Equation (11.6) can be loosely interpreted as the square of the aggregate noise-to-signal ratio of the betas. The higher the ratio, the smaller the scaling factor.
- Consider the edge case where the true betas are all zero and errors are independent. Then $\hat{\beta} = \eta$ and the expected value³ of the denominator is $E[(\mathbf{w}'\hat{\beta})^2] = \mathbf{w}'\Omega_\eta\mathbf{w}$. In expectation, numerator and denominator are equal, and $y_h^* = 1$. On average, we do not hedge, which is the correct course of action.

In practice, we recommend the following steps:

1. Estimate the time-series $\hat{\beta}_i$ and its standard error τ_i . Define Ω_η as the diagonal matrix whose i th term is τ_i^2 .
2. Compute y_h^* using Formula (11.6).
3. Buy $x_h^* = -(y_h^*)^+ \times (\hat{\beta}' \mathbf{w})$ of the hedging instrument. The lower bound at zero is meant to avoid the situation where we hedge in the opposite direction.
4. (optional). It is difficult to estimate the correlations between estimation errors, especially in periods of market stress. You can simulate their impact by assuming constant correlations between them and then defining

$$\Omega_\eta = \begin{pmatrix} \tau_1^2 & 0 & \dots & \dots \\ 0 & \tau_2^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \tau_n^2 \end{pmatrix} \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} \tau_1^2 & 0 & \dots & \dots \\ 0 & \tau_2^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \tau_n^2 \end{pmatrix}$$

and testing the sensitivity for different values of ρ . The hedging ratio decreases linearly as ρ increases.

³Informally, if the number of assets is large, we should expect the variance of $(\mathbf{w}'\hat{\beta})^2$ to be small, so that the expected value is a good proxy for $(\mathbf{w}'\hat{\beta})^2$.

5. (simplifying Formulas (11.6, 11.7)). Assume that the terms w_i^2 are uncorrelated with τ_i^2 . Then⁴

$$\begin{aligned}\mathbf{w}'\boldsymbol{\Omega}_\eta\mathbf{w} &= \sum_i w_i^2 \tau_i^2 \\ &= \frac{1}{n} \sum_i w_i^2 \sum_i \tau_i^2 + n\text{cov}((w_1^2, \dots, w_n^2), (\tau_1^2, \dots, \tau_n^2)) \\ &= \|\mathbf{w}\|^2 \hat{E}(\boldsymbol{\tau}^2)\end{aligned}$$

An analogous simplification occurs for the denominator. Then the formula for the optimal hedge ratio becomes

$$y^\star = 1 - \frac{\hat{E}(\boldsymbol{\tau}^2) \|\mathbf{w}\|^2}{(\hat{\beta}' \mathbf{w})^2} \quad (11.8)$$

Higher standard errors τ_i imply greater shrinkage. Lower dollar exposure to the tradable factor also means greater shrinkage. Finally, to simplify things dramatically, consider the case where all $\hat{\beta}_i$ are identical, and the portfolio is long only. The shrinkage factor simplifies further:

$$y_h^\star = 1 - \underbrace{\hat{\beta}^{-2} \hat{E}(\boldsymbol{\tau}^2)}_{(squared\ noise-to-signal)} \times \underbrace{H(\mathbf{w})}_{(portfolio\ concentration)}$$

The ratio $H(\mathbf{w}) := \|\mathbf{w}\|_2^2 / \|\mathbf{w}\|_1^2$ is a measure of portfolio concentration⁵. A portfolio that has maximum diversification has n positions with identical NMV, and has $H(\mathbf{w}) = 1/n$, while a maximally concentrated portfolio has all NMV concentrated in a single stock, so that $H(\mathbf{w}) = 1$. The interpretation here is that the shrinkage factor is small when the portfolio is more concentrated. The intuition is that the estimation error of the beta averages out more in diversified portfolios.

11.4 Factor-Mimicking Portfolios of Time Series

A problem related to hedging a portfolio using a tradable security is that of trading a portfolio that is close to a *non-tradable* security. Such time series

⁴For a vector \mathbf{x} , define $\hat{E}(\mathbf{x})$ the average of the values x_1, \dots, x_n .

⁵The Herfindahl Index is usually defined for a set of n nonnegative numbers x_i that sum to one: $H := \sum_i x_i^2$. It can be extended to arbitrary sets of numbers y_i , by defining $x_i := |y_i| / \sum_j |y_j|$ and applying the original definition..

abound in real life. A quantitative portfolio manager may be interested in trading them for a few reasons. First, the time series may show high correlation to the securities in her investment universe; and therefore the time series could serve as a useful hedging instrument. Another use case is that of the macroeconomic systematic investor⁶ who has some well-informed reason to trade a theme. Developing a tradable portfolio that “tracks” the time has real value for her. Lastly, just verifying how well can we track a time series is interesting in itself. It shows us whether the time series is of concrete use. Only that that can be traded exists. The occasional analysts that hawk non-tradable themes are full of sound and fury, usually signifying sell-side research fees.

We introduce the ingredients for our problem earlier in the chapters. We have n assets with returns r_i , and a time-series with return r_h ; we keep the original subscript, since subscripts should not be multiplied beyond necessity. In its simplest form, we have the following ingredients:

- We have two periods and one realized return. Investment decisions are made in period one, profits are realized in period two.
- There are n assets with returns r_i , with covariance matrix Ω_r .
- There are n loadings $\hat{\beta}_i = \beta_i + \eta_i$ where η_i is the estimation error of β_i , with $E(\eta_i^2) = \tau_i^2$. We denote $\Gamma := \text{diag}(\tau_1^2, \dots, \tau_n^2)$.

The problem asks to minimize the tracking error between the time series and a portfolio: $\min_{\mathbf{w}} E[(\mathbf{r}'\mathbf{w} - r_h)^2]$. We condition on $\boldsymbol{\eta}$, as we did earlier in the chapter.

$$\begin{aligned} E[(\mathbf{r}'\mathbf{w} - r_h)^2] &= E_{\boldsymbol{\eta}} [E[(\mathbf{r}'\mathbf{w} - r_h)]^2 + \text{var}(\mathbf{r}'\mathbf{w} - r_h)|\boldsymbol{\eta}]] \\ &= E_{\boldsymbol{\eta}} [E[((\boldsymbol{\beta} - \boldsymbol{\eta})\mu_h + \boldsymbol{\epsilon})'\mathbf{w} - r_h]^2] \\ &\quad + \mathbf{w}'\Omega_r\mathbf{w} - 2\sigma_h^2(\boldsymbol{\beta} - \boldsymbol{\eta})'\mathbf{w} + \sigma_h^2 \\ &= \mu_h^2(\boldsymbol{\beta}'\mathbf{w} - 1)^2 + \mu_h^2\mathbf{w}'\Gamma\mathbf{w} + \mathbf{w}'\Omega_r\mathbf{w} - 2\sigma_h^2\boldsymbol{\beta}'\mathbf{w} + \sigma_h^2 \\ &= \mathbf{w}'(\Omega_r + \mu_h^2\Gamma + \mu_h^2\boldsymbol{\beta}\boldsymbol{\beta}')\mathbf{w} - 2(\mu_h^2 + \sigma_h^2)\boldsymbol{\beta}'\mathbf{w} + \mu_h^2 + \sigma_h^2 \end{aligned}$$

And the first-order condition on this unconstrained problem gives the optimal portfolio, which we transform by means of the Woodbury-Sherman-Morrison

⁶A species who I have ignored in this book, because I have not been lucky enough to meet its members in the wild.

lemma of the inverse matrix; see Equation (14.36):

$$\begin{aligned}\mathbf{w}^* &= (\mu_h^2 + \sigma_h^2)(\boldsymbol{\Omega}_r + \mu_h^2 \boldsymbol{\Gamma} + \mu_h^2 \boldsymbol{\beta} \boldsymbol{\beta}')^{-1} \boldsymbol{\beta} \\ &= (\mu_h^2 + \sigma_h^2) \left(1 + \frac{\boldsymbol{\beta}'(\boldsymbol{\Omega}_r + \mu_h^2 \boldsymbol{\Gamma})^{-1} \boldsymbol{\beta}}{\mu_h^{-2} + \boldsymbol{\beta}'(\boldsymbol{\Omega}_r + \mu_h^2 \boldsymbol{\Gamma})^{-1} \boldsymbol{\beta}} \right) (\boldsymbol{\Omega}_r + \mu_h^2 \boldsymbol{\Gamma})^{-1} \boldsymbol{\beta}\end{aligned}$$

Having done most of the heavy lifting, we close with a few remarks:

- The beta estimation error $\boldsymbol{\Gamma}$ serves as a regularizer for the covariance matrix. The larger the expected returns, the higher the importance of the regularization term.
- When $\boldsymbol{\Gamma} = 0$ (no estimation error), and $\mu_h = 0$ (zero return), the optimal portfolio is $\mathbf{w}^* = \sigma_h^2 \boldsymbol{\Omega}_r^{-1} \boldsymbol{\beta}$, which is, up to a scaling factor, the minimum-variance portfolio. A minor point: it seems that the scaling factor is σ_h^2 , which would make no sense. The covariance matrix does contain σ_h , though, so that dependency is effectively linear.
- When $|\mu_h| \rightarrow \infty$, the optimal portfolio approaches, up to a constant, $\boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}$.
- Once we have the optimal portfolio \mathbf{w}^* , hedging is straightforward, in the sense that we can employ Equation (11.4) to reduce the core portfolio's risk.

Exercise 11.4. 15 *Describe how you would hedge to a time-series factor (or a FMP of a time series) on top of equity FMPs for a preexisting model (Hint: orthogonalization).*

11.5 ★Appendix

Proof of Exercise 11.1. We replace the decision variable $\mathbf{w}_h = \mathbf{P}\mathbf{x}$. From the definition of \mathbf{P} , it follows that $\mathbf{B}'\mathbf{P}\mathbf{x} = \mathbf{x}$, and $\mathbf{x}'\mathbf{P}'\boldsymbol{\Omega}_\epsilon\mathbf{P}\mathbf{x} = \mathbf{x}'(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x}$.

It follows that the optimization problem (11.5) can be rewritten

$$\begin{aligned}& \max \boldsymbol{\alpha}'_\perp (\mathbf{w}_c + \mathbf{w}_h) - \frac{1}{2} \rho (\sigma_f^2 + \sigma_i^2) - f(\mathbf{w}_h - \mathbf{w}_{h,0}) \\ \text{s.t. } & \mathbf{b} = \mathbf{b}_c + \mathbf{x} \\ & \sigma_f^2 = \mathbf{x}'\boldsymbol{\Omega}_f\mathbf{x} \\ & \sigma_i^2 = \sigma_{\epsilon,c}^2 + \mathbf{x}'(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x} + \mathbf{b}_c'(\mathbf{B}'\boldsymbol{\Omega}_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x} + \mathbf{w}_c'\boldsymbol{\Omega}_\epsilon\mathbf{y} \\ & \mathbf{w}_h \in \mathbb{R}^n\end{aligned}$$

Assume that $\mathbf{y} = 0$, $\boldsymbol{\mu} = 0$ and transaction costs equal to 0. The objective function becomes

$$\begin{aligned}\alpha'_\perp \mathbf{w}_c - \frac{1}{2}\rho(\mathbf{x}'\Omega_f\mathbf{x} + \sigma_{\epsilon,c}^2 + \mathbf{x}'(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x} + \mathbf{b}'_c(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x}) \\ \equiv -\frac{1}{2}\rho[\mathbf{x}'(\Omega_f + (\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1})\mathbf{x} + \mathbf{b}'_c(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{x}]\end{aligned}$$

which is minimized at

$$\begin{aligned}\mathbf{x}^* &= -(\Omega_f + (\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1})^{-1}(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}\mathbf{b}_c \\ &= -[\mathbf{I} + (\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})\Omega_f]^{-1}\mathbf{b}_c\end{aligned}$$

□

Part III

After the Trade

Chapter 12

Dynamic Risk Allocation

12.1 The Kelly Criterion

So far we have focused exclusively on single-period portfolio optimization. This may be appropriate for one-off investment decisions, but is inadequate for long-term investment strategies. There is a rich academic literature on inter-temporal choice theory, which aims at modeling the interplay between consumption and investment in the long run, both at the level of the individual consumer and at the aggregate level. Much of the literature has been ignored by asset managers for their investment decisions, for reasons we conjecture below. First, these models require the specification of a principled utility function and of an intertemporal tradeoff (in the form of a discount factor for future utility), something that no investment manager would or could specify. If quadratic utility had been the main justification of mean-variance optimization, it would probably have never been adopted. Secondly, these models don't capture well the institutional setting of asset managers. "Consumption" for asset managers corresponds to outflows, and asset managers don't receive any utility from it. Moreover, outflows do not bear a direct relationship to the principals' utilities (i.e., those who provide the investment capital to the managers). The reason for this is that inflows and outflows occur at low rates (they are "sticky"), due to inertia of the principals (who resist changing their asset allocations) and to contractual obligations with the asset managers (who require long advance notice for capital withdrawal).

One line of research in multi-period investing has been relevant to practitioners. It takes various names: Kelly gambling, Optimal Growth Portfolios, or Universal Portfolios. To introduce the concept, we consider first a very simple example. You have one risky asset in which to invest, which returns 100% or 50% with equal probability. The single-period expected return of the asset is $5/4$, and its volatility is $3/\sqrt{8}$. You have to decide how to invest your initial

capital in this asset. Consider two alternatives:

1. (*Constant Capital Allocation*) Every day you allocate the same amount of capital to the risky asset. This approach is consistent with solving a mean-variance optimization problem in each period. The problem faced in every period is

$$\max_w \frac{5}{4}w - \frac{\lambda}{2} \frac{9}{16}w^2 \Rightarrow w^* = \frac{20}{9\lambda} \quad (12.1)$$

where w is the net amount allocated to the risky asset and is independent of the period.

2. (*Static Allocation*) On day 0, you allocate a fraction x of your capital to the risky asset, and then you let it run. This is consistent with solving a mean-variance optimization in period 0, and letting it run.
3. (*Dynamic Allocation*) Every day, you allocate a fraction x of your capital *on that day* to the risky asset. We have no motivation for this (yet). The intuition however is it seems reasonable to have a volatility proportional to the available capital in each period. The ratio of the strategy's volatility to capital in each period is equal to $3x/4$ and is indeed constant in this approach.

The chart below shows the cumulative returns under the three approaches. The constant capital allocation shows low growth. Independently of x , the static allocation has poor performance, even though the risky asset has positive expected return. Conversely, the dynamic allocation exhibits a variety of behaviors. Its growth rate (equal to the slope of the curve) is not monotonically increasing with x . Risk-adjusted performance is good for low values of x , but the average returns are low. The most profitable strategy corresponds to $x = 1/2$. Higher values detract from performance.

What is remarkable is that, for each strategy, and for each period, the Sharpe Ratio is identical and equal to $5\sqrt{8}/3$, because in each period the portfolio, being the combination of a free-risk asset and a risky asset, is mean-variance efficient (see Ex. ??). This numerical example should warn about the subtleties of using the Sharpe Ratio as a performance measure. We have been able to compute the Sharpe Ratio exactly, thus abstracting away any complication due to performance measurement; and we obtained the same value for all the strategies in our example. Yet, the behavior of the cumulative returns differs wildly among the strategies! We can interpret this as follows: the single-period

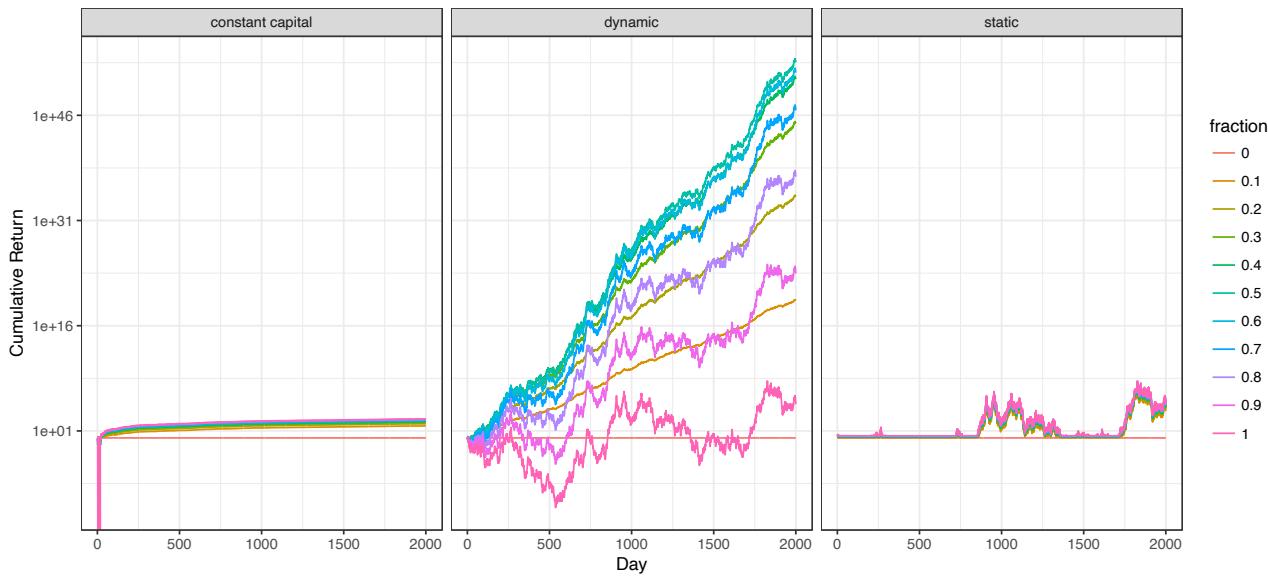


Figure 12.1: Cumulative returns under the dynamic and static policies. All the same curves are based on the same realization of returns of the risky asset. The returns are plotted on a logarithmic scale.

Sharpe Ratio, defined as expected mean return/standard deviation in a single period, is a measure of *investor skill*, but not of *strategy performance*. Averaging the Sharpe Ratio over the life of a strategy can give us a better estimate of skill, but is not telling us much about the risk-adjusted performance of the strategy over its lifetime. If we chose the cumulative returns over the strategy lifetime as a metric – and ignored any drawdown concern – then the dynamic strategy with $x^* = 1/2$ would be the clear favorite.

A second observation is that skill alone, defined as the ability to select a high-Sharpe portfolio in any given period, is necessary but not sufficient to be a successful investor. The size of the overall portfolio over time plays a major role in the long term. Yet, this topic does not receive much attention among academics nor practitioners.

To understand where the value $x = 1/2$ comes from, let $r_t(x)$ be the return of the dynamic allocation strategy. It is

$$r_t(x) = \begin{cases} 1 + x & p = 1/2 \\ 1 - x/2 & p = 1/2 \end{cases} \quad (12.2)$$

The total return of the strategy is $\prod_{t=1}^T r_t(x)$. The average growth rate of the strategy $g(x)$ is such that $\prod_{t=1}^T r_t(x) = \exp(Tg(x))$, or $g(x) = T^{-1} \sum_t \log(r_t(x))$. If we wanted to maximize the expected growth rate of the strategy, we would

solve the problem

$$\max_x \frac{1}{T} \sum_{t=1}^T E[\log(r_t(x))] = \max_x E[\log(r_1(x))] \quad (12.3)$$

The objective function is maximized when the investment fraction x is equal to 1/2.

If we were to maximize the total return of the strategy, we would have solved $\max_x E[\prod_t r_t(x)]$; since returns are iid, this would have been the same as solving $\max_x E[r_1(x)]$, whose optimal point is 1.

From Fig. 12.1, it appears that this strategy performs decidedly worse than other strategies with lower investment fractions. By simulation, one can show that strategy in which $x > 1$ performs even worse; this corresponds to borrowing money to invest in the risky asset. Summing up, it appears that long-term returns are maximized not by maximizing the expected returns, but by maximizing the expected growth rate, which is mathematically equivalent to maximizing an expected utility, with a logarithmic utility function. These strategies go under different names: Kelly strategies, optimal growth strategies, log-optimal strategies, and universal strategies.

Example 12.1 (The Kelly allocation to a single security). *Let's work out in detail an important example. We have only two assets: a risk-free asset and a risky asset. One way to interpret this asset in real-world application is as a portfolio manager to which we need to allocate capital. If we were to maximize the expected growth of the portfolio, then we would solve the problem*

$$\max_x E[\log(W_0(1-x)(1+r_f) + W_0x(1+r+r_f))] \quad (12.4)$$

$$\equiv \max_x E[\log(1+rx)] \quad (12.5)$$

In addition to an exact numerical solution, we also produce an approximate solution based on the quadratic approximation of the logarithm:

$$\log(1+x) = x - \frac{x^2}{2} + o(x^2) \quad (12.6)$$

Then,

$$\max_x E[\log(1+rx)] \simeq \max_x \mu x - \frac{1}{2}(\sigma^2 + \mu^2)x^2 \quad (12.7)$$

$$\Rightarrow x^* = \frac{\mu}{\sigma^2 + \mu^2} \simeq \frac{SR}{\sigma} \quad (\text{assuming that } \mu \ll \sigma) \quad (12.8)$$

This approximate result is reliable when the typical fluctuations of x^* are smaller than 1. A heuristic is to require that the volatility of x^* be smaller than 1: $|x^*\sigma| \ll 1$, i.e., $|SR| \ll 1$. Assume that the risky asset has daily excess returns r that are lognormally distributed with $\log r \sim N(\nu, \tau)$. In Table 12.1 we plot the result $\tau = 0.01$.

ν	$E(r)$	$\text{stdev}(r)$	Sharpe	x^*	x_{approx}^*	Error (%)
0.0001	0.0002	0.010	0.01	1.6	1.5	-3.6
0.0002	0.0003	0.010	0.02	2.5	2.5	-2.0
0.0003	0.0004	0.010	0.03	3.5	3.5	-0.2
0.0004	0.0005	0.010	0.04	4.5	4.5	0.0
0.0005	0.0006	0.010	0.05	5.4	5.5	1.4
0.0006	0.0007	0.010	0.06	6.5	6.5	-0.5
0.0007	0.0008	0.010	0.07	7.4	7.5	0.7
0.0008	0.0009	0.010	0.08	8.5	8.5	0.1
0.0009	0.0010	0.010	0.09	9.3	9.5	1.5
0.0010	0.0011	0.010	0.10	10.0	10.5	4.8

Table 12.1: Optimal allocation for a single bet. $\log r \sim N(\nu, \tau)$ with $\tau = 0.01$, $\nu \in [1E - 4, 1E - 3]$. The error is defined as $(x_{\text{approx}}^* - x^*)/x^*$.

