A Signal Processing Perspective on Financial Engineering

Yiyong Feng
Dept. of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
yiyong@connect.ust.hk

Daniel P. Palomar
Dept. of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
palomar@ust.hk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Y. Feng and D. P. Palomar. *A Signal Processing Perspective on Financial Engineering*. Foundations and Trends® in Signal Processing, vol. 9, no. 1-2, pp. 1–231, 2015.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-119-1

Foundations and Trends® in Signal Processing
Volume 9, Issue 1-2, 2015

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2015, Volume 9, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in Signal Processing
Vol. 9, No. 1-2 (2015) 1–231

DOI: 10.1561/2000000072

now
the essence of knowledge

A Signal Processing Perspective on Financial Engineering

Yiyong Feng
Dept. of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
yiyong@connect.ust.hk

Daniel P. Palomar
Dept. of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
palomar@ust.hk

Contents

Abstract

Financial engineering and electrical engineering are seemingly different areas that share strong underlying connections. Both areas rely on statistical analysis and modeling of systems; either modeling the financial markets or modeling, say, wireless communication channels. Having a model of reality allows us to make predictions and to optimize the strategies. It is as important to optimize our investment strategies in a financial market as it is to optimize the signal transmitted by an antenna in a wireless link.

This monograph provides a survey of financial engineering from a signal processing perspective, that is, it reviews financial modeling, the design of quantitative investment strategies, and order execution with comparison to seemingly different problems in signal processing and communication systems, such as signal modeling, filter/beamforming design, network scheduling, and power allocation.

Y. Feng and D. P. Palomar. *A Signal Processing Perspective on Financial Engineering*. Foundations and Trends® in Signal Processing, vol. 9, no. 1-2, pp. 1–231, 2015.
DOI: 10.1561/2000000072.

1

Introduction

Despite the different natures of financial engineering and electrical engineering, both areas are intimately connected on a mathematical level. The foundations of financial engineering lie on the statistical analysis of numerical time series and the modeling of the behavior of the financial markets in order to perform predictions and systematically optimize investment strategies. Similarly, the foundations of electrical engineering, for instance, wireless communication systems, lie on statistical signal processing and the modeling of communication channels in order to perform predictions and systematically optimize transmission strategies. Both foundations are the same in disguise.

This observation immediately prompts the question of whether both areas can benefit from each other. It is often the case in science that the same or very similar methodologies are developed and applied independently in different areas. The purpose of this monograph is to explore such connections and to capitalize on the existing mathematical tools developed in wireless communications and signal processing to solve real-life problems arising in the financial markets in an unprecedented way.

Thus, this monograph is about investment in financial assets treated as a signal processing and optimization problem. An investment is the current commitment of resources in the expectation of reaping future benefits. In financial markets, such resources usually take the form of money and thus the investment is the present commitment of money in order to reap (hopefully more) money later [27]. The carriers of money