The approximation is quite good for daily Sharpe Ratios up to 0.10, corresponding to an annualized Sharpe of 1.59. Equation 12.8 is also useful to estimate the ratio of the dollar volatility of the risky strategy to the available capital. The dollar volatility is $W_0 x^* \sigma$, and the ratio takes the simple form

$$\frac{(\text{dollar volatility})}{(\text{capital})} = \frac{W_0 x^* \sigma}{W_0} = SR \quad (12.9)$$

This means that if we had a \$1B of capital and an annualized Sharpe of 2, we should have a dollar volatility of \$2B, and we should leverage our strategy accordingly to reach this volatility target.

Example 12.2 (Returns of the Kelly allocation to a single security). Since $(\text{expected PnL}) = (\text{dollar volatility}) \times SR$, the expected return under optimal growth allocation is

$$(\text{expected return}) = \frac{(\text{dollar volatility}) \times SR}{(\text{capital})} = SR^2 \quad (12.10)$$

For a security with an annualized Sharpe Ratio equal to 0.5, the expected return is 25%. [*** CHECK COMPOUNDING ***]

Example 12.3 (The Kelly allocation to the US market). We can specialize the analysis above to the important case in which the risky asset is the US market benchmark. This asset is available to retail investors in the form of low-management fees mutual funds and ETFs, both of which track the US market accurately. Futures for the US markets are also available to sophisticated investors. In Fig. 12.4 we show the time series of the cumulative returns as a function of the fraction x invested in the stock market, and the final cumulative returns as a function of x . Based on the observed realization of the historical returns, the total returns are maximized at $x = 2.2$. This corresponds to an annual growth of the portfolio of 9.7%. Over the same period (February 1926–March 2018) the realized annualized Sharpe Ratio was 0.44, and the realized annualized volatility was 18.3%. From these data, the approximate optimal fraction is $x^* = 2.4$. This is in quite close agreement with the 2.2 computed ex post on empirical data. If we had chosen this fraction, the portfolio would have appreciated at an annualized rate of 9.6%. The annualized return of the market benchmark over the same period was 6.5%. This example suggests that, if our goal was to maximize our long-term returns and the interest rate paid to our lender of choice to borrow money was zero, then it would be optimal to leverage our capital. In practice, this is not the case, and the cost and risks of borrowing may reduce or even nullify the benefits of higher growth rate. In addition, the investor may face institutional constraints. However, this example illustrates of the invested fraction can have a very dramatic impact on capital appreciation.

Exercise 12.1 (Continuous approximation). (30) The goal of this exercise is to approximate this discrete return process with a continuous one.

1. Consider a risky asset with total return y taking two possible values: $e^u > 1$ and e^{-u} , with $u > d$ and $p_u + p_d = 1$. If $E \log y = \nu$ and $\text{var}(y) = \tau^2$, show that $u = \nu + \tau$ and $d = \nu - \tau$.
2. Partition the interval $[0, 1]$ in n consecutive intervals of length $1/n$. Approximate r_t with the product $r := \prod_{k=1}^n y_k$, where y_k are independent and distributed as in point 1 with $\mu := 1/n$, $\tau^2 = 1/n$.

$$\log(r) = \sum_{k=1}^n \log(y_k) = \sum_{k=1}^n (y_k - 1) + \frac{1}{2}(y_k - 1)^2 \quad (12.11)$$

Show that as $n \rightarrow \infty$,

$$\sqrt{n} \left[\frac{1}{n\tau} \left(\sum_{k=1}^n (y_k - 1) \right) - (\mu - 1) \right] \sim N(0, 1) \quad (12.12)$$

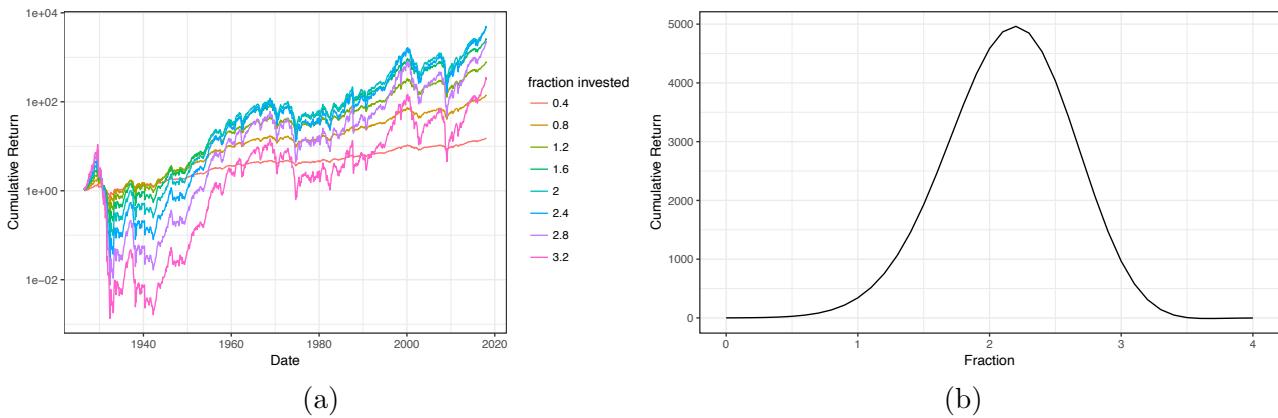


Figure 12.2: (a) Time series of cumulative returns for different fractions of the capital invested in the US market benchmark (cap-weighted average of NYSE, AMEX and NASDAQ-listed companies). Monthly excess returns of the benchmark for the period February 1926–March 2018 are from Ken French’s data library site. (b) Cumulative returns as a function of the fraction invested in the US market benchmark. The optimal fraction is 2.2.

Assume that the return r_t is lognormal with parameters (μ, σ) , so that $\log(r_t) \sim N(\mu, \sigma)$.

The next subsection is devoted to describing the attractive mathematical properties of Optimal Growth Strategies. We close the section and the chapter with a description of implementation of these strategies in real-world portfolios.

12.1.1 Kelly Portfolios: Mathematical properties

We limit our attention to the case in which we can choose in each period among a set of strategies Θ , and the associated returns $r_t(\theta)$ are independent of $r_{t'}(\theta)$, for all $t' < t$. These results were proved first by Breiman [Breiman \(1961\)](#) and Dubins and Savage [Dubins and Savage \(1965\)](#), but some of them have also been established for dependent random variables; see [Algoet and Cover \(1988\)](#).

Theorem 12.1. *Let X_t, Y_t the cumulative returns of the log-optimal strategy and alternative strategy with lower expected growth rate respectively.*

1. *The log-optimal strategy grows faster than any strategy with lower expected growth rate with probability 1. Let X_t, Y_t the cumulative returns of the log-optimal strategy and alternative strategy respectively. Then, with probability*

1,

$$\lim_{t \rightarrow \infty} \frac{X_t}{Y_t} = \infty \quad a.s. \quad (12.13)$$

2. Let $g := E[\log r_1]$ and X_t the associated cumulative return process. Then, with probability 1
 - a) $g > 0 \Rightarrow X_t \rightarrow \infty$
 - b) $g < 0 \Rightarrow X_t \rightarrow -\infty$
 - c) $g = 0 \Rightarrow \limsup_t X_t = \infty, \liminf_t X_t = -\infty.$
3. The expected time to reach capital level C is equal to $\log C/g$ in the limit $C \rightarrow \infty$, hence it is shortest for the log-optimal strategy.

Proof. We only sketch the proofs; detailed proofs are in Breiman (1961).

1. Define $z_t := \log r_t^X - \log r_t^Y$. $\mu := Ez_t = g^X - g^Y > 0$, and $Ez_t^2 < E(\log r_t^X)^2 - E(\log r_t^Y)^2 < \infty$. Hence, by the Strong Law of Large Numbers $t^{-1} \sum_{s=1}^t z_s \rightarrow \mu$ a.s., so that $\sum_{s=1}^t z_s \rightarrow \infty$ a.s..

$$\log(X_t/Y_t) = \sum_{s=1}^t z_s \rightarrow \infty \quad a.s \quad (12.14)$$

Hence $X_t/Y_t \rightarrow \infty$ a.s.

2. These statements follow from the fact that $\sum_{s=1}^t \log r_s^X \rightarrow \infty (-\infty)$ a.s., hence their exponents converge to $\infty (0)$ a.s. The case of $g = 0$ follows from the properties of zero-drift random walks.
3. Let $x_k := \log r_k^X$. Then, by Donsker's theorem, in the limit $T \rightarrow \infty$ Billingsley (1999)

$$\frac{1}{\sqrt{T}} \left(\frac{1}{T} \sum_{k=1}^{\lfloor tT \rfloor} x_k - gt \right) \Rightarrow \sigma B_t \quad t \in [0, 1] \quad (12.15)$$

or equivalently Glynn (1990)

$$\sum_{k=1}^{\lfloor tT \rfloor} x_k \Rightarrow gtT + \sqrt{T}\sigma B_t \quad t \in [0, 1] \quad (12.16)$$

The time to reach capital C_T is defined as

$$\tau := \inf\{t : X_t \geq C_T\} \quad (12.17)$$

$$\tau = \inf \left\{ s : \sum_{k=1}^s x_k \geq \log C_T \right\} \quad (12.18)$$

$$E\tau \rightarrow E \left(\inf \left\{ s : gtT + \sqrt{T}\sigma B_t \geq \log C_T \right\} \right) \quad (12.19)$$

The value on the right is the expected first hitting time of a brownian motion with drift, which is equal to $E\tau = \log C_T/g$.

□

What the theorem says is that a Kelly strategy has many very desirable features. It beats any other strategy that has a different expected growth rate with probability 1. It also reaches a certain cumulative return faster than any other strategy; and the approximate time needed to reach this return can be expressed as a function of g . And it minimizes the probability of a drawdown. Finally, a positive expected growth rate is a necessary and sufficient condition for any strategy to have a growing cumulative return over time.

What the theorem *doesn't say* is that a Kelly Strategy is maximizing the Sharpe Ratio, even if we were able to compute it exactly from knowledge of the true expected return and volatility of the strategy. Nor does it guarantee any lower bound on the maximum drawdowns, which can be severe, as seen in the simulations above. The scale of the y axis is logarithmic. The excursion of the returns is therefore proportional to the drawdown percentage. For example, in example 12.3, the fraction x invested increases from 0 to the growth-maximizing level x^* , both the growth rate and the size of the drawdowns increase with x . Above the optimal level, the growth rate diminishes (as expected) and the drawdowns further increase. For fractions of the invested wealth lower than x^* , there is a trade-off between expected log returns and volatility of the log returns. MacLean, Ziemba and Blazenko [MacLean et al. \(1992\)](#) explore trade-offs between growth and security when the investment strategy is a linear combination of the growth-maximizing one and the safety-maximizing one, i.e., we invest a fixed percentage of our capital in the Kelly strategy, and the remaining in a risk-free asset. This means that we still invest in the risky asset a fixed but lower proportion than the Kelly strategy. In the literature, this is called a *fractional*

Kelly strategy. In their original paper and in successive research, MacLean and co-authors [MacLean et al. \(1992, 2004, 2010\)](#) show that the fractional Kelly strategy does indeed trade off growth for security when we choose reasonable measures for each. The examples in this section already contain instances of fractional Kelly, and the trade-offs there are visible. Does it mean that fractional Kelly is always efficient with respect to criteria of growth vs security? The answer is, in general, negative; in this respect the strategy is just a heuristic. However, there is a regime in which in the subsection below we will make the trade-off precise in the limit of “small bets compared to wealth” and show the optimality of fractional Kelly.

12.2 Log-Return Mean-Variance Optimization

If we quantify the *magnitude* of a drawdown as the second moment of the log-returns and the average growth rate as the expected value of the log-return, then a measure of risk adjusted performance for a strategy over time may be taken as the Sharpe Ratio of the log-returns¹. The formulation of the problem is

$$\max E(\log((W - e'w)(1 + r_f) + \mathbf{w}'(1 + r + r_f))) \quad (12.20)$$

$$\text{s.t.} \text{var}(\log((W - e'w)(1 + r_f) + \mathbf{w}'(1 + r + r_f))) \quad (12.21)$$

In applications, we replace the variance with the second moment, as we did in previous chapters. After rearranging the terms, we obtain

$$\max E(\log(W(1 + r_f) + \mathbf{w}'r)) \quad (12.22)$$

$$\text{s.t.} \text{var}(\log(W(1 + r_f) + \mathbf{w}'r)) \quad (12.23)$$

The penalized formulation is

$$\max E(\log(W(1 + r_f) + \mathbf{w}'r) - \frac{\lambda}{2} \text{var}(\log(W(1 + r_f) + \mathbf{w}'r))) \quad (12.24)$$

We have isolated the risk-free asset in the optimization in order to approximate the logarithmic terms with their second-order polynomial expansion. This will yield two benefits. First, the problem will become numerically more tractable. Second, the proximal problem will yield additional insight in the nature of the problem. Write

$$\log(W(1 + r_f) + \mathbf{w}'r) = \log(W(1 + r_f)) + \log\left(1 + \frac{\mathbf{w}'r}{W(1 + r_f)}\right) \quad (12.25)$$

¹This approach has been recently advocated by Luenberger in [Luenberger \(1993, 2013\)](#). [Williams \(1936\)](#) seems the first one to have proposed it.

We assume that

$$\text{stdev}(\mathbf{w}'r) \ll W(1 + r_f) \quad (12.26)$$

Define $\kappa := \log(W(1 + r_f))$. Under this assumption, the penalized formulation becomes

$$\max E \left(\kappa - \frac{\kappa^2 \lambda}{2} + \frac{(1 - \kappa \lambda)}{e^\kappa} \mathbf{w}'r + \frac{1}{2} \frac{(\kappa - 1)\lambda - 1}{e^{2\kappa}} (\mathbf{w}'r)^2 \right) \quad (12.27)$$

$$\equiv \max E \left(\mathbf{w}'r - \frac{e^{-\kappa}}{2} \left(1 + \frac{\lambda}{1 - \kappa \lambda} \right) (\mathbf{w}'r)^2 \right) \quad (12.28)$$

And, after defining as usual $\alpha := Er$ and $\Omega_r := E(rr')$, we get

$$\max \alpha' \mathbf{w} - \frac{e^{-\kappa}}{2} \left(1 + \frac{\lambda}{1 - \kappa \lambda} \right) \mathbf{w}' \Omega_r \mathbf{w} \quad (12.29)$$

The risk tolerance $\rho(\lambda)$ is

$$\rho(\lambda) = \frac{1}{W(1 + r_f)} \left(1 + \frac{\lambda}{1 - \kappa \lambda} \right) \quad (12.30)$$

and it non-zero even when $\lambda = 0$, i.e., when there is no constraint on the variance of the log-return. The risk tolerance is, to a first approximation, inversely proportional to the wealth. The volatility-to-capital ratio of the optimal portfolio is

$$\frac{\text{stdev}(\mathbf{w}'r)}{W(1 + r_f)} = \frac{\sqrt{\alpha' \Omega_r \alpha}}{\sqrt{W(1 + r_f)(1 + \lambda/(1 - \kappa \lambda))}} \quad (12.31)$$

The portfolio in the absence of constraints on log-return variance is

$$\tilde{\mathbf{w}} = W(1 + r_f) \Omega_r^{-1} \alpha \quad (12.32)$$

and, for a binding constraint,

$$w(\lambda) = \left(1 + \frac{\lambda}{1 - \kappa \lambda} \right)^{-1} \tilde{w} \quad (12.33)$$

so that the constrained problem can be interpreted as a justification for fractional Kelly. A constraint on log-return volatility results in an optimal portfolio with the same relative weights as Kelly, but with lower gross market value reduced by a factor $\rho(0)/\rho(\lambda)$.

12.3 Fractional Kelly and Drawdown Control

In an influential paper, Grossman and Zhou [Grossman and Zhou \(1993\)](#) address a question related to that of identifying a growth-optimal strategy and of constrained growth-optimal optimization. In the optimization problem, the constraint was on the volatility of log returns, and implicitly on the probability of drawdown; in the Grossman-Zhou formulation, the investor wants to maximize the long-term growth and with probability one avoid reaching a drawdown threshold. As formulated in their original paper, the model only consider a risk-free asset and a risky one with mean μ and volatility σ . In order to formulate the policy we define the *high watermark* as $M_t = \max\{W_s : t \in [0, t]\}$. Let d_t the current percentage drawdown from the high watermark $d_t = 1 - W_t/M_t$. Let the maximum allowed drawdown be D . The optimal policy gives the optimal fraction invested in the risky asset and is given by

$$f_t = \frac{\alpha}{\sigma^2} \left(1 - \frac{1 - D}{1 - d_t} \right) \quad (12.34)$$

This policy is elegant and intuitive. For some intuition, fix first, $D = 1$; i.e., we can tolerate infinite drawdown. Then the strategy is the one we identified in Eq.12.8: invest a fixed fraction $x^* = \mu/\sigma^2$. If $0 < D < 1$, then the optimal strategy is to invest a fraction x^*D when we are at the high watermark $d_t = 0$. This means that we are more prudent than in the simple Kelly scenario, and we are more prudent if our threshold is conservative. Moreover, we decrease the invested fraction as we approach the drawdown threshold, and we liquidate the risky asset Figure 12.3 shows the optimal fraction as a function of the threshold. The reduction rate is nearly constant over the range of allowed drawdowns. To understand the trade-offs between optimizing for variance control and optimizing for drawdown control, it is useful to compare the Grossman-Zhou and Fractional Kelly strategies in a numerical examples. Specifically, we consider the case of a risky asset with independent identically distributed returns. Its expected daily return is 0.08% and its daily volatility is 1%, corresponding to a Sharpe Ratio of 1.27. The two strategies are parametrized by the Kelly fraction and the drawdown threshold respectively, i.e.

$$f_t(p) = p \frac{\alpha}{\sigma^2} \quad (\text{Fractional Kelly}) \quad (12.35)$$

$$f_t(D) = \frac{\alpha}{\sigma^2} \left(1 - \frac{1 - D}{1 - d_t} \right) \quad (\text{Grossman-Zhou}) \quad (12.36)$$

with $p \in (0, 1)$, $D \in (0, 1)$. I then simulate the performance of the two strategies over a 100-year period (i.e., 25,200 days) and compare the realized volatility and

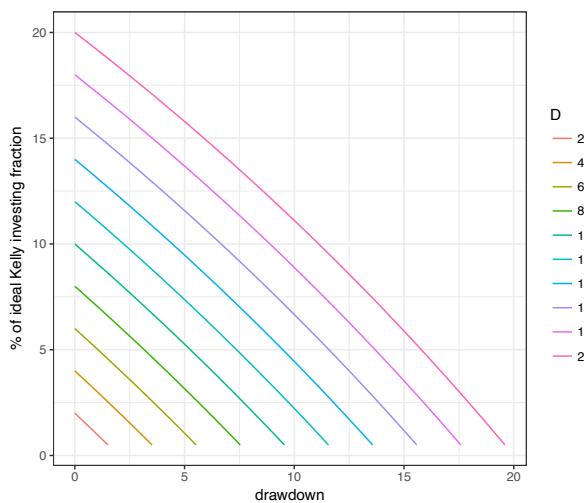


Figure 12.3: Reduction factor $1 - (1 - D)/(1 - d_t)$, for various values of D and d_t .

the maximum drawdown for strategies having the same expected log-return. Fig. 12.4 shows the results. As expected, the fractional Kelly strategy has a better profile than the Grossman-Zhou in the mean-volatility plane, and a worse one in the mean-maximum drawdown one. In this numerical example, the reduction in drawdown of Grossman-Zhou seems more marked than the associated increase in volatility. For example, consider a max tolerated drawdown of 30%. Grossman-Zhou achieves an average daily return of approximately 0.09%, while fractional Kelly achieves an average daily return of 0.075%, a 20% increase. More importantly, Grossman-Zhou controls the maximum drawdown *ex ante*, with probability one and independently of misspecification of the problem. In the mean-variance approach, we can only provide a probabilistic bound on the drawdown; moreover, if the parameters in the optimization problem are incorrect, this bound will be incorrect as well. These considerations suggest that the Grossman-Zhou strategy may be preferable. There are a few qualifications to this statements. First, we have ignored the role played by transaction costs. In Grossman-Zhou, the amount invested fluctuates heavily over time, since we may force a complete liquidation of the risky asset when we reach the threshold. This in turn may affect the profitability of the strategy and make the approach less attractive. It is beyond the scope of this chapter to extend the analysis to the case of transaction costs which, in the absence of analytical results, may only be tractable with numerical experiments. A second issue with the Grossman-Zhou strategy is that it does not give us the same modeling flexibility of an optimization formulation. This book makes the case that additional penalty terms and constraints are justified by concern with regulatory requirements,

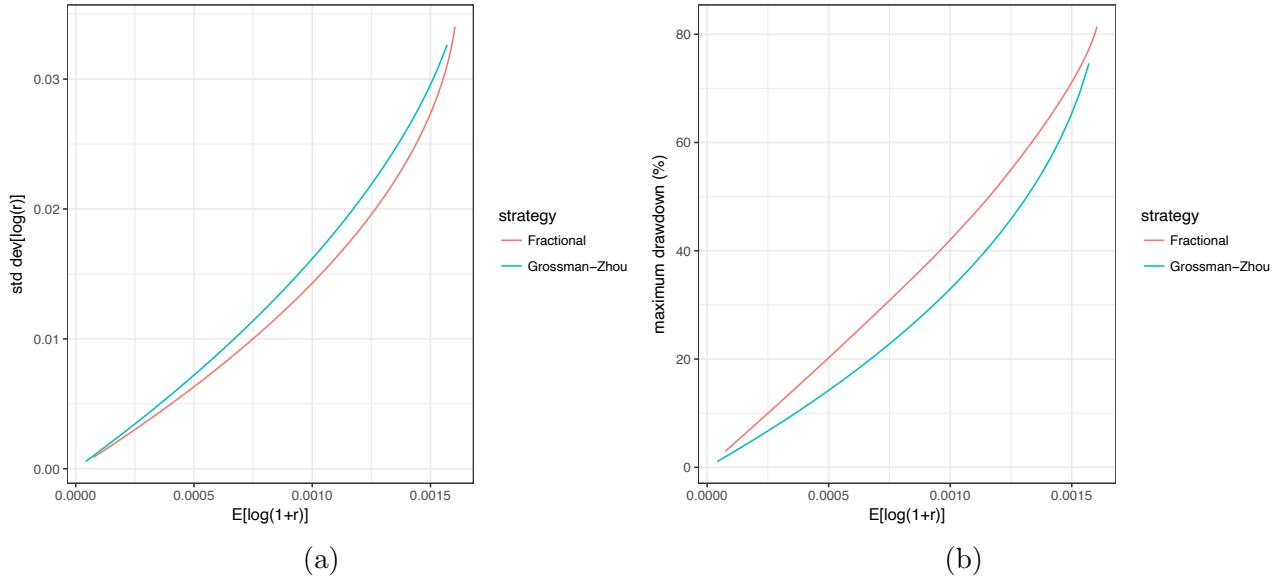


Figure 12.4: Comparison of fractional Kelly and Grossman-Zhou strategies. Both strategies' performance measures are estimated over the same sequence of 25,200 returns, but with different parameters p , D . (a) Standard Deviation of daily log-returns vs mean log-return. (b) Maximum drawdown.

fiduciary mandates, model mis-estimation and many other considerations; all of which fit naturally in the fractional Kelly framework, which can be derived as the outcome of an optimization. The same constraints and penalties in the Grossman-Zhou dynamic problem poses formidable challenges. These objections notwithstanding, Grossman-Zhou is a useful heuristic that can be used as an overlay to a Kelly-like strategy.

12.4 Variants of Fractional Kelly: Finite Horizon, Transaction Costs, and Heuristics

The previous models of investment opportunities are idealized. The investor has an infinite investment horizon, and the drawdown is computed from the high watermark. In practice, an investor may be interested in maximizing the growth rate over a finite time interval, say annual; in addition, investors in a fund are often sensitive to drawdowns compared to a reference date. We formulate the problem as a stochastic dynamic program. Although the problem does not admit a closed-form solution, it is possible to solve the equations numerically.

The policy recommends a fraction of the wealth invested in the risky asset. In the GZ policy, f depends on the high-watermark and the current wealth level. In the problem variant, f depends on the previous value of f , as well as the current level of wealth and on the epoch.

The initial wealth is w_0 , and the state of the portfolio at time t is given by the pair (wealth, fraction invested). In every period, the fraction invested changes from f to $f + a$. The traded amount is wa , and the associated transaction costs are a function $c(wa)$ of the traded amount. There is only one reward at the end of T periods: the logarithm of the wealth. If the wealth is below the threshold $w_0(1 - D)$, the investor receives a reward $q \ll w_0(1 - D)$. By recursion, if the investor has state (w, f) at time t , she chooses the fraction invested that gives the highest expected reward in period $t + 1$. The Hamilton-Jacobi-Bellman equations are

$$V_T(w, f) = \begin{cases} T^{-1} \log(w/w_0) & \text{if } w/w_0 \geq 1 - D \\ q & \text{otherwise} \end{cases} \quad (12.37)$$

$$V_t(w, f) = \begin{cases} \max_a E_r[V_{t+1}(w - c(wa))((f + a)r + 1), f + a] & \text{if } w/w_0 \geq 1 - D \\ q & \text{otherwise} \end{cases} \quad (12.38)$$

The expected growth rate is given by $V_0(w_0, 0)$, and in each epoch t , there is an optimal fraction invested $f^*(w_t, f_t, t) = f_t + a^*(w_t, f_t, t)$.² Visualizing the policy is more complicated, since the fractions depends on three parameters rather than one. We illustrate it for the case of a risky asset with identical features to the example above: expected daily return of 0.08% and daily volatility of 1%. We fix an investment horizon $T = 252$, corresponding to a trading calendar year for equities, set the maximum drawdown percentage to $D = 5\%$, and fix the epoch to $t = 126$. As a transaction cost function, we set $c(x) = 6e - 4 \times |x|^{3/2}$. Figure xxx shows the optimal fraction invested for various values of the drawdown d_t , and for various values of the invested fraction at time

²We have assumed that the strategy is non-stochastic, i.e., there is a unique optimal value a per stage and state. This is dictated by practical considerations. In principle, it is possible that the optimal strategy be stochastic, i.e., it may be optimal to sample a from a probability distribution. The optimality of a stochastic policy depends on the properties of the return r and of the transaction cost function c , and is beyond the scope of this treatment.

12.5 Further Reading*

On intertemporal choice theory, see Cochrane (2005). Basics: De Finetti (1940)Markowitz (1952)Markowitz (2014)Kolm et al. (2014)

Alpha decom: Black et al. (1972) Fama and MacBeth (1973)Lee and Stefek (2008)

Robustness: Maccheroni et al. (2013) Fabozzi et al. (2007) Kim et al. (2014b) Kim et al. (2014a) Lai et al. (2011) DeMiguel et al. (2009b) Schoettle and Werner (2009) Jagannathan and Ma (2003) Saxena and Stubbs (2013) Ceria et al. (2012) Bender et al. (2009) Huberman et al. (1987) Wang (2005)

Market Impact:Harris (2003) Hasbrouck (2007) Almgren et al. (2005) Bouchaud (2010) Almgren (2009) Huberman and Stanzl (2004)

Kelly:Breiman (1961), Ottusák and Vajda (2007),Thorp (2006), MacLean et al. (1992),MacLean et al. (2004),MacLean et al. (2010), Grossman and Zhou (1993), Cvitanić and Karatsas (1995), Luenberger (1993)

References on standard topics:

- Optimization theory: Boyd and Vandenberghe (2004); Bazaraa et al. (2006); Luenberger and Ye (2008); Luenberger (1969).
- Utility theory: the relationship, preferences, utility functions and decision making under uncertainty is covered inHuang and Litzenberger (1988); Kreps (1988); A. Mas-Colell and Green (1995). (Cochrane, 2005; Huang and Litzenberger, 1988) cover the case of quadratic utility as a sufficient condition for MVO.

Constraints: Jagannathan and Ma (2003), DeMiguel et al. (2009a)

Constraint in estimation (add to fund model): Kozak et al. (2020); Chinco et al. (2019)

Chapter 13

Ex Post Performance Attribution

“After the rain has fallen, we return/ To a plain sense of things”. So begins a famous poem¹ which describes well the spirit of this chapter. Out of metaphor, the “rain” is the realized performance of our strategy, and the plain sense of things is our ability to understand what happened after the fact, namely:

- Is our performance due to luck or skill?
- How did we make or lose money? What is the contribution of factor PnL and idiosyncratic PnL?
- In idiosyncratic space, what drove our PnL? Asset selection or sizing? The first is being on the right side of a bet; the second one is the ability to size appropriately asset bets that yield higher returns.
- How can we explain factor PnL concisely and insightfully, i.e., using only factors that are of interest to us?

Performance attribution offers numerous advantages. First, it provides the portfolio manager with a much-needed reality check. If she lost money, maybe she can explain the source of the loss, and identify countermeasures to apply going forward; sometimes the remedies are straightforward and contained in the output of the performance attribution itself. If she made money, maybe she did so as the result of unintended bets on factors that were not included in the strategy scope. “The first principle is that you must not fool yourself—and you are the easiest person to fool.” This statement, made by Richard Feynman in his 1974 CalTech commencement address, holds true for scientists and traders alike. Secondly, performance attribution empowers the *principal* to reward an agent appropriately. The principal may be the hedge fund manager and the

¹W. Stevens, “The Plain Sense of Things” in [Stevens \(1990\)](#).

agent the portfolio manager, or, descending one step down in the decision-making hierarchy, the principal may be the portfolio manager, and the agent may be the analyst who works in the portfolio manager's team. There are other benefits. A portfolio manager is bound to use a specific factor model for *ex ante* portfolio construction. No *ex post* limitation exists after the trade, though: she can look at her performance under the magnifying glass of different risk models. For example, a global risk model, sometimes unsuited for country-specific investing, could reveal cross-country exposures. We can use statistical models in addition to fundamental models.

Performance attribution is conceptually simple but it is not trivial. The remainder of this chapter is broadly organized in two parts. First, we introduce *time-series performance attribution*, and review the concept of *characteristics-based* (also known as *holdings-based*) dd

13.1 Performance Attribution: The Basics

Recall the short introduction to performance attribution in Section 3.5.1: the PnL can be decomposed into the sum of factor and idiosyncratic components. The performance decomposition *process* is slightly more involved. Trading time is not discrete, whereas performance attribution occurs in discrete time. To reconcile the two views, The time axis is partitioned in intervals delimited by epochs τ_i . Denote the PnL_i the PnL in interval $[\tau_{i-1}, \tau_i]$, and \mathbf{r}_i , \mathbf{f}_i , and $\boldsymbol{\epsilon}_i$ the total returns, factor returns and idiosyncratic returns respectively. Also, define, as we have done previously, $\mathbf{b}_i := \mathbf{B}'\mathbf{w}_i$. Then we can isolate the *trading PnL* with the decomposition:

$$\begin{aligned} \text{PnL} &= \sum_t (\text{PnL}_t - \mathbf{r}_t \mathbf{w}_t) + \mathbf{r}_t \mathbf{w}_t \\ &= \underbrace{\sum_t (\text{PnL}_t - \mathbf{r}_t \mathbf{w}_t)}_{\text{trading PnL}} + \underbrace{\sum_t \mathbf{b}'_t \mathbf{f}_t}_{\text{factor PnL}} + \underbrace{\sum_t \boldsymbol{\epsilon}'_t \mathbf{w}_t}_{\text{idiosyncratic PnL}} \\ &\quad \underbrace{\qquad\qquad\qquad}_{\text{position PnL}} \end{aligned}$$

The sum of factor and idiosyncratic PnL is sometimes referred to as *position PnL*. This is the PnL we would experience if we could instantaneously trade, with no transaction costs, so that the PnL is resulting from the application to the portfolio of the interval's total returns. To fix ideas on the interpretation of the trading PnL, it is helpful to consider the case of an idealized high-frequency

trader (HFT). Let the epochs be the close of trading days. The high-frequency trader ends the day flat²: $\mathbf{w}_t = 0$. The accounting PnL is zero, but the trading PnL is not. It originates from three terms: intraday alpha, i.e. “price discovery”; compensation for providing liquidity by submitting limit orders and receiving a fraction of the bid-ask spread; and costs incurred by taking liquidity by submitting market orders.

The factor PnL can be decomposed in separate time series for the contribution of each factor:

$$\text{Factor PnL} = \sum_{j=1}^m \left(\sum_{t=1}^T [\mathbf{b}_t]_j [\mathbf{f}_t]_j \right) \quad (13.1)$$

This could be the end of a simple story: take portfolio snapshots at each epoch, decompose PnL into three terms, and then dive into the contribution of individual factors and of individual securities to idiosyncratic PnL. Reality, however, is more complex. First, we need to unveil the illusion of certainty that comes with the simple decomposition of Equation (13.1).

13.2 Performance Attribution with Errors

13.2.1 Two Paradoxes

To motivate the importance of having a more nuanced view of factor-based performance attribution we introduce two paradoxical facts about performance attribution, both related to factor-mimicking portfolios:

- (*Factor-Mimicking Portfolios have idiosyncratic risk but not PnL.*) Each factor-mimicking portfolio \mathbf{v}_i has by necessity a non-zero idiosyncratic variance $\sigma_{\mathbf{v}_i}^2 := \mathbf{v}_i' \boldsymbol{\Omega}_\epsilon \mathbf{v}_i$. However, the factor-mimicking portfolio has no idiosyncratic PnL whatsoever. This can be seen intuitively by the fact that the return of the factor is the return of the portfolio itself. More rigorously, let \mathbf{P} be the matrix whose columns are the factor-mimicking portfolios, as defined in Equation (4.11). The

$$\mathbf{P}' \boldsymbol{\epsilon} = \mathbf{B} (\mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{B})^{-1} \mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} (\mathbf{I}_n - \mathbf{B} (\mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{B})^{-1} \mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1}) \mathbf{r} = 0$$

and therefore the idio PnL is null. This holds for *all* factor portfolios, including those have an idiosyncratic variance percentage close to 50%.

²Non-idealized HFTs do not necessarily close the day flat, but instead rebalance the book and/or partially hedge it.

and for all periods. This is especially concerning given that factor model performance is often evaluated on factor portfolios.

- (*Factor-Neutral Portfolios*) On the other side, consider a portfolio \mathbf{w} with no factor exposures, i.e., $\mathbf{B}'\mathbf{w} = 0$. Hence, its entire volatility is its idiosyncratic volatility $\sigma_{\mathbf{w}}^2 := \mathbf{w}'\Omega_{\epsilon}\mathbf{w}$. Now, consider the portfolio $\mathbf{w} + \lambda\mathbf{v}$, where \mathbf{v} is a factor-mimicking portfolio and $\lambda \in \mathbb{R}$. The idiosyncratic PnL of this portfolio is the same for any value of λ , since \mathbf{v} has no idio PnL. However, the idiosyncratic volatility of $\mathbf{w} + \lambda\mathbf{v}$ depends on λ , and is equal to $\sigma_{\mathbf{w}}^2 + \lambda^2\sigma_{\mathbf{v}}^2 + 2\lambda\mathbf{w}'\Omega_{\epsilon}\mathbf{v}$. Hence we have *exactly* the same sequence of residual PnL generated by a continuum of portfolios with possibly very different volatilities. We can make the realized idiosyncratic volatility of the portfolio arbitrarily different than the predicted idiovolt, thus greatly undermining the credibility of the model. How can this be?

One could object that in practice, factor portfolios do not have zero idiosyncratic PnL. This is due primarily to the nonstationarity of the process, so that factor portfolios as of time t are slightly stale when applied to time $t+1$. This criticism doesn't address the concerns exemplified by the paradoxes for two reasons. First, because even in the ideal case in which the model is stationary, and we have accurately estimated its parameters, we do have these paradoxes. Secondly, because the idiosyncratic PnL would be in any event much smaller than what would be compatible with the idiosyncratic volatility predicted by the model.

In the next three sections I present a possible solution to these paradoxes. The overall take-away in the analysis is that the returns of the factor-mimicking portfolios are *estimates* of the true factor returns from the model. Once we account rigorously for the estimation error, the factor PnL and idiosyncratic PnL can be characterized as random variables whose first and second moments can be obtained from model and portfolio data. The next section lays out some basic facts about model estimation; the last section derives the main formulas. We then give explanations for the paradoxes.

13.2.2 Estimating Attribution Errors

Let us rewrite the attribution equations, but paying attention to the fact that we are using factor and idiosyncratic return estimates $\hat{f}_t, \hat{\epsilon}_t$. We consider the case of a time-independent factor model, but Recall from Section 8.3.1 that the factor returns can be written as

$$\hat{f}_t = \mathbf{f}_t + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim N(0, (\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1})$$

Analogously, for the idiosyncratic returns, we have

$$\begin{aligned}\hat{\epsilon}_t &= \mathbf{r}_t - \mathbf{B}\hat{\mathbf{f}}_t \\ &= \epsilon_t - \mathbf{B}\boldsymbol{\eta}_t \quad \mathbf{B}\boldsymbol{\eta}_t \sim N(0, \mathbf{B}(\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1}\mathbf{B}')\end{aligned}$$

$$\begin{aligned}(estimated \ factor \ PnL)_t &= \mathbf{w}'_t \mathbf{B}\hat{\mathbf{f}}_t \\ &= (true \ factor \ PnL)_t + \mathbf{w}'_t \mathbf{B}\boldsymbol{\eta}_t \\(estimated \ idiosyncratic \ PnL)_t &= \mathbf{w}'_t \hat{\epsilon}_t \\ &= (true \ idiosyncratic \ PnL)_t - \mathbf{w}'_t \mathbf{B}\boldsymbol{\eta}_t \\ \mathbf{w}'_t \mathbf{B}\boldsymbol{\eta}_t &\sim N(0, \mathbf{b}'_t (\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1}\mathbf{b}_t)\end{aligned}$$

When we attribute the PnL over multiple periods, we have

$$\begin{aligned}(true \ factor \ PnL) &= (estimated \ factor \ PnL) - \sum_t \mathbf{w}'_t \mathbf{B}\boldsymbol{\eta}_t \\(true \ idiosyncratic \ PnL) &= (estimated \ idiosyncratic \ PnL) + \sum_t \mathbf{w}'_t \mathbf{B}\boldsymbol{\eta}_t\end{aligned}$$

Finally, this gives us two useful results. First, it provides confidence intervals around the attributed PnL:

$$\begin{aligned}(true \ factor \ PnL) &\sim N\left(\sum_t \mathbf{b}'_t \hat{\mathbf{f}}_t, \sum_t \mathbf{b}'_t (\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1}\mathbf{b}_t\right) \\(true \ idiosyncratic \ PnL) &\sim N\left(\sum_t \mathbf{w}'_t \hat{\epsilon}_t, \sum_t \mathbf{b}'_t (\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1}\mathbf{b}_t\right)\end{aligned}$$

If, for example, we observe a negative idiosyncratic PnL over a given time interval, we can determine whether \$0 falls inside the 95% confidence interval or not. The same applies to factor PnL. An additional result is that the time series of factor and idiosyncratic PnL are in general negatively correlated. Take the case of a constant portfolio, and constant factor exposures \mathbf{b} . The covariance between factor and idiosyncratic PnL is given by $-\mathbf{b}'(\mathbf{B}'\Omega_{\epsilon}^{-1}\mathbf{B})^{-1}\mathbf{b}$. This is sometimes observed in practice.

13.2.3 Paradox Resolution

We first discuss the paradoxes introduced in the first section, both analytically and numerical examples.

- (*Factor Portfolios*) Factor portfolio i has exposure vector $\mathbf{b}_i = (0, \dots, 0, 1, 0, \dots, 0)$, where the 1 is in the i th position, so $\|\mathbf{b}_i\| = 1$. Therefore

$$(true \ factor \ PnL) \sim N \left(\sum_{t=1}^T \hat{f}_{t,i}, T[(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}]_{i,i} \right)$$

$$(true \ idiosyncratic \ PnL) \sim N(0, T[(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}]_{i,i})$$

So the factor portfolio has a random zero-mean idiosyncratic PnL whose variance grows linearly in T .

- (*Factor-Neutral Portfolios*) Let \mathbf{w} a portfolio with no exposure to any factor, i.e., $\mathbf{B}'\mathbf{w} = 0$. The portfolio $\mathbf{w} + \lambda\mathbf{v}_i$ (where \mathbf{v}_i is the first factor-mimicking portfolio) has exposure $\mathbf{b}_i = (0, \dots, \lambda, \dots, 0)$. The factor and idiosyncratic PnL are:

$$(true \ factor \ PnL) \sim N \left(\lambda \sum_{t=1}^T \hat{f}_{t,i}, \lambda^2 T[(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}]_{i,i} \right)$$

$$IP \sim N \left(\sum_{t=1}^T \mathbf{w}'\hat{\epsilon}_t, \lambda^2 T[(\mathbf{B}'\Omega_\epsilon^{-1}\mathbf{B})^{-1}]_{i,i} \right)$$

The idiosyncratic PnL is no longer independent of the hedge $\lambda\mathbf{v}_i$. A greater hedge makes the idiosyncratic attribution more uncertain, and the uncertainty is linear in the hedge.

Summing up, the current factor-based attribution methodology universally assigns a numeric factor and idiosyncratic PnL to a strategy; these are deterministic functions of the portfolios over time, the stock returns, and additional available data, such as asset characteristics. Ignoring the estimation error of these attributions leads to inconsistencies. These inconsistencies are not edge cases. Attributing performance of factor portfolios and of hedged portfolios is central to the practice of risk management and to understanding the performance of a strategy. As a simple resolution to these paradoxes, we saw that, even if are employing the true factor model, the returns of the factor-mimicking portfolios are unbiased estimates of the actual factor returns, and that we can characterize the estimation error. Given this characterization, one can propagate its impact in the performance attribution process, and view the factor and idiosyncratic PnL as random variables for which we have the full distributions (under the assumption of normality of returns) and the confidence intervals.

Insight 13.1: Reporting Standard Errors for Attributions

When reporting factor-based performance attributions, always include (either graphically, or in tabular form), the standard errors of the factor and idiosyncratic PnL:

$$(true \text{ factor } PnL) \sim N\left(\sum_t \mathbf{b}'_t \hat{\mathbf{f}}_t, \sum_t \mathbf{b}'_t (\mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{B})^{-1} \mathbf{b}_t\right)$$

$$(true \text{ idiosyncratic } PnL) \sim N\left(\sum_t \mathbf{w}'_t \hat{\epsilon}_t, \sum_t \mathbf{b}'_t (\mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{B})^{-1} \mathbf{b}_t\right)$$

This will help the portfolio manager better understand the uncertainty associated with her attributed performance.

13.3 Maximal Performance Attribution

A different way to summarize the previous section is: do performance attribution, but use caution. The coming section admits a similarly concise summary: do performance attribution, but trying to reduce confusion. If had to attempt a parallel to real life, performance attribution is like falling in love: fundamentally *good*, but certainly dangerous, and potentially confusing. Where does the confusion come from? Consider the following scenario. A portfolio manager has positive momentum exposure and loses a large sum due to a negative momentum return. He then cuts momentum exposure to zero, as a defensive measure. The day after, the factor has a very large negative return. We ask: is the portfolios *expected* PnL equal to zero? The answer is no. Another way to state this fact is that the relationship between asset performance and factor returns is mediated by betas, not exposures. The beta of a portfolio to momentum is given by the covariance between portfolio returns and the factor's returns, divided by the factor's variance. In formulas³:

$$\beta(\mathbf{w}' \mathbf{r}, f_i) = \frac{\mathbf{b}' \boldsymbol{\Omega}_{\mathbf{f}} \mathbf{e}_i}{\sigma_i^2}$$

The beta is in general non zero even if $b_i = 0$, because $\beta = \sum_{k \neq 0} \rho_{k,i} b_k (\sigma_k / \sigma_i)$. Factors other than momentum, but that are correlated to it, are responsible for

³We use the notation \mathbf{e}_i for the vector having a 1 in the i th position: $(0, \dots, 1, \dots, 0)$. We also assume that the factor portfolio has negligible idiosyncratic variance.

the transmission of the shock.

Let us go through another example. You are developing a risk model with a country factor (whose loadings are all ones) and a historical beta factor. You have the option of z-scoring the historical beta loadings. The choice to z-score does not affect the performance of the risk model, i.e., z-scoring is a model rotation if a country factor is present, and aggregated factor risk does not change if you z-score or not; nor does the aggregate performance attribution. However, PnL attributions of the individual beta and country factor change. Z-scoring makes attribution in the beta factor much smaller. What is the “right” choice? What criteria should we use? This is relevant for *ex post* analysis. If a portfolio has a large factor drawdown, it is possible that the PnL be spread across multiple factors and that no factor stands out. It is also possible that all these factor losses may be correlated. For example, losses in many industries could be “explained” as Momentum losses, or crowding losses. The problem is not only associated with performance analysis. The *ex ante* risk associated with a factor depends on the representation of the factor itself in the risk model. By this, we mean that the information contained in a given set of factors can be represented in different ways. The same factor may have zero correlation to other factors in one representation, and positive correlation in another. The central question is then: pick a *subset* of factors. Is there a single, non-ambiguous way to assign performance attribution and risk to this subset, such that it explains the PnL and the risk of the portfolio as much as possible?

The answer is in the affirmative: there is a procedure to assign unequivocally maximum risk and PnL to a subset of factors. There are four different ways to formulate and model the problem, all yielding the same result.

We introduce some notation. Denote the sets

$$\begin{aligned}\mathcal{U} &:= \{1, \dots, m\} \\ \mathcal{S} &:= \{1, \dots, p\} \\ \bar{\mathcal{S}} &:= \{p + 1, \dots, m\}\end{aligned}$$

so we write $\mathbf{f}_{\bar{\mathcal{S}}}$ instead of $\mathbf{f}_{(p+1):m}$ or $\boldsymbol{\Omega}_{\mathcal{U}, \bar{\mathcal{S}}}$ instead of $\boldsymbol{\Omega}_{1:p, (p+1):m}$.

1. *Maximal Cross-Sectional Return Explanation.* Consider the problem of describing the asset returns as function of the returns of factor u as well as possible, i.e.,

$$\mathbf{r} = \beta \mathbf{f}_{\mathcal{S}} + \boldsymbol{\eta} \tag{13.2}$$

where $\beta \in \mathbb{R}^{n \times p}$ and η is independent of \mathbf{f}_S . By construction, this is the maximum amount of returns we can attribute to factors \mathcal{U} . Once we identify beta, the return attributed to the factors is $\beta' \mathbf{f}_S$, which is in general different than $\mathbf{B}_{S,\mathcal{U}} \mathbf{f}_S$. We solve the problem

$$\begin{aligned} & \min E \|\eta\|^2 \\ \text{s.t. } & \mathbf{r} = \mathbf{B}\mathbf{f} + \epsilon \\ & \mathbf{r} = \beta \mathbf{f}_S + \eta \\ & \beta \in \mathbb{R}^{n \times p} \end{aligned} \tag{13.3}$$

This is equivalent to

$$\min_{\beta} E \|\mathbf{B}\mathbf{f} - \beta \mathbf{f}_S\|^2$$

which is solved by $\beta = \mathbf{B}\Omega_{\mathcal{U},S}\Omega_{S,S}^{-1}$. Then the attribution using factor set S is given by

$$\begin{aligned} \mathbf{w}'\mathbf{r} &= \mathbf{w}'\beta \mathbf{f}_S + (\text{Pnl independent of } \mathbf{f}_S) \\ \mathbf{w}'\beta \mathbf{f}_S &= \mathbf{b}'\Omega_{\mathcal{U},S}\Omega_{S,S}^{-1}\mathbf{f}_S \\ &= \mathbf{b}'_S\Omega_{S,S}\Omega_{S,S}^{-1}\mathbf{f}_S + \mathbf{b}'_{\bar{S}}\Omega_{\bar{S},S}\Omega_{S,S}^{-1}\mathbf{f}_S \\ &= \mathbf{b}'_S\mathbf{f}_S + \mathbf{b}'_{\bar{S}}\Omega_{\bar{S},S}\Omega_{S,S}^{-1}\mathbf{f}_S \end{aligned} \tag{13.4}$$

The term $\mathbf{w}'\beta \mathbf{f}_S$ is the *maximal* attribution to factors in S . When factors in u are uncorrelated to factors in S , the factor covariance matrix has a blockwise structure, with $\Omega_{S,U} = 0$, and then $\beta = \mathbf{B}_{\cdot,S}$, so that standard performance attribution and maximal attribution are the same. But in general, the factors in S and in \bar{S} are correlated and $\Omega_{S,U} \neq 0$. Maximal attribution shifts the PnL attributable to factors in S from the other factors.

2. *Conditional Expectation.* There is another way to interpret these formulas, based on conditional distribution of the multivariate Gaussian distribution. Given returns \mathbf{f}_S , the conditional expected returns of factors in \bar{S} are known analytically and are given by the vector

$$E(\mathbf{f}_{\bar{S}}|\mathbf{f}_S) = \Omega_{\bar{S},S}\Omega_{S,S}^{-1}\mathbf{f}_S \tag{13.5}$$

The formula for the normal performance attribution is

$$\mathbf{b}'_S\mathbf{f}_S + \mathbf{b}'_{\bar{S}}E(\mathbf{f}_{\bar{S}}|\mathbf{f}_S) = \mathbf{b}'_S\mathbf{f}_S + \mathbf{b}'_{\bar{S}}\Omega_{\bar{S},S}\Omega_{S,S}^{-1}\mathbf{f}_S$$

and this is identical to the maximal attribution term in Equation (13.4).

3. *Maximal Portfolio PnL Explanation.* Start with the factor PnL of the portfolio \mathbf{w}' , with factor exposure \mathbf{b} . Try to explain as much of this PnL by means of the returns of factors in set \mathcal{S} . In formulas, we solve the problem

$$\begin{aligned} \min_{\tilde{\mathbf{b}} \in \mathbb{R}^p} E \left\| \mathbf{b}' \mathbf{f} - \tilde{\mathbf{b}}' \mathbf{f}_{\mathcal{S}} \right\|^2 &= \min_{\tilde{\mathbf{b}} \in \mathbb{R}^p} \mathbf{b}' \Omega_{\mathcal{U}, \mathcal{U}} \mathbf{b} + \mathbf{x}' \Omega_{\mathcal{S}, \mathcal{S}} \mathbf{x} - 2\mathbf{b}' \Omega_{\mathcal{U}, \mathcal{S}} \tilde{\mathbf{b}} \\ &\Rightarrow \tilde{\mathbf{b}} = \Omega_{\mathcal{S}, \mathcal{S}}^{-1} \Omega_{\mathcal{S}, \mathcal{U}} \mathbf{b} \end{aligned}$$

and the PnL attribution is $\tilde{\mathbf{b}}' \mathbf{f}_{\mathcal{S}} = \mathbf{b}' \Omega_{\mathcal{U}, \mathcal{S}} \Omega_{\mathcal{S}, \mathcal{S}}^{-1} \mathbf{f}_{\mathcal{S}}$, which is, again, we obtain Equation (13.4).

This suggests an interpretation of the vector $\tilde{\mathbf{b}}$ as the *adjusted-dollar betas* of the portfolio to factors in set \mathcal{S} .

4. *Uncorrelated Factor Rotation.* We have seen in Section 3.4.1 that factor models are not uniquely determined. One can transform the loadings matrix by right-multiplying it by a non-singular square matrix \mathbf{C} , and correspondingly transform the factor returns by left-multiplying them by \mathbf{C}^{-1} . The resulting risk model has factor covariance matrix $\mathbf{C}^{-1} \Omega (\mathbf{C}^{-1})'$. It makes the same predictions as the original risk model, in the sense that the factor variance predicted by the two models is identical, and so is the total factor PnL attribution. However, the PnL attributed to the individual factors will change. $[\mathbf{w}' \mathbf{B}]_i f_i$ is not the same as $[\mathbf{w}' \mathbf{B} \mathbf{C}]_i [\mathbf{C}^{-1} \mathbf{f}]_i$. We ask whether there is an equivalent model that yields the above “maximal attribution” for the first p factors, and what is its interpretation. The answer to the first question is simple, given the previous derivations. We need to find \mathbf{C} such that

$$\sum_{i \in \mathcal{S}} [\mathbf{b}' \mathbf{C}]_i [\mathbf{C}^{-1} \mathbf{f}]_i = \mathbf{b}' \Omega_{\mathcal{U}, \mathcal{S}} \Omega_{\mathcal{S}, \mathcal{S}}^{-1} \mathbf{f}_{\mathcal{S}} \quad (13.6)$$

define the matrix $\mathbf{A} := \Omega_{\mathcal{U}, \mathcal{S}} \Omega_{\mathcal{S}, \mathcal{S}}^{-1}$ and the rotation matrix \mathbf{C} as

$$\begin{aligned} \mathbf{C} &:= \begin{pmatrix} \mathbf{I}_{\mathcal{S}, \mathcal{S}} & 0 \\ \mathbf{A} & \mathbf{I}_{\bar{\mathcal{S}}, \bar{\mathcal{S}}} \end{pmatrix} \\ \Rightarrow \quad \mathbf{C}^{-1} &= \begin{pmatrix} \mathbf{I}_{\mathcal{S}, \mathcal{S}} & 0 \\ -\mathbf{A} & \mathbf{I}_{\bar{\mathcal{S}}, \bar{\mathcal{S}}} \end{pmatrix} \end{aligned}$$

Direct calculation shows that

$$[\mathbf{b}' \mathbf{C}]_{\mathcal{S}} = \mathbf{b}'_{\mathcal{S}} + \mathbf{b}'_{\bar{\mathcal{S}}} \Omega_{\mathcal{U}, \mathcal{S}} \Omega_{\mathcal{S}, \mathcal{S}}^{-1} = \mathbf{b}' \Omega_{\mathcal{U}, \mathcal{S}} \Omega_{\mathcal{S}, \mathcal{S}}^{-1} \mathbf{f}_{\mathcal{S}} \quad (13.7)$$

which is the same as Equation (13.3).

In the rotated risk model the covariance matrix is

$$\begin{aligned}\mathbf{C}^{-1}\boldsymbol{\Omega}(\mathbf{C}^{-1})' &= \begin{pmatrix} \mathbf{I}_{\mathcal{S},\mathcal{S}} & 0 \\ \mathbf{A} & \mathbf{I}_{\bar{\mathcal{S}},\bar{\mathcal{S}}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}} & \boldsymbol{\Omega}_{\mathcal{S},\bar{\mathcal{S}}} \\ \boldsymbol{\Omega}_{\bar{\mathcal{S}},\mathcal{S}} & \boldsymbol{\Omega}_{\bar{\mathcal{S}},\bar{\mathcal{S}}} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{\mathcal{S},\mathcal{S}} & 0 \\ -\mathbf{A} & \mathbf{I}_{\bar{\mathcal{S}},\bar{\mathcal{S}}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}} & 0 \\ 0 & \boldsymbol{\Omega}_{\bar{\mathcal{S}},\bar{\mathcal{S}}} - \boldsymbol{\Omega}_{\bar{\mathcal{S}},\mathcal{S}}\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}^{-1}\boldsymbol{\Omega}_{\mathcal{S},\bar{\mathcal{S}}} \end{pmatrix}\end{aligned}\quad (13.8)$$

The interpretation of the transformation is that it makes the first p factors independent from the remaining ones. The returns and volatilities of the first p factors are unchanged, and the volatilities of the remaining ones are reduced. This is unintuitive at first sight, but has a simple interpretation: we have orthogonalized the factors in the set $\bar{\mathcal{S}}$, and pushed the explanatory power in the first p ones. The dollar exposures of a portfolio for the first p factors, on the other side, are *changing* from \mathbf{b} to $\mathbf{C}'\mathbf{b}$, as per Equation (13.7) so that the volatility and the performance attributable to them is increasing as well.

Procedure 13.1: *Maximal Attribution*

1. **Inputs:** Factor covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{m \times m}$; loadings matrix $\mathbf{B} \in \mathbb{R}^{n \times p}$, factor universe $\mathcal{U} := \{1, \dots, m\}$; sets $\mathcal{S}, \bar{\mathcal{S}} := \mathcal{U}/\mathcal{S}$; portfolio \mathbf{w} .
2. Set

$$\begin{aligned}\mathbf{b} &:= \mathbf{B}'\mathbf{w} \\ \mathbf{A} &:= \boldsymbol{\Omega}_{\mathcal{U},\mathcal{S}}\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}^{-1} \\ \mathbf{C} &:= \begin{pmatrix} \mathbf{I}_{\mathcal{S},\mathcal{S}} & 0 \\ \mathbf{A} & \mathbf{I}_{\bar{\mathcal{S}},\bar{\mathcal{S}}} \end{pmatrix}\end{aligned}$$

3. **Output:**

Per-factor maximal PnL: $\text{PnL}_k = [\mathbf{b}'\mathbf{A}]_k f_k$, for all $k \in \mathcal{S}$.
Rotated Factor covariance matrix: $\tilde{\boldsymbol{\Omega}} := \mathbf{C}^{-1}\boldsymbol{\Omega}(\mathbf{C}^{-1})'$

Let us go through an example. We have a sector strategy for which we performance daily factor performance attribution, which is shown in Figure 13.1 (top).

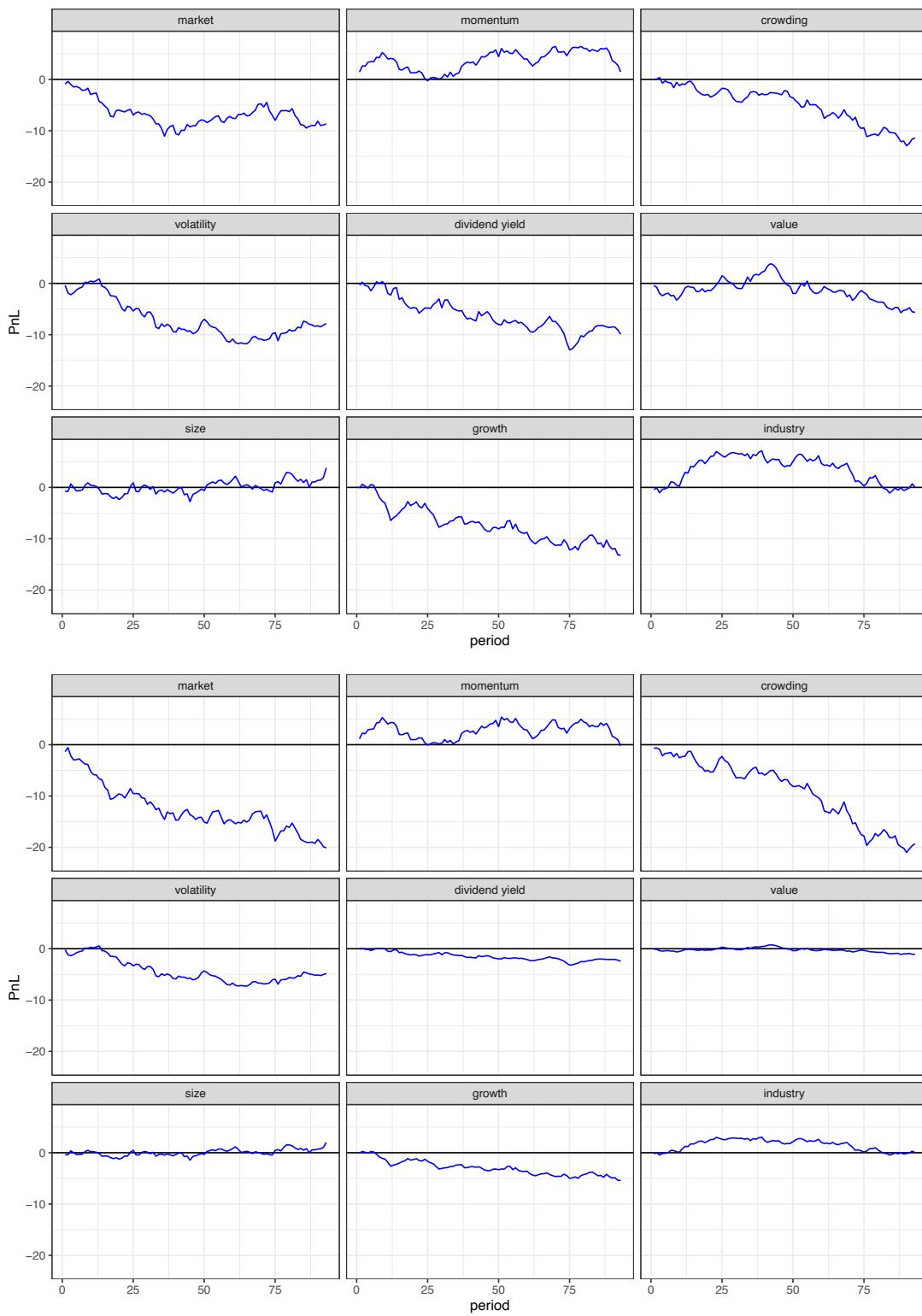


Figure 13.1: Top: PnL base factor performance attribution. Bottom: Maximal attribution on three factors: market, momentum, and crowding.

We select as maximal factors the market, momentum, and crowding factors. After rotating the remaining factors, the performance attribution changes significantly and is shown in Figure 13.1 (bottom). Market is responsible for a higher loss; crowding is “flatter” than in the regular attribution, whereas momentum changes sign: it had a cumulative PnL of \$2.6M in the regular attribution, but it has \$-5M in the maximal attribution.

Procedure 13.2: *Nested Maximal Attribution*

1. **Inputs:** Factor covariance matrix $\Omega \in \mathbb{R}^{m \times m}$; $\mathcal{U} := \{1, \dots, m\}$; set partition $\mathcal{S}_1, \dots, \mathcal{S}_p$ of \mathcal{U} ; portfolio \mathbf{w} .
2. For $i = 1, \dots, p$:
 - a) Perform Maximal Attribution (Procedure 13.1) on Ω , \mathbf{B} , \mathcal{U} , \mathcal{S}_i , \mathbf{w} .
 - b) Set $\Sigma^{(i)} := \Omega_{\mathcal{S}_i, \mathcal{S}_i}$, $\mathcal{U} \leftarrow \mathcal{U}/\mathcal{S}_i$, $\mathbf{B} \leftarrow \mathbf{B}_{\cdot, \mathcal{U}}$, $\Omega \leftarrow \tilde{\Omega}_{\mathcal{U}, \mathcal{U}}$.
3. Return PnL_k , for $k \in \mathcal{U}$, and the rotated risk model

$$\begin{pmatrix} \Sigma^{(1)} & 0 & \dots & 0 & 0 \\ 0 & \Sigma^{(2)} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \Sigma^{(p-1)} & 0 \\ 0 & \dots & \dots & 0 & \Omega \end{pmatrix}$$

We close this section with two observations. First, we focused on the “maximal attribution” factors. Alternatively, we could focus on the factors in the set $\bar{\mathcal{S}}$. If the performance attributable to these factor is small, then we have identified “minimal attribution” factors. The model has been rotated so that the portfolio performance has been described by a smaller dimensional space.

Secondly, we can perform a *nested* maximal performance attribution. Instead of having a “maximal attribution” set and a “minimal attribution” set, we extend the approach to a partition of the factor set $\{1, \dots, m\}$ by factor sets \mathcal{S}_i . Factor set \mathcal{S}_1 gets the maximal attribution; set \mathcal{S}_2 gets the maximal attribution of the remaining PnL; and so on. The most granular instance is that where $\mathcal{S}_i = \{i\}$, so that we orthogonalize the model sequentially one factor at a time. In practice,

however, it may be more sensible to create a coarser partition, every element of which describes a common theme. For example, we may have a “market factor” set composed of country, market and volatility factors; then a “value factor” set composed of earnings yield, earning variation, dividend yield, book-to-price, and quality; a “sentiment factor” set, an “industry set”, and so on. The steps involved in simple maximal attribution and nested attribution are described in Procedures 13.1 and 13.2.

13.4 Selection vs. Sizing Attribution

In factor-based attribution, the idiosyncratic profit and loss (PnL) of a strategy is the most crucial performance term, representing the PnL that cannot be explained by factor exposure. While factor-based attribution identifies the non-idiosyncratic portion of the PnL, it fails to explain the source of idiosyncratic performance. Portfolio managers often consider asset selection and sizing as the primary sources of their skills. Selection skill refers to the ability to be long on stocks with positive returns and short on those with negative returns. Sizing skill means being more profitable when right than when wrong. These skills have practical implications for portfolio construction and can lead to improved risk-adjusted performance. Quantitative analysts have developed “hitting” and “slugging” metrics to quantify selection and sizing. Hitting is the percentage of profitable single-asset investments, while slugging is the ratio between the average PnL of profitable and unprofitable investments. Despite their intuitive appeal, these measures have two drawbacks: they lack a direct relationship with profitability measures like Information Ratio, and do not provide clear guidance for portfolio managers.

This section aims to address these problems. We show how a new selection-sizing decomposition achieves three objectives:

1. It links through an analytical, interpretable formula the Information Ratio (IR) of a strategy to selection, sizing and breadth of a portfolio;
2. It provides guidance for portfolio managers, both in the case that the strategy has positive sizing skill and that it has negative sizing skill.

The IR is the expected value of the idiosyncratic PnL divided by its standard deviation. If we restrict our attention to a single period, an estimate of the IR is

$$\widehat{IR}_t = \frac{(Idio\ PnL)_t}{(Idio\ Vol)_t}$$

An estimate for the IR that employs the available time series of portfolios in epochs $1, 2, \dots, T$ is

$$\widehat{\text{IR}} = \frac{1}{T} \sum_{t=1}^T \widehat{\text{IR}}_t$$

The IR can be expressed as a simple combination of intuitive terms. The decomposition is

$$\widehat{\text{IR}} = \frac{1}{T} \sum_{t=1}^T [(selection)_t \times (diversification)_t + (sizing)_t]$$

The terms in the identity are:

- A *selection* skill.

$$(selection)_t := \frac{1}{n} \sum_{i=1}^n \tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i}) .$$

We z-score the idiosyncratic return $\epsilon_{t,i}$ of an asset to obtain $\tilde{\epsilon}_{t,i} := \epsilon_{t,i}/\sigma_i$ and multiply it by the sign of that asset's holding. If holding and return have the same sign, the portfolio manager was on the right side of a security bet in a specific period and the contribution to selection is positive. Z-scoring puts assets with different volatility on the same scale, so that selection does not reward the magnitude of the return.

- *Diversification*. Instead of reasoning about the notional value of positions, we use the dollar volatility of each position, defined as $\tilde{w}_{t,i} := \sigma_i w_{t,i}$. Then we define

$$(diversification)_t := \frac{\|\tilde{\mathbf{w}}_t\|_1}{\|\tilde{\mathbf{w}}_t\|_2}$$

When all the dollar volatilities are identical, then the portfolio diversification is \sqrt{n} . At the other end, if the portfolio has a single position, then the portfolio diversification is 1. The diversification squared ranges between 1 and n , and can be interpreted as the effective number of assets. This diversification term has a well-known connection to the Herfindahl Index, which is a measure of concentration. To be more specific, define weights $x_i := |\tilde{w}_{t,i}| / \sum_j |\tilde{w}_{t,j}|$. The Herfindahl Index is defined as $H := \sum_i x_i^2$. The relationship is then $(diversification)_t = 1/\sqrt{H}$. The relationship between diversification and portfolio construction was first explored by [Bouchaud et al. \(1997\)](#).

- The last term is *sizing*. It is equal to here, $\widehat{\text{cov}}$ is a cross-sectional covariance, where we treated the quantities associated to individual assets as empirical observations⁴. The interpretation of sizing is that it measures the correlation between being on the right side of a bet $\tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i})$, and the bet size $|\sigma_i w_{t,i}|$. Sizing is positive if, when the portfolio manager is right about the *side* of a position, they are right about its *size* by having a relatively large position. In formulas, we first we define

$$(sizing)_t := \frac{n}{\|\tilde{\mathbf{w}}_t\|} \widehat{\text{cov}}(\underbrace{\tilde{\epsilon}_t \circ \text{sgn}(\tilde{\mathbf{w}}_t)}_{(right-side \ index)}, \underbrace{|\tilde{\mathbf{w}}_t|}_{(bet \ size)})$$

This equation can be used in several ways. To achieve a higher IR, a portfolio manager has the following three options:

- *Increase diversification.* Markowitz famously said that diversification is the only free lunch in investing. This equation shows that benefits from diversification are accrued via selection skill, i.e., selection is the marginal benefit obtained by increasing diversification. This reasoning is not entirely correct, however. Managers can increase diversification in two ways. The first one is by making portfolio positions more equal. This does not require additional effort⁵. Alternatively, the portfolio manager could add stocks to the investment universe. This operation is not costless, since it would involve spending less time on each stock, and possibly cover less desirable stocks not in the primary universe. When increasing diversification, the manager may want to consider the impact on stock selection from this decision.
- *Improve selection skill.* The decomposition helps by providing a simple measure, which makes use of the entire dataset at a manager's disposal: daily positions, PnL, and idiosyncratic risk of the individual positions. Once selection can be measured, several actions are possible. For example, the portfolio manager can track selection skill at the sub-industry or at the thematic level; or the portfolio manager can compare performance during earnings versus outside earnings. Improving portfolio selection is not easy, but is possible.

⁴More rigorously, for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\widehat{\text{cov}}(\mathbf{x}, \mathbf{y}) := n^{-1} \sum_i x_i y_i - n^{-2} \sum_j x_j \sum_k y_k$.

⁵The analysis presented here does not take into account transaction costs. This is a reasonable approximation for small portfolios. A more comprehensive model is possible, but outside of this article's scope.

- *Improve sizing skill.* There is value already in having portfolio managers assess their sizing skill relative to selection; most portfolio managers overestimate their sizing skill, and find the low sizing skill, or even the absence thereof, instructive. If their sizing skill is *negative*, the portfolio manager should not differentiate positions according to size. In doing so, they will eliminate the drag from negative sizing and magnify the benefit of stock selection, by maximizing breadth. If there is *positive* sizing skill, the portfolio manager can optimize the size of the high-conviction positions to maximize the IR. This is the subject of the next subsection.

13.4.1 Connection to the Fundamental Law of Active Management

This formula bears some resemblance to Grinold and Kahn's Fundamental Law of Active Management ([Grinold and Kahn, 1999](#)). That formula stated that the IR is the product of the Information Coefficient and the breadth of the portfolio \sqrt{n} . This formula uses a different portfolio breadth –the effective breadth–which treats not all positions as equal. For example, a portfolio of 100 stocks with a gross notional of \$1 million in each of them does not have the same breadth of a portfolio of 100 stocks where one position is \$999,901 in one stock and \$1 in the remaining 99. In their seminal article, [Bouchaud et al. \(1997\)](#) present a modified mean-variance portfolio formulation that puts a lower bound on our definition of diversification. This results in using a shrunked covariance matrix. The same approach has been advocated using robust portfolio construction models ([Stubbs and Vance, 2005](#); [V. et al., 2013](#); [Pedersen et al., 2021](#)) and penalized covariance estimation methodologies ([Ledoit and Wolf, 2003b](#)).

13.4.2 Long-Short Performance Attribution

The selection component of our performance attribution is linear, and therefore lends itself naturally to be further partitioned in different performance subclasses. A natural partition is long versus short; that is, the fraction of selection skill that arises from being on the right side of returns when positions are long, versus when positions are short. The decomposition follows from the chain of equalities below.

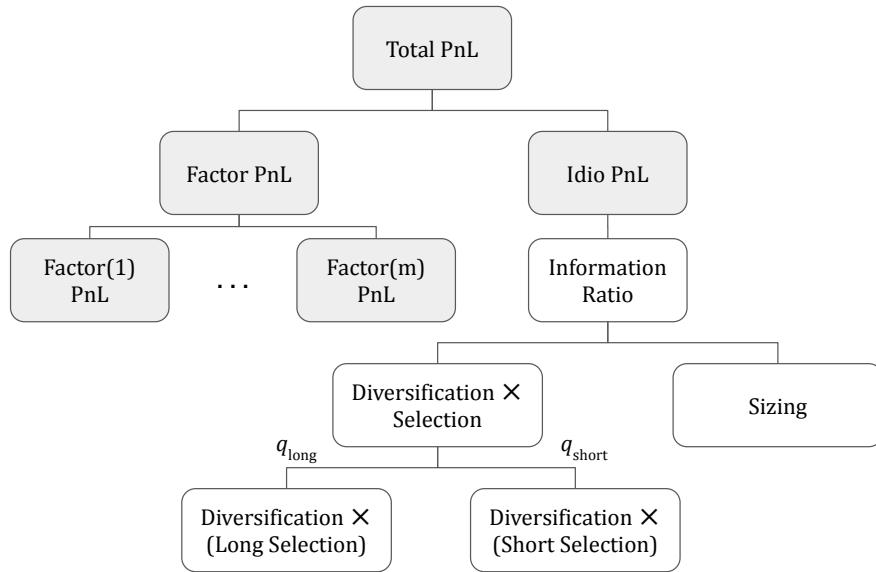


Figure 13.2: A Taxonomy of Performance Attribution.

$$\begin{aligned}
 (selection)_t &= \frac{1}{n} \sum_{i:w_{t,i}>0} \tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i}) + \frac{1}{n} \sum_{i:w_{t,i}<0} \tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i}) \\
 &= \frac{n_{\text{long}}}{n} \frac{1}{n_{\text{long}}} \sum_{i:w_{t,i}>0} \tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i}) + \frac{n_{\text{short}}}{n} \frac{1}{n_{\text{short}}} \sum_{i:w_{t,i}<0} \tilde{\epsilon}_{t,i} \text{sgn}(w_{t,i}) \\
 &= q_{\text{long}} \times (selection)_{L,t} + q_{\text{short}} \times (selection)_{S,t}
 \end{aligned}$$

where n_{long} , n_{short} are the number of long and short positions, and q_{long} , q_{short} are the fraction of the total portfolio positions that are long and short, respectively.

Summing up, in Figure 13.2 we show the dependency tree of the decomposition terms.

13.5 Time-Series Performance Attribution

Current Status: there is nothing interesting here so far.

Start with a generic factor model. Let $\mathbf{R} \in \mathbb{R}^{n \times T}$ the matrix of returns. The factor-mimicking portfolio returns are $\mathbf{F} = (\mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{B})^{-1} \mathbf{B}' \boldsymbol{\Omega}_\epsilon^{-1} \mathbf{R}$. Given these returns, let us estimate the time-series beta of returns to factor returns. Denote

these $\tilde{\mathbf{B}}$. They are the solution to the problem

$$\begin{aligned}\min \left\| \Omega_{\epsilon}^{-1/2} (\mathbf{R} - \tilde{\mathbf{B}} \mathbf{F}) \right\|_F^2 &= \min \left\| \mathbf{R}' \Omega_{\epsilon}^{-1/2} - \mathbf{F}' \tilde{\mathbf{B}}' \Omega_{\epsilon}^{-1/2} \right\|^2 \\ \tilde{\mathbf{B}}' \Omega_{\epsilon}^{-1/2} &= (\mathbf{F} \mathbf{F}')^{-1} \mathbf{F} \mathbf{R}' \Omega_{\epsilon}^{-1/2} \\ \tilde{\mathbf{B}}' &= (\mathbf{F} \mathbf{F}')^{-1} \mathbf{F} \mathbf{R}'\end{aligned}$$

Start with the usual model $\mathbf{r} = \mathbf{Bf} + \epsilon$ with $\Omega_f = \mathbf{I}_m$, and $\Omega_{\epsilon} = \sigma^2 \mathbf{I}_n$, estimate $\mathbf{F} = (\mathbf{B}' \mathbf{B})^{-1} \mathbf{B}' \mathbf{R}$. Estimate $\tilde{\mathbf{B}}$ by minimizing $\left\| \mathbf{R} - \tilde{\mathbf{B}} \mathbf{F} \right\|^2$. The solution is

$$\tilde{\mathbf{B}}' = (\mathbf{F} \mathbf{F}')^{-1} \mathbf{F} \mathbf{R}'$$

This follows from Equation (8.4): $\arg \min_{\mathbf{X}} \left\{ \|\mathbf{A} - \mathbf{B} \mathbf{X} \right\|_F^2 \} = (\mathbf{B}' \mathbf{B})^{-1} \mathbf{B}' \mathbf{A}$, applied to $\left\| \mathbf{R}' - \mathbf{F}' \tilde{\mathbf{B}}' \right\|^2$.

$$\begin{aligned}\mathbf{R} \mathbf{F}' &= \mathbf{R} \mathbf{R}' \Omega_{\epsilon}^{-1} \mathbf{B} (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \\ &= T \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B} (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \\ \mathbf{F} \mathbf{F}' &= T (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B} (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \\ \tilde{\mathbf{B}} &= (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B}) (\mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B}) (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \\ &= (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B}) (\mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r\end{aligned}$$

Under which condition $\tilde{\mathbf{B}} = \mathbf{B}$?

$$\begin{aligned}\mathbf{B}' &= (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B}) (\mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \\ \mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B} &= (\mathbf{B}' \Omega_{\epsilon}^{-1} \mathbf{B}) (\mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B})^{-1} \mathbf{B}' \Omega_{\epsilon}^{-1} \Omega_r \Omega_{\epsilon}^{-1} \mathbf{B}\end{aligned}$$

Which is always satisfied.

13.6 Appendix*

13.6.1 Proof of the Selection vs. Sizing Decomposition

Theorem 13.1. Consider a portfolio sequence $\mathbf{w}_t \in \mathbb{R}^n$, and a sequence of iid idiosyncratic returns ϵ_t taking values in \mathbb{R}^n , with $\text{cov}(\epsilon_t) = \Omega$. Define the Empirical Information Ratio:

$$\widehat{\text{IR}} = \frac{1}{T} \sum_{t=1}^T \frac{(Idio \ PnL)_t}{(Idio \ Vol)_t}$$

Then the identity holds

$$\widehat{IR} = \frac{1}{T} \sum_{t=1}^T [(selection)_t \times (diversification)_t + (sizing)_t] \quad (13.9)$$

where the terms in the equation above are defined as follows:

$$\begin{aligned}\tilde{\mathbf{w}}_t &:= \boldsymbol{\Omega}^{1/2} \mathbf{w}_t \\ \tilde{\mathbf{u}}_t &:= \sqrt{n} \frac{\tilde{\mathbf{w}}_t}{\|\tilde{\mathbf{w}}_t\|} \\ \tilde{\boldsymbol{\epsilon}}_t &:= \boldsymbol{\Omega}^{-1/2} \boldsymbol{\epsilon}_t \\ (selection)_t &:= \hat{E}(\tilde{\boldsymbol{\epsilon}}_t \circ sgn(\tilde{\mathbf{u}}_t)) \\ (diversification)_t &:= \sqrt{n} \hat{E}(|\tilde{\mathbf{u}}_t|) \\ &= \frac{\|\tilde{\mathbf{w}}_t\|_1}{\|\tilde{\mathbf{w}}_t\|_2} \\ (sizing)_t &:= \sqrt{n} \widehat{cov}(\tilde{\boldsymbol{\epsilon}}_t \circ sgn(\tilde{\mathbf{u}}_t), |\tilde{\mathbf{u}}_t|) \\ &= \frac{n}{\|\mathbf{w}_t\|} \widehat{cov}(\tilde{\boldsymbol{\epsilon}}_t \circ sgn(\tilde{\mathbf{w}}_t), |\tilde{\mathbf{w}}_t|)\end{aligned}$$

Proof. In period t , the risk-adjusted PnL of the portfolio at time t is given by

$$\widehat{IR}_t = \frac{\mathbf{w}'_t \boldsymbol{\epsilon}_t}{\sqrt{\mathbf{w}'_t \boldsymbol{\Omega} \mathbf{w}_t}}$$

Set $\tilde{\mathbf{w}}_t := \boldsymbol{\Omega}^{1/2} \mathbf{w}_t$, and $\tilde{\boldsymbol{\epsilon}}_t := \boldsymbol{\Omega}^{-1/2} \boldsymbol{\epsilon}_t$. The vector $\tilde{\mathbf{w}}_t$ has a familiar interpretation. It is a portfolio whose positions are not expressed as NMV but rather as dollar volatilities in each asset. The return vector $\tilde{\boldsymbol{\epsilon}}_t$ contains the z-scored asset returns. Its covariance matrix is the identity. With these transformations, the sample IR takes a simpler form:

$$\widehat{IR}_t = \frac{\tilde{\mathbf{w}}'_t \tilde{\boldsymbol{\epsilon}}_t}{\|\tilde{\mathbf{w}}_t\|}$$

This follows from the fact that the numerator is

$$\begin{aligned}\mathbf{w}'_t \boldsymbol{\epsilon}_t &= \sum_i w_{t,i} \epsilon_{t,i} = \sum_i (\sigma_i w_{t,i}) (\epsilon_{t,i} / \sigma_i) \\ &= \sum_i \tilde{w}_{t,i} \tilde{\epsilon}_{t,i}\end{aligned}$$

and the denominator is

$$\sqrt{(\mathbf{w}'_t \boldsymbol{\Omega}^{1/2})(\boldsymbol{\Omega}^{1/2} \mathbf{w}_t)} = \|\boldsymbol{\Omega}^{1/2} \mathbf{w}_t\| = \|\tilde{\mathbf{w}}_t\|$$

We can further simplify the formula by considering a breadth-rescaled percentage of the total dollar volatility:

$$\begin{aligned}\widehat{\text{IR}}_t &= \sum_i \tilde{\epsilon}_i \frac{\tilde{w}_i}{\|\tilde{\mathbf{w}}\|} \\ &= \sum_i \tilde{\epsilon}_i \text{sgn}(\tilde{w}_i) \frac{|\tilde{w}_i|}{\|\tilde{\mathbf{w}}\|} \\ &= \sqrt{n} \frac{1}{n} \sum_i \tilde{\epsilon}_i \text{sgn}(\tilde{w}_i) \frac{\sqrt{n} |\tilde{w}_i|}{\|\tilde{\mathbf{w}}\|} \\ &= \sqrt{n} \frac{1}{n} \sum_i \tilde{\epsilon}_i \text{sgn}(\tilde{u}_i) |\tilde{u}_i|\end{aligned}$$

where we set

$$\tilde{\mathbf{u}}_t := \sqrt{n} \frac{\tilde{\mathbf{w}}_t}{\|\tilde{\mathbf{w}}_t\|}$$

We denote the *cross-sectional empirical average* and the *cross sectional empirical covariance*

$$\begin{aligned}\hat{E}(\mathbf{x}) &:= n^{-1} \sum_i x_i \\ \widehat{\text{cov}}(\mathbf{x}, \mathbf{y}) &:= \hat{E}[(\mathbf{x} - \hat{E}\mathbf{x})^2 (\mathbf{y} - \hat{E}\mathbf{y})^2]\end{aligned}$$

The formula becomes:

$$\widehat{\text{IR}}_t = \sqrt{n} \hat{E}(\tilde{\epsilon}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t) \circ |\tilde{\mathbf{u}}_t|)$$

where we have used the notation ‘ \circ ’ to denote the element-wise (Hadamard) product of two vectors, i.e.: $(\mathbf{x} \circ \mathbf{y})_i := x_i y_i$. Finally, in the last step we use the identity $\hat{E}(\mathbf{x} \circ \mathbf{y}) = \widehat{\text{cov}}(\mathbf{x}, \mathbf{y}) + \hat{E}(\mathbf{x}) \hat{E}(\mathbf{y})$ with $\mathbf{x} = \tilde{\epsilon}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t)$ and $\mathbf{y} = |\tilde{\mathbf{u}}_t|$. It follows that

$$\widehat{\text{IR}}_t = \sqrt{n} \left[\hat{E}(\tilde{\epsilon}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t)) \hat{E}(|\tilde{\mathbf{u}}_t|) + \widehat{\text{cov}}(\tilde{\epsilon}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t), |\tilde{\mathbf{u}}_t|) \right]$$

A possible interpretation of the above formula is as a sample of the realized IR over a single observation, or period. An estimate of the IR over the period $1, \dots, T$ is then given by its time-series average:

$$\begin{aligned}\widehat{\text{IR}} &= \frac{1}{T} \sum_{t=1}^T \frac{\tilde{\mathbf{w}}_t' \tilde{\boldsymbol{\epsilon}}_t}{\|\tilde{\mathbf{w}}_t\|} \\ &= \sqrt{n} \left[\frac{1}{T} \sum_{t=1}^T \hat{E}(\tilde{\boldsymbol{\epsilon}}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t)) \hat{E}(|\tilde{\mathbf{u}}_t|) + \frac{1}{T} \sum_{t=1}^T \widehat{\text{cov}}(\tilde{\boldsymbol{\epsilon}}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t), |\tilde{\mathbf{u}}_t|) \right]\end{aligned}$$

This is equal to Equation (13.9) once we define

$$\begin{aligned}(\text{selection})_t &= \hat{E}(\tilde{\boldsymbol{\epsilon}}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t)) \\ (\text{diversification})_t &= \sqrt{n} \hat{E}(|\tilde{\mathbf{u}}_t|) \\ &= \frac{\sum_{i=1}^n |\tilde{w}_{t,i}|}{\sqrt{\sum_{i=1}^n \tilde{w}_{t,i}^2}} \\ (\text{sizing})_t &= \sqrt{n} \widehat{\text{cov}}(\tilde{\boldsymbol{\epsilon}}_t \circ \text{sgn}(\tilde{\mathbf{u}}_t), |\tilde{\mathbf{u}}_t|) \\ &= \frac{n}{\sqrt{\sum_{i=1}^n \tilde{w}_{t,i}^2}} \widehat{\text{cov}}(\tilde{\boldsymbol{\epsilon}}_t \circ \text{sgn}(\tilde{\mathbf{w}}_t), |\tilde{\mathbf{w}}_t|)\end{aligned}$$

□

13.7 Exercises

Exercise 13.1. (35) Solve explicitly optimization problem (13.3). (Hint: solve first for T empirical observations of factor returns using Equation (8.4), and then send $T \rightarrow \infty$).

Chapter 14

★Appendix

14.1 Realized Variance of Minimum Variance Portfolios

Theorem 14.1. Let $\hat{\Omega}_r \in \mathbb{R}^{n \times n}$ be a candidate covariance matrix and Ω_r be the true covariance matrix. Let $\mathbf{b} \in \mathbb{R}^n$, and solve the risk minimization problem

$$\begin{aligned} & \min \mathbf{w}' \hat{\Omega}_r \mathbf{w} \\ & \text{s.t. } \mathbf{b}' \mathbf{w} = 1 \end{aligned} \tag{14.1}$$

and let $\mathbf{w}(\hat{\Omega}_r)$ be its solution. Denote the realized variance of the portfolio $\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)$.

The realized volatility of portfolio $\mathbf{w}(\hat{\Omega}_r)$ is greater than the one of $\mathbf{w}(\Omega_r)$, and the two are identical if and only if $\Omega_r \propto \hat{\Omega}_r$.

Proof. The solution of Problem (14.1) is $\mathbf{w}(\hat{\Omega}_r) = (\mathbf{b}' \hat{\Omega}_r^{-1} \mathbf{b})^{-1} \hat{\Omega}_r^{-1} \mathbf{b}$. The ratio between realized variance of the portfolios constructed on $\hat{\Omega}_r$ and on Ω_r is

$$\frac{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)}{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)} = \frac{\mathbf{b}' \Omega_r^{-1} \mathbf{b}}{\mathbf{b}' \hat{\Omega}_r^{-1} \mathbf{b}} \frac{\mathbf{b}' \hat{\Omega}_r^{-1} \Omega_r \hat{\Omega}_r^{-1} \mathbf{b}}{\mathbf{b}' \hat{\Omega}_r^{-1} \mathbf{b}}$$

One can verify directly that if $\hat{\Omega}_r^{-1} \propto \Omega_r$ the ratio is one. Let $\hat{\Omega}_r = \hat{\mathbf{U}} \hat{\mathbf{S}} \hat{\mathbf{U}}'$, $\Omega_r = \mathbf{U} \mathbf{S} \mathbf{U}'$. Let $\mathbf{x} := \hat{\mathbf{S}}^{-1/2} \hat{\mathbf{U}}' \mathbf{b}$. Let $\mathbf{H} := \hat{\mathbf{S}}^{1/2} \hat{\mathbf{U}}' \mathbf{U} \mathbf{S}^{-1} \mathbf{U}' \hat{\mathbf{U}} \hat{\mathbf{S}}^{1/2}$. Then we rewrite the variance ratio as

$$\begin{aligned} \frac{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)}{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)} &= \frac{\mathbf{x}' \hat{\mathbf{S}}^{1/2} \hat{\mathbf{U}}' \mathbf{U} \mathbf{S}^{-1} \mathbf{U}' \hat{\mathbf{U}} \hat{\mathbf{S}}^{1/2} \mathbf{x}}{\|\mathbf{x}\|^2} \frac{\mathbf{x}' \hat{\mathbf{S}}^{-1/2} \hat{\mathbf{U}}' \mathbf{U} \mathbf{S}^{-1} \mathbf{U}' \hat{\mathbf{U}} \hat{\mathbf{S}}^{-1/2} \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \frac{\mathbf{x}' \mathbf{H} \mathbf{x}}{\|\mathbf{x}\|^2} \frac{\mathbf{x}' \mathbf{H}^{-1} \mathbf{x}}{\|\mathbf{x}\|^2} \end{aligned}$$

Consider now the SVD of $\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}'$ and define $\mathbf{y} := \mathbf{V}'\mathbf{x}$. We have

$$\frac{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)}{\text{var}(\mathbf{w}(\hat{\Omega}_r), \Omega_r)} = \left(\sum_i \frac{y_i^2}{\sum_j y_j^2} d_i \right) \left(\sum_i \frac{y_i^2}{\sum_j y_j^2} d_i^{-1} \right)$$

The term on the RHS can be interpreted as $E(\xi)E(1/\xi)$, where ξ is a random variable taking value d_i is state i with probability $p_i := y_i^2 / \sum_j y_j^2$. By Jensen's inequality, $E(1/\xi) \geq 1/E(\xi)$ and the result follows. \square

14.2 Asymptotic Properties of Principal Component Analysis

This is a summary of the asymptotic properties of PCA in the regime where the number of variables n is constant and the number of observations T goes to infinity. We have T realizations of iid random vectors $\mathbf{x}_t \sim N(\mathbf{0}, \Sigma)$, from which we want to estimate Σ . We assume that the $\mathbf{x}_{t,i}$ have finite fourth moments. Let $\hat{\Sigma}_T := T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$. By the Law of Large Numbers, $\hat{\Sigma}_T \rightarrow \Sigma$ almost surely. Both eigenvalues and eigenvectors converge to the covariance matrix. [Anderson \(1963\)](#) proves a CLT for the eigenvalues of the covariance matrix. Decompose the empirical and true covariance matrices into their eigenvalues and eigenvectors:

$$\Sigma = \mathbf{U} \Lambda \mathbf{U}' \quad (14.2)$$

$$\hat{\Sigma}_T = \hat{\mathbf{U}} \hat{\Lambda} \hat{\mathbf{U}}' \quad (14.3)$$

with $\lambda_1 > \lambda_2 > \dots > \lambda_n$; all eigenvalues are assumed to be distinct. Anderson proves that, as $T \rightarrow \infty$,

$$\sqrt{T}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \sim N(\mathbf{0}, 2\Lambda) \quad (14.4)$$

$$\sqrt{T}(\hat{\mathbf{u}}_i - \mathbf{u}_i) \sim N(\mathbf{0}, \mathbf{E}_i) \quad (14.5)$$

$$\mathbf{E}_i := \mathbf{U} \begin{pmatrix} \frac{\lambda_1 \lambda_i}{(\lambda_1 - \lambda_i)^2} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2 \lambda_i}{(\lambda_2 - \lambda_i)^2} & \dots & 0 \\ 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\lambda_n \lambda_i}{(\lambda_n - \lambda_i)^2} \end{pmatrix} \mathbf{U}' \quad (14.6)$$

where the i th row has all zeros. Therefore:

1. the standard error on $\hat{\lambda}_i$ is $2\lambda_i/\sqrt{T}$.
2. the standard error on the principal components, defined as $\sqrt{E(\|\hat{\mathbf{u}}_i - \mathbf{u}_i\|^2)}$, is

$$\frac{1}{\sqrt{T}} \sqrt{\sum_{k=1, k \neq i}^n \frac{\lambda_k \lambda_i}{(\lambda_k - \lambda_i)^2}} \quad (14.7)$$

The relative error depends on the separation between eigenvalues.

14.3 The Linear-Quadratic Regulator

This section covers a model used in optimal trading problems. A few historical notes are in order. A special instance of this problem came to the fore in 2013 in a much-cited paper ([Gârleanu and Pedersen, 2013](#)). In a recent interview, Bouchaud said to have had the solution as far back as 2007 and blamed himself for not having published it. In both cases, they were referring to a special instance of 40-year old results from control theory, which are widely used in applications and even known in Economics¹. I cover the theory and provide one application (out of several) in finance. There are two wrinkles. First, an extension of the theory to more general quadratic forms. Second, the formulation of the trading problem in a way that makes it suitable for solution. The result allows for general alpha decays (not just exponential, as in the Gârleanu-Pedersen paper) and for general quadratic costs (not just with characteristic matrix proportional to the asset covariance matrix, as in the same paper).

14.4 The Discounted Linear-Quadratic Regulator

The theory is based on a fundamental family of models ubiquitous in control theory. The Linear-Quadratic Regulator (LQR) is a system in which the per-period cost is quadratic in the state vector and in the control vector, and the control impact is linear on the state. The problem has been analyzed in discrete and continuous time, and in deterministic, stochastic and robust settings. I present it here for two reasons. First, completeness; second, because for applications we need to use an extended cost function that has interaction terms in the state and the control. This formulation can't be found in the extant literature on the subject.

Consider the following process:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{C}\mathbf{w}_{t+1} \quad (14.8)$$

with $\mathbf{x}_t \in \mathbb{R}^n$, $\mathbf{u}_t \in \mathbb{R}^m$, $\mathbf{w}_t \sim N(0, \mathbf{I}_n)$ and A, B, C are of suitable dimension and full-rank. \mathbf{x}_t is the state vector, \mathbf{u}_t the control vector. The discounted

¹The two better-known proponents of these methods being perhaps Sargent and Hansen, both Nobel recipients in Economics.

infinite-horizon objective function is:

$$V := \sum_{t=1}^{\infty} \beta^t (\mathbf{x}'_t \mathbf{Q} \mathbf{x}_t + \mathbf{u}'_t \mathbf{R} \mathbf{u}_t + 2\mathbf{x}'_t \mathbf{W} \mathbf{u}_t + \mathbf{a}' \mathbf{x}_t + \mathbf{b}' \mathbf{u}_t) \quad (14.9)$$

where we assume \mathbf{Q}, \mathbf{R} to be symmetric positive definite. The Hamilton-Jacobi-Bellman equation is

$$V(\mathbf{x}) = \min_{\mathbf{u}} [\mathbf{x}' \mathbf{Q} \mathbf{x} + \mathbf{u}' \mathbf{R} \mathbf{u} + 2\mathbf{x}' \mathbf{W} \mathbf{u} + \mathbf{a}' \mathbf{x} + \mathbf{b}' \mathbf{u}] \quad (14.10)$$

$$+ \beta E_{\mathbf{w}} V(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{w})] \quad (14.11)$$

We solve the problem by “guessing” the solution and identifying conditions for it. The uniqueness of the solution relies on additional properties of the Hamilton-Jacobi-Bellman equation and we omit its proof.

Our candidate value function is

$$V(\mathbf{x}) = \mathbf{x}' \mathbf{P} \mathbf{x} + \mathbf{q}' \mathbf{x} + \rho \quad (14.12)$$

for some symmetric $\mathbf{P} \in \mathbb{R}^{n \times n}$, $\mathbf{q} \in \mathbb{R}^n$, $\rho \in \mathbb{R}$.

$$E_{\mathbf{w}} V(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{w}) = \mathbf{x}' \mathbf{A}' \mathbf{P} \mathbf{A} \mathbf{x} + \mathbf{u}' \mathbf{B}' \mathbf{P} \mathbf{B} \mathbf{u} \quad (14.13)$$

$$+ 2\mathbf{x}' \mathbf{A}' \mathbf{P} \mathbf{B} \mathbf{u} \quad (14.14)$$

$$+ \mathbf{q}' \mathbf{A} \mathbf{x} + \mathbf{q}' \mathbf{B} \mathbf{u} + \text{trace}(\mathbf{C}' \mathbf{P} \mathbf{C}) + \rho \quad (14.15)$$

Replacing the expectation in the HJB equation we get

$$V(\mathbf{x}) = \min_{\mathbf{u}} [\mathbf{x}' (\mathbf{Q} + \beta \mathbf{A}' \mathbf{P} \mathbf{A}) \mathbf{x} + \mathbf{u}' (\mathbf{R} + \beta \mathbf{B}' \mathbf{P} \mathbf{B}) \mathbf{u}] \quad (14.16)$$

$$+ (\mathbf{b} + \beta \mathbf{B}' \mathbf{q})' \mathbf{u} + (\mathbf{a} + \beta \mathbf{A}' \mathbf{q})' \mathbf{x} + 2\mathbf{x}' (\mathbf{W} + \beta \mathbf{A}' \mathbf{P}' \mathbf{B}) \mathbf{u} \quad (14.17)$$

$$+ \beta \text{trace}(\mathbf{C}' \mathbf{P} \mathbf{C}) + \beta \rho] \quad (14.18)$$

Let F be the function being minimized on the right-hand side. The quadratic

minimization problem has solution and optimal value

$$\mathbf{u}^* = -\frac{1}{2}(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}[(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) + 2(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})\mathbf{x}] \quad (14.19)$$

$$F(\mathbf{u}^*) = \mathbf{x}'(\mathbf{Q} + \beta\mathbf{A}'\mathbf{P}\mathbf{A})\mathbf{x} \quad (14.20)$$

$$+ \frac{1}{4}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) \quad (14.21)$$

$$+ \mathbf{x}'(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})\mathbf{x} \quad (14.22)$$

$$- \frac{1}{2}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) \quad (14.23)$$

$$- (\mathbf{b} + \beta\mathbf{B}'\mathbf{q})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})\mathbf{x} \quad (14.24)$$

$$+ (\mathbf{a} + \beta\mathbf{A}'\mathbf{q})'\mathbf{x} \quad (14.25)$$

$$- \mathbf{x}'(\mathbf{W}' + \beta\mathbf{A}'\mathbf{P}'\mathbf{B})(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}[2(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})\mathbf{x}] \quad (14.26)$$

$$- \mathbf{x}'(\mathbf{W}' + \beta\mathbf{A}'\mathbf{P}'\mathbf{B})(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}[(\mathbf{b} + \beta\mathbf{B}'\mathbf{q})] \quad (14.27)$$

$$+ \beta\text{trace}(\mathbf{C}'\mathbf{P}\mathbf{C}) + \beta\rho \quad (14.28)$$

$$(14.29)$$

We set $\mathbf{x}'\mathbf{P}\mathbf{x} + \mathbf{q}'\mathbf{x} + \rho = F(\mathbf{u}^*)$ and equate the terms of equal degree.

$$\mathbf{P} = \mathbf{Q} + \beta\mathbf{A}'\mathbf{P}\mathbf{A} + (\mathbf{W}' + \beta\mathbf{A}'\mathbf{P}\mathbf{B})(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A}) \quad (14.30)$$

$$\mathbf{q} = -2(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) + \mathbf{a} + \beta\mathbf{A}'\mathbf{q} \quad (14.31)$$

$$\rho = \beta\text{trace}(\mathbf{C}'\mathbf{P}\mathbf{C}) + \beta\rho - \frac{1}{4}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q})'(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) \quad (14.32)$$

The first equation is an algebraic Riccati equation. The other two equations are linear. The optimal control is

$$\mathbf{u}_t = \kappa - \mathbf{H}\mathbf{x}_t \quad (14.33)$$

with

$$\kappa = -\frac{1}{2}(\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{b} + \beta\mathbf{B}'\mathbf{q}) \quad (14.34)$$

$$\mathbf{H} = (\mathbf{R} + \beta\mathbf{B}'\mathbf{P}\mathbf{B})^{-1}(\mathbf{W}' + \beta\mathbf{B}'\mathbf{P}\mathbf{A}) \quad (14.35)$$

From the form of the solution, one can see that the control is *linear* in the state and therefore continuous: every day we trade a little.

In numerical experiments, a Python implementation of a gradient-based algorithm² for the Riccati Equation can solve an instance with a thousand

²H. M. Amman and H. Neudecker. (1997) Numerical solutions of the algebraic matrix Riccati equation. *Journal of Economic Dynamics and Control* **21**, 363-369.

variables within a minute in double precision on a single-core 2.5GHz i7 processor, with 16GB 1.6GHz DDR3 RAM. Real-world applications require a smaller state vector and/or faster hardware and implementations; moreover, double precision is not needed. It is likely that the problem should be solvable within a second.

14.5 Spiked Covariance Matrix: Basic Results

14.5.1 Some Useful Results from Linear Algebra

Spiked Covariance Matrices are the sum of a full-rank sparse (possibly diagonal) matrix and a low-rank matrix. For this class of matrices, there are useful, computationally cheap ways to compute inverse and determinant: the Woodbury-Sherman-Morrison Lemma, and the Matrix Determinant Lemma.

Woodbury-Sherman-Morrison Lemma. Useful to compute the inverse of a matrix (e.g., min-variance portfolio and log-likelihood).

$$(\mathbf{D} + \mathbf{B}\Omega\mathbf{B}')^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{B}(\Omega^{-1} + \mathbf{B}'\mathbf{D}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{D}^{-1} \quad (14.36)$$

Determinant Lemma. Useful in log-likelihood calculations.

$$\det(\mathbf{D} + \mathbf{B}\Omega\mathbf{B}') = \det(\mathbf{D}) \det(\Omega) \det(\Omega^{-1} + \mathbf{B}'\mathbf{D}^{-1}\mathbf{B}) \quad (14.37)$$

14.6 Optimal Trading: The Single-Signal Case

The primitives are:

1. a discount rate $\beta \in (0, 1)$.
2. a sequence of $n \times n$ asset covariance matrices Ω_t . At time t , the variance of a portfolio w is $w'\Omega_tw$.
3. a sequence of $n \times n$ diagonal, positive definite cost matrices C_t . At time t , the cost of trading a basket of assets w is $w'C_tw$.
4. a stationary sequence of signals $\{w_t\}$ in \mathbb{R}^n , such that these expectations exist:

$$E(r_{t+i}w_t) \quad (14.38)$$

$$E(w'_{t+i-j}\Omega_tw_t) \quad (14.39)$$

$$E(w'_{t+i-j}C_tw_t) \quad (14.40)$$

These can be estimated by replacing expectations with empirical averages. The first quantity can be interpreted as an alpha decay curve for the signal. The second one is usually a function of $i - j$ and is the autocovariance of the signals. There is no obvious interpretation for the last one, other than it is a measure of impact-weighted portfolio overlap. We are concerned with the optimal execution of *portfolio-based strategies* where $t \in \mathbb{N}$ is a time index, and w_{ti} is the net value in a numeraire of asset i at epoch t (close of trading day $t - 1$); only signal prior to epoch t are known by the investor.

Any portfolio v_t can be represented as the linear combination of the previous τ signals and an orthogonal component. We choose τ large enough so that $a_i \simeq 0$ for $i \geq \tau$. Moreover, we expect it to be the case that $\beta^\tau \simeq 0$.

$$v_t = \mathbf{W}_t^\tau s_t \quad (14.41)$$

with

$$\mathbf{W}_t^\tau = (w_t | w_{t-1} | \dots | w_{t-\tau+1}) \quad (14.42)$$

The interpretation is that $(s_t)_1$ is the quantity of portfolio w_t held at time t ; $(s_t)_2$ is the quantity of portfolio w_{t-1} held at time t , and so on.

The state of the system is

$$\mathbf{x}_t = \begin{pmatrix} s_t \\ s_{t-1} \end{pmatrix} \quad (14.43)$$

and the state equation is $\mathbf{x}_{t+1} = \mathbf{x}_t + B\mathbf{u}_t$ with

$$B = \begin{pmatrix} I_\tau & 0 \\ 0 & 0 \end{pmatrix} \quad (14.44)$$

Define the matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{R}^{(n+1) \times n}$ as

$$\mathbf{H}_1 = \begin{pmatrix} I_\tau \\ 0 \end{pmatrix} \quad \mathbf{H}_2 = \begin{pmatrix} 0 \\ I_\tau \end{pmatrix} \quad (14.45)$$

The portfolio in two consecutive dates is

$$v_t = \mathbf{W}_t^\tau s_t = \mathbf{W}_t^\tau B \mathbf{x}_t \quad (14.46)$$

$$v_{t-1} = \mathbf{W}_{t-1}^\tau s_{t-1} \quad (14.47)$$

$$v_t - v_{t-1} = \mathbf{W}_t^{\tau+1} (\mathbf{H}_1 s_t - \mathbf{H}_2 s_{t-1}) \quad (14.48)$$

$$= \mathbf{W}_t^{\tau+1} (\mathbf{H}_1 | -\mathbf{H}_2) \mathbf{x}_t \quad (14.49)$$

The transaction cost term is given by

$$(v_t - v_{t-1})' C_t (v_{t+1} - v_t)$$

The variance is

$$v_t' \Omega_t v_t$$

The expected return is

$$E[\mathbf{r}_t' W_t^\tau s_t] = E[\mathbf{r}_t' W_t^\tau]' B s_t = a' \mathbf{x}_t$$

The quadratic terms can be rewritten as function of the state: $(v_t - v_{t-1})' C_t (v_{t+1} - v_t) + v_t' \Omega_t v_t = \mathbf{x}_t' \mathbf{Q} \mathbf{x}_t$ where

$$\mathbf{Q} = B'(W_t^\tau)' \Omega_t W_t^\tau B + \begin{pmatrix} H_1' \\ -H_2' \end{pmatrix} (W_t^{\tau+1})' C_t W_t^{\tau+1} \begin{pmatrix} H_1 \\ -H_2 \end{pmatrix}$$

$$\mathbf{x}_t = x_{t-1} + B u_{t-1} \tag{14.50}$$

$$V = \sum_{t=1}^{\infty} \beta^t (a' \mathbf{x}_t - \mathbf{x}_t' \mathbf{Q} \mathbf{x}_t) \tag{14.51}$$

which is a special case of the LQR.

14.7 Conditioning

The objective of this section is to identify the probability distribution conditional on linear constraints on the random vectors. We have a multivariate random vector taking values in \mathbb{R}^n with mean μ_ξ and covariance matrix Σ_ξ . We are given a matrix $A \in \mathbb{R}^{n \times m}$ and a vector $b \in \mathbb{R}^m$. We assume that the constraints

$$A' \xi = b \tag{14.52}$$

hold with probability one. We denote $\tilde{\xi}$ a random variable distributed as ξ conditional on $A' \xi = b$. We prove that $\tilde{\xi}$ is multivariate normal with mean and covariance matrix equal to

$$\begin{aligned} \mu_{\tilde{\xi}} &= \mu_\xi + \Sigma_\xi A \Sigma_{vv}^{-1} (b - A' \mu_\xi) \\ \Sigma_{\tilde{\xi}} &= \Sigma_\xi - \Sigma_\xi A (A' \Sigma_\xi A)^{-1} A' \Sigma_\xi \end{aligned} \tag{14.53}$$

Let $v := A'\xi$, and $y \in \mathbb{R}^{n+m}$ be a random vector defined as a linear transformation of ξ :

$$y := \begin{pmatrix} \xi \\ v \end{pmatrix} \quad (14.54)$$

Since y is a linear transformation of a multivariate normal, it is also multivariate normal with mean and covariance matrix³

$$\mu_y = \begin{pmatrix} I_n \\ A' \end{pmatrix} \mu_\xi \quad \Omega_y = \begin{pmatrix} \Sigma_\xi & \Sigma_\xi A \\ A' \Sigma_\xi & A' \Sigma_\xi A \end{pmatrix} \quad (14.55)$$

The distribution of ξ conditional on the constraint (14.52) is the same as its distribution conditional on $v = b$. We use the well known distribution of a multivariate Gaussian conditional on a subset of its variables taking a certain value; in this case, v .

$$\begin{aligned} \mu_{\tilde{\xi}} &= \mu_\xi + \Sigma_{\xi v} (A' \Sigma_\xi A)^{-1} (b - \mu_v) \\ \Sigma_{\tilde{\xi}} &= \Omega_{\xi \xi} - \Omega_{\xi v} \Omega_{\xi \xi}^{-1} \Omega_{v \xi} \end{aligned}$$

which in turn gives Equations (14.53). The only case of practical interest is $\mu_\xi = 0$, for which the conditional mean simplifies to

$$\mu_{\tilde{\xi}} = \Sigma_\xi A (A' \Sigma_\xi A)^{-1} b$$

If $A := (I_p | 0_{1:p, 1:(p-m)})$, then $\mu_{\tilde{\xi}} = \Sigma_{1:m, 1:p} \Sigma_{1:p, 1:p}^{-1} b$.

³If $x \sim N(\mu, \Sigma)$ and $y = Ax$, then $y \sim N(A\mu, A\Sigma A')$.

14.8 Three Papers on Backtesting Tests

14.8.1 White (2000)

White's [White \(2000\)](#) "reality check" is a simultaneous test of hypothesis on the n time series. First, the time series are studentized⁴. The null hypothesis⁵ is

$$H_0 : \theta \leq 0 \quad (14.56)$$

To answer the question, White studies the asymptotic properties of $\hat{\theta}(X)$. Under the null hypothesis,

$$P_D(\max_n \hat{\theta}_n(X) \geq x) \leq P_D(\max_n (\hat{\theta}_n(X) - \theta_n(X)) \geq x) \quad (14.57)$$

because all $\theta_n(X) \leq 0$. The latter probability can be estimated via bootstrap methods:

$$P_D(\max_n (\hat{\theta}_n(X) - \theta_n(X)) \geq x) \simeq P_{D^*}(\max_n (\hat{\theta}_n(X^*) - \theta_n(X)) \geq x) \quad (14.58)$$

Let

$$\xi := \max_n (\hat{\theta}_n(X^*) - \theta_n(X)) \quad (14.59)$$

and F_ξ be its distribution. If $\max_n \hat{\theta}_n(X) > F_\xi^{-1}(1 - \alpha)$, with α being the significance level, reject the null.

N	ρ	0.05	0.01	0.001
10	0	0.0620	0.0140	0.004
10	0.9	0.0640	0.0180	0.002
100	0	0.0840	0.0260	0.004
100	0.9	0.0860	0.0260	0.002
1000	0	0.1100	0.0200	0.010
1000	0.9	0.0780	0.0120	0.002

Table 14.1: Simulation results.

Like all the results in this section, the test is only asymptotically valid when $T \rightarrow \infty$ and N is constant. However, the results can be inaccurate in the

⁴This step is not advocated by the author, but is common in all subsequent papers. It puts all the observation on the same scale.

⁵White subtracts from the observed statistic for every subject the statistic of a "benchmark" in the same period. We can interpret x_{tn} as this difference.

case of $N \gg T$. Consider the following: $T = 100$, $N \in \{10, 100, 1000\}$, and $x_{nt} \sim N(0, 1)$, with $\text{cor}(x_{mt}, x_{nt}) = \rho$, and $\rho \in \{0, 0.9\}$. For each pair N, ρ we generate 1000 variates. The last three columns show the percentage of false positives for $\alpha = 0.01, 0.05, 0.001$ respectively. The results are displayed in Table 14.1. One can see that the fraction of false positives exceeds α when $N \rightarrow \infty$, and the correlation among variates is low. For high values of ρ the effective number of strategies is reduced.

14.8.2 Romano and Wolf (2005)

Romano and Wolf [Romano and Wolf \(2005\)](#) consider a sequence of tests $n = 1, \dots, N$:

$$H_{0n} : \theta_n \leq 0 \quad (14.60)$$

$$H_{an} : \theta_n > 0 \quad (14.61)$$

Let S_0 be the set of strategies for which the null is true and S_a the set of strategies for which H_{an} is true. The goal of the authors is to identify S_a , while establishing an upper bound α for the family-wise error rate (FWE), defined as the probability of rejecting *at least* one of the true nulls; i.e., the probability of having one “false positive” or more:

$$\text{FWE}_D := P_D(\text{reject at least one } H_{0n}, n \in S_0) \quad (14.62)$$

This condition may be trivially satisfied by accepting all the nulls, but that is so conservative as to be useless. The authors choose to maximize the percentage of false null hypotheses that are correctly rejected. Let denote \hat{S} be the set of strategies identified by the decision rule. We restate these conditions as follows

$$\max E_D(\hat{S} \cap S_a) \quad (14.63)$$

$$\text{s.t. } P_D(\hat{S} \cap S_0 \neq \emptyset) \leq \alpha \quad (14.64)$$

The steps of the method are:

1. Set $\hat{S} = \emptyset$ and studentize the columns of X .
2. Let F_ξ be the bootstrap distribution of

$$\xi := \max_n (\hat{\theta}_n(X^*) - \hat{\theta}_n(X)) \quad (14.65)$$

Add to \hat{S} all i such that $\hat{\theta}_i(X) > F_\xi^{-1}(1 - \alpha)$ and remove these strategies from X .

3. If no strategies are removed, stop. Otherwise repeat step 2.

Under this procedure, the authors prove that the final set \hat{S}_a satisfies

$$\limsup_T P_D(S_0 \cap \hat{S}_a \neq \emptyset) \leq \alpha \quad (14.66)$$

$$\lim_T P(i \in \hat{S}_a) = 1 \quad \forall i \in S_a \quad (14.67)$$

The FWE control approach in Romano-Wolf has a serious drawback, in that can be too conservative when the number of null hypotheses is very large. It is essential that any procedure work well for large (but not infinite) sets of hypotheses, so this drawback limits the usefulness of the test. An alternative approach (not pursued by the authors) is to control the false discovery rate (FDR) expected ratio of false rejections divided by the total number of rejections.

14.8.3 Hansen, Lunde and Nason (2011)

Hansen *et al.*[Hansen et al. \(2011\)](#) deal with a different question: given a data matrix X , we would like to identify the best subset of superior strategies S^* , for which

$$\theta_i \geq \theta_j \quad \forall i \in S^*, j \in S \quad (14.68)$$

To achieve this goal, the authors devise two steps that are applied iteratively. The first is an *equivalence test*, which tests whether a candidate set satisfies the null hypothesis of Eq.(14.68). If the equivalence test rejects the null, then an *elimination rule* is applied, which discards one strategy from \hat{S} ; otherwise it stops. The procedure converges to a limit set \hat{S}^* and the authors show that

$$\liminf_T P(S^* \subset \hat{S}^*) \geq 1 - \alpha \quad (14.69)$$

$$\lim_T P(i \in \hat{S}^*) = 0 \quad \forall i \notin S^* \quad (14.70)$$

1. Set $\hat{S} := S$.
2. For $i, j \in \hat{S}$, define the performance difference between each strategy and the rest of the set; then studentize it and finally compute the maximal

statistic:

$$\delta_i = N^{-1} \sum_{k \in \hat{S}} \hat{\theta}_k - \hat{\theta}_i \quad (\text{demeaning}) \quad (14.71)$$

$$t_i = \frac{\delta_i}{\sqrt{\hat{\sigma}^2(\delta_i)}} \quad (\text{studentization}) \quad (14.72)$$

$$T_{\max} = \max_{i \in \hat{S}} t_i \quad (\text{maximal statistic}) \quad (14.73)$$

Let $Z = (\delta_1, \dots, \delta_m)$. Via bootstrap, estimate the distribution of \hat{F} of T_{\max} . If $T_{\max} < \hat{F}^{-1}(1 - \alpha)$ the equivalence test passes and the procedure stops with \hat{S}^* . If not, remove from \hat{S} $\arg \max_{i \in \hat{S}} t_i$ and repeat the step.

Hansen *et al.* results are symmetrical to those in Romano and Wolf's paper. The former avoid false positives while keeping under control the probability of a false negative, whereas the latter avoid false negatives while keeping under control the probability of a false positive. To clarify the connection, we denote $S_0 = S - S^*$, $\hat{S}_0 = S - \hat{S}^*$ and rewrite them

$$\limsup_T P(S^* \cap \hat{S}_0 \neq \emptyset) \leq \alpha \quad (14.74)$$

$$\lim_T P(i \in \hat{S}^*) = 0 \quad \forall i \in S_0 \quad (14.75)$$

In RW, we include in the final set any “null strategy” with low probability, and we include all the “good” strategies. In HLN, we exclude from the final set any “good” strategy with low probability, and exclude all the “null” strategies. Unlike RW, HLN does not suffer from excess conservatism in the case of many null strategies. It does in the scenario where $|S^*| \rightarrow \infty$ as $N \rightarrow \infty$, which seems implausible, as the number of “best” models should stay constant, and be possibly equal to 1.

Chapter 15

Bibliography

15.1 Bibliography

A. Mas-Colell, M. D. W. and Green, J. R. (1995). *Microeconomic theory*. Oxford University Press.

Agarwal, V. and Naik, N. Y. (2004). Risks and portfolio decisions involving hedge funds. *The Review of Financial Studies*, 17(1):63–98.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, 18(2):351–416.

Algoet, P. H. and Cover, T. M. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Annals of Probability*, 16(2):876–898.

Almgren, R. (2009). Execution costs. In Cont, R., editor, *Encyclopedia of Quantitative Finance*. Wiley.

Almgren, R., Thum, C., Hauptmann, H. L., and Li., H. (2005). Equity market impact. *Risk*, pages 57–62.

Amihud, Y., Mendelson, H., and Pedersen, L. H. (2012). *Market liquidity*. Cambridge University Press.

Andersen, T. G. and Benzoni, L. (2009). Realized volatility. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 555–575. Springer.

- Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2006). Volatility and correlation forecasting. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 15, pages 777–878. Elsevier.
- Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2013). Financial risk measurement for financial risk management. In Constantinides, G. M., Harris, M., and Stulz, R. M., editors, *Handbook of the Economics of Finance*, volume 2, part B, chapter 17, pages 1127–1220. Elsevier.
- Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors (2009). *Handbook of Financial Time Series*. Springer.
- Anderson, T. W. (1963). Theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.
- Arnott, R., Harvey, C. R., and Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science*, 1(1):64–74.
- Asness, C. S., Frazzini, A., and Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, 24(34-112).
- Bacidore, J. M. (2020). *Algorithmic Trading: A Practitioner’s Guide*. TBG Press.
- Bacon, C. R. (2005). *Practical Portfolio Performance Measurement and Attribution*. Wiley.
- Bai, Z. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):447–474.
- Bai, Z. and Yao, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Annales de l’Institut Henri Poincaré*, 44(3):447–474.
- Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697.
- Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408.

- Barber, B. M. and Odean, T. (2013). *Handbook of the Economics of Finance*, volume 2(B), chapter The Behavior of Individual Investors, pages 1533–1570. Elsevier.
- Barenblatt, G. I. (2003). *Scaling*. Cambridge University Press.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, 12:C1–C32.
- Barndorff-Nielsen, O. E. and Shephard, N. (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, 17(5):457–477.
- Baron, D. P. (1977). On the utility theoretic foundations of mean-variance analysis. *Journal of Finance*, 32(5):1683–1697.
- Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006). *Nonlinear programming: theory and algorithms*. Wiley, 3rd edition.
- Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521.
- Bender, J., Briand, R., Melas, D., and Subramanian, R. A. (2013). Foundations of factor investing. Technical report, MSCI Research Insight.
- Bender, J., J.-H, L., and Stefek, D. (2009). Refining portfolio construction when alphas and risk factors are misaligned. Technical report, MSCI Research Insight.
- Berben, R.-P. and Jansen, W. J. (2005). Comovement in international equity markets: A sectoral view. *Journal of International Money and Finance*, 24(5):832–857.
- Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, (70-83).

- Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.
- Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, 2nd edition.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Black, F., Jensen, M. C., and Scholes, M. (1972). The capital asset pricing model: Some empirical tests. In Jensen, M. C., editor, *Studies in the Theory of Capital Markets*, pages 79–121. Praeger Publishing Co.
- Bluman, G. W. and Kumei, S. (1989). *Symmetries and Differential Equations*. Springer.
- Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194.
- Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics*, 72(3):498–505.
- Bouchaud, J. (2010). Price impact. In Cont, R., editor, *Encyclopedia of Quantitative Finance*. Wiley.
- Bouchaud, J., Bonart, J., Donier, J., and Gould, M. (2018). *Trades, Quotes and Prices*. Cambridge University Press.
- Bouchaud, J. and Potters, M. (2020). *A First Course in Random Matrix Theory*. Cambridge University Press.
- Bouchaud, J.-P., Potters, M., and Aguilar, J.-P. (1997). Missing information and asset allocation. cond-mat/9707042.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Boyd, S., Busseti, E., Diamond, S., Kahn, R. N., Koh, K., Nystrup, P., and Speth, J. (2016). Multi-period trading via convex optimization. *Foundations and Trends in Optimization*, 3(1):1–76.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

- Breiman, L. (1961). Optimal gambling systems for favorable games. In Neyman, J., editor, *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 65–78. University of California Press.
- Brooks, R. and Del Negro, M. (2005). Country versus region effects in international stock returns. *Journal of Portfolio Management*, 31(4):67–72.
- Brownlees, C., Engle, R., and Kelly, B. (2011). A practical guide to volatility forecasting through calm and storm. *Journal of Risk*, 14(2):3–22.
- Bryzgalova, S., Huang, J., and Julliard, C. (2022). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance*, 78(1).
- Bun, J., Bouchaud, J., and Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109.
- Buraczewski, D., Damek, E., and Mikosch, T. (2016). *Stochastic models with power-law tails*. Springer.
- Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10:1–59.
- Calvino, I. (1999). *Six Memos for the Next Millennium*. Harvard University Press.
- Cattel, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276.
- Cavaglia, S., Brightman, C., and Aked, M. (2000). The increasing importance of industry factors. *Financial Analysts Journal*, 56(5):41–54.
- Ceria, S., Saxena, A., and Stubbs, R. A. (2012). Factor alignment problems and quantitative portfolio management. *Journal of Portfolio Management*, 28(2):29–43.
- Cerqueira, V., Torgo, L., and Soares, C. (2023). Model selection for time series forecasting an empirical analysis of multiple estimators. *Neural Processing Letters*, 55:10073–10091.
- Chamberlain, G. (1983). A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory*, 29:184–201.

- Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis of large asset markets. *Econometrica*, 51(5):1281–1305.
- Chen, A. Y. and Velikov, M. (2019). Accounting for the anomaly zoo: a trading cost perspective. *working paper*.
- Chincarini, L. B. and Kim, D. (2007). Another look at the information ratio. *Journal of Asset Management*, 8(5):284–295.
- Chincarini, L. B. and Kim, D. (2022). *Quantitative Equity Portfolio Management*. McGraw Hill, 2nd edition.
- Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. *Journal of Finance*, 74(1):449–492.
- Chinco, A. and Sammon, M. (2023). The passive-ownership share is double what you think it is.
- Chopra, V. K. and W.Ziemba (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11.
- Cižek, P., Härdle, W. K., and Weron, R. (2011). *Statistical Tools for Finance and Insurance*. Springer, 2nd edition.
- Clarke, R., de Silva, H., and Thorley, S. (2002). Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, 58(5):48–66.
- Cochrane, J. H. (2005). *Asset Pricing*. Princeton University Press.
- Cochrane, J. H. (2008). The dog that did not bark: a defense of stock return predictability. *The Review of Financial Studies*, 21(4):1533–1575.
- Cohen, K. J., Hawawini, G. A., Maier, S. F., Schwartz, R. A., and Whitcomb, D. K. (1983). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics*, 12:263–278.
- Connor, G., Goldberg, L. R., and Korajczyk, R. A. (2010). *Portfolio risk analysis*. Princeton University Press.
- Connor, G. and Korajczyk, R. A. (2010). Factor models of asset returns. In Cont, R., editor, *Encyclopedia of Quantitative Finance*. Wiley.

- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236.
- Cvitanić, J. and Karatsas, I. (1995). On portfolio optimization under “drawdown” constraints. *IMA Volumes in Mathematics and Applications*, 65:35–46.
- Davis, R. A. and Mikosch, T. (2009). Extreme value theory for GARCH processes. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 187–200. Springer.
- De Finetti, B. (1940). Il problema dei pieni. *Giornale dell'Istituto Italiano degli Attuari*, 11:1–88.
- DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Addison-Wesley, 4th edition.
- DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009a). A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.
- DeMiguel, V., Garlappi, L., and Uppal, R. (2009b). Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953.
- SIAM Review, 41(1):45–76.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*.

Ding, J. and Meade, N. (2010). Forecasting accuracy of stochastic volatility, garch and ewma models under different volatility scenarios. *Applied Financial Economics*, (10):1742–1778.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(4):1742–1778.

Dubins, L. and Savage, L. J. (1965). *How to gamble if you must: inequalities for stochastic processes*. McGraw-Hill.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

- Eckart, G. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.
- El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6):2757–2790.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(987-1007).
- Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5(1):1–50.
- Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. (2007). Robust portfolio optimization. *The Journal of Portfolio Management*, 33(3):40–48.
- Fama, E. F. and French, K. R. (1993). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.
- Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.
- Fan, J., Li, R., Zhang, C., and Zou, H. (2020). *Statistical Foundations of Data Science*. CRC Press.
- Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 16:C1–C32.
- Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.
- Farmer, S. (1971). An investigation into the results of principal component analysis of data derived from random numbers. *Statistician*, 20(4):63–72.
- Feng, G., Giglio, S., and D.Xiu (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*.
- Ferré, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Computational Statistics & Data Analysis*, 19(19):669–682.
- Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1):1–25.

- French, K. R., Schwert, G. W., and Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, 19(1):3–29.
- Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5):2326–2377.
- Galilei, G. (1623). *Il Saggiatore*.
- Gârleanu, N. B. and Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *Journal of Finance*, 68(6):2309–2340.
- Gatheral, J. (2016). Three models of market impact.
- Gelman, A., Hill, J., and Vehtari, A. (2022). *Regression and Other Stories*. Cambridge University Press.
- Gerakos, J. and Linnainmaa, J. T. (2021). Asset managers: Institutional performance and factor exposures. *Journal of Finance*, 76(4):2035–2075.
- Gibbins, J. C. (2011). *Dimensional Analysis*. Springer.
- Glynn, P. W. (1990). Diffusion approximations. In Heyman, D. and Sobel, M. J., editors, *Handbooks on OR and MS*, volume 2, chapter 4, pages 145–198. Elsevier.
- Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*. Johns Hopkins University Press, 4th edition.
- Granger, C. W. J. and Ding, Z. (1995). Some properties of absolute return: an alternative measure of risk. *Annales d' Economie et de Statistique*, (40):67–91.
- Green, J., Hand, J. R. M., and Zhang, X. F. (2013). The supraview of return predictive signals. *Review of Accounting Studies*, 18:692–730.
- Grinold, R. C. (1989). The fundamental law of active management. *Journal of Portfolio Management*, 15(3):30–37.
- Grinold, R. C. and Kahn, R. N. (1999). *Active Portfolio Management*. McGraw-Hill Education, 2nd edition.
- Grossman, S. J. and Zhou, Z. (1993). Optimal investment strategies for controlling drawdowns. *Mathematical Finance*, 3(3):241–276.

- Guéant, O. (2016). *The Financial Mathematics of Market Liquidity*. Chapman & Hall/CRC.
- Hansen, B. (2022). *Econometrics*. Princeton University Press.
- Hansen, L. P. and Sargent, T. J. (2008). *Robustness*. Princeton University Press.
- Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.
- Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a *garch*(1, 1). *Journal of Applied Econometrics*, 20(7):873–889.
- Hansen, P. R. and Lunde, A. (2006a). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2):97–121.
- Hansen, P. R. and Lunde, A. (2006b). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics*, 24(2):127–161.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.
- Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer, 2nd edition.
- Harris, L. (2003). *Trading and Exchanges*. Oxford University Press.
- Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.
- Harvey, A. C. and Shephard, N. (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business & Economic Statistics*, 14(4):429–424.
- Harvey, C. H. and Liu, Y. (2020). A census of the factor zoo. *preprint*.
- Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–61.
- Hasbrouck, J. (2007). *Empirical Market Microstructure*. Oxford University Press.

- Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, 2nd edition.
- He, C. and Teräsvirta, T. (1999). Fourth moment structure of the GARCH(p, q) process. *Econometric Theory*, 15(6):824–846.
- Heston, S. L. and Rouwenhorst, K. G. (1994). Does industrial structure explain the benefits of industrial diversification? *Journal of Financial Economics*, 36:3–27.
- Heston, S. L. and Rouwenhorst, K. G. (1995). Industry and country effects in international stock returns. *The Journal of Portfolio Management*, 21(3):53–58.
- Horn, R. A. and Johnson, C. (2012). *Matrix Analysis*. Cambridge University Press, 2nd edition.
- Huang, C.-F. and Litzenberger, R. H. (1988). *Foundations for financial economics*. Prentice-Hall.
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.
- Huberman, G., Kandel, S., and Stambaugh, R. F. (1987). Mimicking portfolios and exact arbitrage pricing. *Journal of Finance*, 42(1):1–9.
- Huberman, G. and Stanzl, W. (2004). Price manipulation and quasi-arbitrage. *Econometrica*, 74(4):1247–1276.
- Ilmanen, A. (2011). *Expected Returns*. Wiley.
- Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8):696–701.
- Isichenko, M. (2021). *Quantitative Portfolio Management*. Wiley.
- Jacobs, H. (2015). What explains the dynamics of 100 anomalies? *Journal of Banking and Finance*, 57:65–85.
- Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58(4):1651–1683.

- Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance*, 48(1):65–91.
- Jegadeesh, N. and Titman, S. (2011). Momentum. *Annual Review of Financial Economics*, 3(1):493–509.
- Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson, 6th edition.
- Johnstone, D. J. and Lindley, D. V. (2011). Elementary proof that mean–variance implies quadratic utility. *Theory and Decision*, 70(2):149–155.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327.
- Johnstone, I. M. and Paul, D. (2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292.
- Jolliffe, I. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202.
- Jolliffe, I. T. (2010). *Principal Component Analysis*. Springer, 2nd edition.
- Kagan, L. and Tian, M. (2017). Firm characteristics and empirical factor models: a data-mining experiment.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.
- Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108.
- Kelley, R. L. (1955). *General Topology*. Van Nostrand.
- Kesten, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131(207-248).
- Kim, W. C., Kim, J. H., and Fabozzi, F. J. (2014a). Deciphering robust portfolios. *Journal of Banking and Finance*, 45:1–8.

- Kim, W. C., Kim, M. J., Kim, J. H., and Fabozzi, F. J. (2014b). Robust portfolios that do not tilt factor exposure. *European Journal of Operational Research*, 234(411-421).
- Kolanovic, M. and Krishnamachari, R. T. (2017). Big data and ai strategies: Machine learning and alternative data approach to investing. Technical report, J. P. Morgan.
- Kolm, P. and Westray, N. (2021). A principled approach to clean-up costs in algo trading. *Risk Magazine*.
- Kolm, P. N., Tütüncü, R., and Fabozzi, F. J. (2014). 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234:356–371.
- Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.
- Kreps, D. M. (1988). *Notes on the theory of choice*. Routledge.
- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 54(6):1315–1335.
- Lai, T. L., Xing, H., and Chen, Z. (2011). Mean-variance portfolio optimization when means and covariances are unknown. *Annals of Applied Statistics*, 5(2A):798–823.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford Science Publications.
- Ledoit, O. and Wolf, M. (2003a). Honey, I shrunk the sample covariance matrix: Problems in mean-variance optimization. *Journal of Portfolio Management*, 30:110–119.
- Ledoit, O. and Wolf, M. (2003b). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621.
- Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.
- Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

- Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360–384.
- Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, 48(5):3043–3065.
- Lee, J. and Stefek, D. (2008). Do risk factors eat alphas? *Journal of Portfolio Management*, 34(4):12–25.
- Li, S. (2021). Should passive investors actively manage their trades?
- Lindner, A. M. (2009). Stationarity, mixing, distributional properties and moments of garch(p, q)-processes. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 43–70. Springer.
- Litterman, R. and Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income*, 1(1):54–61.
- Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1):293–311.
- Lo, A. W. (2002). The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52.
- López de Prado, M. (2020). *Machine Learning for Asset Managers*. Cambridge University Press.
- Luenberger, D. G. (1969). *Optimization by vector space methods*. Wiley.
- Luenberger, D. G. (1993). A preference foundation for log mean-variance criteria in portfolio choice problems. *Journal of Dynamics and Control*, 17(5-6):887–906.
- Luenberger, D. G. (2013). *Investment Science*. Oxford University Press, 2nd edition.
- Luenberger, D. G. and Ye, Y. (2008). *Linear and Nonlinear Programming*. Springer, 3rd edition.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

- Maccheroni, F., Marinacci, M., and Ruffino, D. (2013). Alpha as ambiguity: robust mean-variance portfolio analysis. *Econometrica*, 81(3):1075–1113.
- MacKinlay, A. C. (1995). Multifactor models do not explain deviations from capm. *Journal of Financial Economics*, 38(1):3–28.
- MacLean, L. C., Sanegre, R., Zhao, Y., and Ziembra, W. T. (2004). Capital growth with security. *Journal of Economic Dynamics and Control*, 28:937–954.
- MacLean, L. C., Thorp, E. O., and Ziembra, W. T. (2010). Good and bad properties of the kelly criterion. In MacLean, L. C., Thorp, E. O., and Ziembra, W. T., editors, *The Kelly Capital Growth Investment Criterion*, pages 563–574. World Scientific.
- MacLean, L. C., Ziembra, W. T., and Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, 38(11):1562–85.
- Mahajan, S. (2014). *The Art of Insight in Science and Engineering*. MIT Press.
- Malkiel, B. (1987). *The New Palgrave Dictionary of Economics*, chapter Efficient Market Hypothesis, pages 1–7. Palgrave MacMillan.
- Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2):270–296.
- Mancini, C. (2011). The speed of convergence of the threshold estimator of integrated variance. *Stochastic Processes and their Applications*, 121(4):845–855.
- Markowitz, H. (2014). Mean–variance approximations to expected utility. *European Journal of Operational Research*, 234:346–355.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Basil Blackwell, 2nd edition.
- Mestre, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11):5113–5129.

- Michaud, R. O. (1989). The markowitz optimization enigma: Is ‘optimized’ optimal? *Financial Analysts Journal*, 45(1):31–42.
- Mikosch, T. and Stărică, C. (2000). Limit theory for the sample autocorrelations and extremes of a garch(1, 1) process. *Annals of Statistics*, 28(5):1427–1451.
- Miralles Marcelo, J. L., Miralles Quirós, J. L., and Martins, J. L. (2013). The role of country and industry factors during volatile times. *Journal of International Financial Markets, Institutions and Money*, 26:273–290.
- Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59.
- Models, G. (2004). M. i. jordan. *Statistical Science*, 19(1):140–155.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. The MIT Press, 2nd edition.
- Murphy, K. P. (2012). *Machine Learning*. MIT Press.
- Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290):299–306.
- Narang, R. K. (1990). *Inside the Black Box*. Wiley, 2nd edition.
- Nelson, D. R. (1990). Stationary and persistence in the GARCH(1, 1) model. *Econometric Theory*, 6:318–334.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Novick, B. (2017). Index investing supports vibrant capital markets. *Blackrock Viewpoint*.
- Obizhaeva, A. A. and Wang, J. (2013). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32.
- Olivares-Nadal, A. V. and DeMiguel, V. (2018). Technical note—a robust perspective on transaction costs in portfolio optimization. *Operations Research*, 66(3):733–739.
- Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016.

- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Ottusák, G. and Vajda, I. (2007). An asymptotic analysis of the mean-variance portfolio selection. *Statistics & Decisions*, 25:63–88.
- Pardo, R. (2007). *The Evaluation and Optimization of Trading Strategies*. Wiley, 2nd edition.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160:246–256.
- Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 801–838. Springer.
- Paul, D. (2017). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(1617-1642).
- Pav, S. E. (2023). *The Sharpe Ratio*. Chapman & Hall/CRC.
- Pedersen, L. H. (2015). *Efficiently Inefficient: How Smart Money Invests and Market Prices are Determined*. Princeton University Press.
- Pedersen, L. H., Babu, A., and Levine, A. (2021). Enhanced portfolio optimization. *Financial Analysts Journal*, 77(2):124–151.
- Podolskij, M. and Vetter, M. (2009). Bipower-type estimation in a noisy diffusion setting. *Stochastic Processes and their Applications*, 119(9):2803–2831.
- Pohl, M., Ristig, A., and Schachmayer, W. (2017). The amazing power of dimensional analysis: quantifying market impact. *Market Microstructure and Liquidity*, 3(4):1850004.
- Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. Wiley.
- Puchkov, A. V., Stefek, D., and Davis, M. (2005). Sources of return in global investing. *Journal of Portfolio Management*, 31(2):12–21.
- Qian, E. E., Hua, R. H., and Sorensen, E. H. (2007). *Quantitative Equity Portfolio Management*. Chapman & Hall/CRC.
- Ratliff-Crain, E., Oort, C. M. V., Bagrow, J., Koehler, M. T. K., and Tivnan, B. F. (2023). Revisiting stylized facts for modern stock markets.

- Rencher, A. C. and Christensen, W. F. (2012). *Methods of multivariate analysis*. Wiley.
- Robert, C. P. (2007). *The Bayesian Choice*. Springer, 2nd edition.
- Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, 39(4):1127–39.
- Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.
- Ross, S. M. (2023). *Introduction to Probability Models*. Academic Press, 13th edition.
- Ruppert, D. and Matteson, D. S. (2015). *Statistics and Data Analysis for Financial Engineering*. Springer, 2nd edition.
- Saxena, A. and Stubbs, R. A. (2013). The alpha alignment factor: a solution to the underestimation of risk for optimized active portfolios. *Journal of Risk*, 15(3):3–37.
- Schoettle, K. and Werner, R. (2009). Robustness properties of mean-variance portfolios. *Optimization*, 58(6):646–663.
- Scholes, M. and Williams, J. (1977). Estimating beta from nonsynchronous data. *Journal of Financial Economics*, 5(3):309–327.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442.
- Sharpe, W. F. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37.
- Sharpe, W. F. (1966). Equilibrium in a capital asset market. *Econometrica*, 34(4):768–783.
- Shen, D., Shen, H., Zhu, H., and Marron, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26:1747–1770.

- Shephard, P. G. (2009). Second order risk.
- Simon, D. (2006). *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. Wiley.
- Skillicorn, D. (2007). *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. Chapman & Hall/CRC.
- Stevens, W. (1990). *The Collected Poems*. Vintage.
- Stock, J. H. and Watson, W. M. (2016). *Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics*, volume 2A, chapter 8, pages 415–525.
- Strang, G. (2019). *Linear Algebra and Learning from Data*. Wellesley - Cambridge Press.
- Stubbs, R. A. and Vance, P. (2005). Computing return estimation error matrices for robust optimization. Technical Report 1, Axioma Research Paper.
- Suzuki, S. (1970). *Zen Mind, Beginner's Mind*. Weatherhill.
- Taylor, S. J. (1986). *Modelling Financial Time Series*. Wiley.
- Taylor, S. J. (2007). *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press.
- Teräsvirta, T. (2009a). An introduction to univariate GARCH models. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 17–42. Springer.
- Teräsvirta, T. (2009b). Multivariate GARCH models. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*. Springer.
- Thorp, E. O. (2006). The kelly criterion in blackjack sports betting, and the stock market. In Zenios, S. and Ziembka, W., editors, *Handbook of asset and liability management*, volume 1. Elsevier.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B*, 61(3):611–622.
- Trefethen, L. and Bau, D. (1997). *Numerical linear algebra*. SIAM.

- Tsay, R. S. (2010). *Analysis of financial time series*. Wiley, 3rd edition.
- V., A. D., Martin-Utrera, and Nogales., F. (2013). Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of Banking and Finance*, 37(8):3018–3034.
- Vause, N. (2010). Counterparty risk and contract volumes in the credit default swap market. *BIS Quarterly Review*.
- Velu, R., Hardy, M., and Nehren, D. (2020). *Algorithmic Trading and Quantitative Strategies*. CRC Press.
- Vershinin, R. (2018). *High-dimensional probability*. Cambridge University Press.
- Wahba, G. (1965). A least squares estimate of satellite attitude. *SIAM Review*, 7(3):384.
- Wang, S., Luo, Y., Alvarez, M., Jussa, J., Wang, A., and Rohal, G. (2014). Seven sins of quantitative investing. Technical report, Deutsche Bank.
- Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of Statistics*, 45(3):1342–1374.
- Wang, Z. (2005). A shrinkage approach to model uncertainty and asset allocation. *Review of Financial Studies*, 18(2):673–705.
- Wasserman, L. (2004). *All of Statistics*. Springer.
- Webster, K. T. (2023). *Handbook of Price Impact Modeling*. Chapman & Hall/CRC.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.
- Whittle, P. (1996). *Optimal Control. Basics and Beyond*. Wiley.
- Williams, J. B. (1936). Speculation and the carryover. *The Quarterly Journal of Economics*, 50(3):436–455.
- Yao, J., Zheng, S., and Bai, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press.

- Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.
- Zivot, E. (2009). Practical issues in the analysis of univariate garch models. In Andersen, T. G., Davis, R. A., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*, pages 113–155. Springer.
- Zivot, E. and Wang, J. (2003). *Modeling Financial Time Series with S-Plus*. Springer.

Index

- R^2 , see Coefficient of Determination 114
- Active Managers, 27
- Aggregational Gaussianity, 44
- Alpha, 74, 86
- Intraday, 303
 - Orthogonal, 78, 111, 274
 - Spanned, 77, 78, 111, 274
- Alternate Trading System, 257
- American Depository Receipt (ADR), 228
- Annihilator Matrix, 94
- Arbitrageurs, 28
- Asset Allocators, 28
- Assets Under Management, 35
- Autocorrelation
 - Absolute univariate returns, 44
 - Univariate Returns, 42
- Backtesting
 - Cross-Validation, 162
 - Protocol, 157
 - Rademacher Anti-Serum (against backtesting bites), 162
 - Walk Forward, 162
 - Walk-Foward, 161
- Beginner's Mind, 10
- Beta, 85
- Betas
 - Adjusted-Dollar, 310
- Bid-Ask Spread, 41
- Bloomberg, 24
- Bonds, 21
- Broker-Dealers, 26
- Brokers, 24
- Buy Side, 23, 26
- CA, *see* Closing Auction
- Cantelli's inequality, 118
- Capital Asset Pricing Model, 117
- CAPM, *see* Capital Asset Pricing Model
- Cash Equivalents, 28
- CDS, *see also* Credit Default Swaps
- Characteristics, 33
- Child Order, 257
- CHM, *see* Conditional Heteroscedastic Models
- Clearing, 25
- Closing Auction (CA), 184
- Clustering
 - K-Means, 235
- Coefficient of Determination, 114, 173
- Complementary slackness conditions, 124
- Conditional Heteroscedastic Models, 46
- Constrained Regression, 191
- Constraint
 - Long-Only, 134
 - Market Beta, 135
 - on Volatility, 106
 - Portfolio Turnover, 135

- Tracking Error, 136
- constraint
 - Pricing Out, 108
 - Constraint Qualification, 123
- Constraints
 - Quadratic, 135
- Consumer Price Index, 33
- Correlation
 - Thresholding, 199
 - Cosine Similarity, 231, 242
 - Covariance Matrix, 15
 - Autocorrelation Correction, 197
 - Empirical, 87
 - Credit Default Swaps, 21
 - Cross-Sectional Empirical Average, 321
 - Cross-Sectional Empirical Covariance, 321
 - Cross-Validation, 158
 - Currency
 - Base, 205
 - Quote, 205
 - Dark Pools, 23
 - Data
 - Categorical, 185
 - Structured, 33
 - Unstructured, 33
 - Data Leakage, 153, 154
 - Dataset
 - Training, 158
 - Validation, 158
 - Dealers, 24
 - Inventory, 24
 - Dealerweb, 24
 - Deleveraging Spirals, 31
 - Determinant Lemma, 329
 - Diversification, 315
 - Dividend, 38
 - Dual Traders, *see* Broker-Dealers
 - Duality
 - Strong, 123
 - Weak, 123
 - Efficient Market, 29
 - Eigenfactors, 214
 - Eigenvalues
 - Bulk, 220
 - Spike, 220
 - Eigenvectors, 214
 - Equities, 21
 - Estimation Universe, 184, 186
 - ETF, *see* Exchange-Traded Funds
 - EWMA, *see also* Exponentially Weighted Moving Average
 - Exchange Rate, 205
 - Indirect, 205
 - Exchange-Traded Funds, 21
 - Exchanges, 22
 - Exponentially Weighted Moving Average, 56
 - Exposure
 - to Factors, 83
 - to Systematic Risk, 34
 - Factor Model, 72
 - Approximate, 72
 - as a superposition of loadings, 76
 - Characteristic, 88, 183
 - Definition, 72
 - Graphical Model, 75
 - Idiosyncratic Component, 72
 - Interpretation, 74
 - Interpretations, 73
 - Macroeconomic, 88
 - Projections, 81
 - Pushout, 82
 - Rotation, 213

- Statistical, 88, 211
- Strict, 72
- Systematic Component, 72
- Transformations, 73
- Types, 88
- Uses, 73
- Factor Models
 - Currency Rebasing, 205
- Factor returns, 72
- Factors
 - Factor-Mimicking Portfolios, 111
 - Mimicking Portfolios, 111
 - Unpriced, 274
- Feasible region, 122
- First Order Necessary Conditions, 107
- Flow predictability, 31
- FONC, *see* First Order Necessary Conditions 107
- Frisch-Waugh-Lovell Theorem, 93
- Front running, 31
- Front-Running, 26
- Fundamental Law of Active Management, 115
- Futures, 21
- GARCH models, 46–48
- General Autoregressive Conditional Heteroskedastic, *see also* GARCH models 46
- Global Depository Receipt (GDR), 228
- GMV, *see also* Gross Market Value, *see also* Gross Market Value
- Graphical model, 73
- Gross Market Value, 35
- Hat Matrix, 91
- Heavy tails, 42
- Hedge Funds, 28
- Hedging, 34
 - Shrinkage Factor, 277
- Herfindahl Index, 279, 315
- heteroskedastic noise, 92
- HFT, *see also* High-Frequency Trader
- High-Frequency Trader, 303
- Idempotent Operator, 82
- Idiosyncratic returns, 72
- iid rv, 16
- Index
 - Rebalancing, 31
 - Reconstitution, 31
- Index Huggers, 28
- Information Coefficient, 113, 160, 163
- Information Ratio, 115, 116, 314, 320
 - Empirical, 319
- Information Set, 29
- Interest Rate Swaps, 21
- IR, *see* Information Ratio
- IRS, *see* Interest Rate Swaps
- Kalman Filter, 42, 56, 62
 - Optimal Gain, 64
- Karush-Kuhn-Tucker conditions, 124
- Kelly Criterion, 105
- Kolmogorov-Smirnov distance, 49
- Lagrange multiplier, 108
- Lagrange multipliers, 122
- Lagrangian, 122
- Leverage, 35
- Leverage Effect, 44
- Likelihood, 90
- Limit Order Book, 41, 184
- Limit-Order Book, 22
- Linear Regression, 89
 - As Projection, 91
 - Decomposition, 93
- Frisch-Waugh-Lovell Theorem, 93

- Random Design, 188
- Linear State-Space Models, 38, 56, 59
- Linear-Quadratic Regulator, 326
- Liquidity, 21, 30
- Loadings
 - Definition, 72
 - Orthonormal, 80
 - Z-scored, 81
 - Z-scoring, 81
- LOB, *see also* Limit Order Book 22,
see Limit Order Book
- Mahalanobis distance, 180
- Market Efficiency, 29
- Market Impact, 34, 256
- Market Makers, 24
- Market orders, 24
- Marketable Order, 258
- Meta-Order, *see also* Parent Order
- MIFID II, 26
- Moore-Penrose Pseudoinvers, 15
- Net Market Value, 39
- NMV, *see also* Net Market Value
- Norm
 - n -Norm, 15
 - n -norm, 15
 - Frobenius, 212
 - Operator, 15, 144
 - Unitarily Invariant, 212
- Numeraire, 38, 77
- Optimization
 - Mean-variance, 105
- OTC, *see* Over-the-Counter
- Over-the-Counter, 23, 257
- Parent Order, 257
- Partial Correlation, 110
- Participation Rate, 259
- Passive Investing, 31
- Payment for Order Flow, 26, 29
- PCA
 - K-Means, 235
 - pd, 16
- Performance Attribution
 - Maximal
 - Portfolio-Based, 310
 - Performance Attribution Maximal, 307
 - As Conditional Expectation, 309
 - As Model Rotation, 310
 - Cross-Sectional, 308
 - Nested, 313
 - Selection-Sizing, 314
 - Time Series, 302
 - Permanent Market Impact, 256, 257
- PFOF, *see also* Payment for Order Flow
- PnL, 34
 - Position, 302
 - Trading, 302
- Portfolio
 - basis, 177
 - Construction, 73
 - Eigenportfolios, 177
 - Factor-Mimicking, 111, 194, 273, 303
 - Minimum-Variance, 323
 - Production, 176
 - Portfolio Management, 86
- POV, *see also* Participation Rate 259
- Precision Matrix, 86, 108
- Price
 - Ask, 24
 - Bid, 24
- Price Discovery, 28, 303
- Principal Component, 213, 214

- Principal Component Analysis, 213, 244, 325
Principal Trading Firms, 28
Product
 Scalar, 16
 Hadamard, 16, 321
Projection, 91
Projection Matrix, 91
Projections, 82
Publicly Available Information, 30

R Squared, see Coefficient of Determination114
R squared, seealso Coefficient of Determination173
Rademacher Complexity, 164
Random Recursive Equations, 47, 61
Random Variables
 Heavy Tails, 49
 Heavy-Tailed, 45
 iid, 45, 47, 48, 52, 55, 56, 59, 61
Regression, 236
 Ordinary Least Squares, 90
 Weighted Least Squares, 92, 193
Return, 38
 Dividend-Adjusted, 38
 Excess, 100
Returns, 38, 116
 Compounding, 40
 Excess, 39
 Factor, 72
 Logarithmic, 40
 Risk-Free, 39, 72
 Stylized Facts, 42
Riccati Equation
 Discrete Time Algebraic, 64
 Recursion Formula, 64
Risk, 30, 33
 Marginal Contribution, 85
Risk Model
 Integrated, 232
Risk-Free Rate, 39
rv, 16

Scalar Product, 216
Scree plot, 227
Secured Overnight Financing Rate, 39
Secured Overnight Lending Rate, 25
Sell Side, 23
Settlement, 25
Shadow Price, 108, 138
Sharpe Ratio, 52, 106–108, 110, 116, 117
 Confidence Interval, 120
 Dimensions, 119
 Efficiency, 143
 Sensitivity, 86
Short-Term Factor Updating, 198
Short-Term Idio Updating, 198
Shrinkage, 224
Signal, 34
Singular Value Decomposition, 80, 96, 99, 144, 213, 244
Singular Values, 213
Skill
 Selection, 315
Skill vs. Luck, 74
Slippage, *see also* Temporary Market Impact
Smart Beta, 87
SOFR, *see* Secured Overnight Financing Rate, *see also* Secured Overnight Financing Rate
Spiked Covariance Model, 219
Spread, 24
Spread Cost, 256, 257
SRE, *see* Sharpe Ratio Efficiency143

- STFU, *see also* Short-Term Factor Updating 195
- Strategy
 - 130/30, 134
 - Capacity, 118
 - Long-Only, 134
- Studentization, *see* Z-scoring
- Subsampling, 53
- Subspace
 - Column, 236
 - Similarity Between Two Subspaces, 232
 - Spanned by eigenvectors, 223
- SVD, *see* Singular Value Decomposition
- Tail
 - GARCH(1, 1) processes, 49
 - Gaussian, 45
- Temporary Market Impact, 257
- Time series, 33
- Tracking Error, 135
- tracking volatility, 133
- Trading
 - Informational Effect, 256
 - Mimetic Effect, 256
 - Strategy, 256
- Trading Cost
 - Single-Period, 135
- Turnover
 - Linear, 230
 - Quadratic, 230
- UAM, *see also* Gross Market Value 35
- Unintended Bets, 301
- Utility Theory, 104
- Vanilla Options, 21
- Volatility, 116
 - Ex Ante*, 84
- Realized, 52
- Short-Term Factor Updating, 195
- Wirehouses, *see* Broker-Dealers
- Woodbury-Sherman Morrison Lemma, 329
- Woodbury-Sherman-Morrison Lemma, 281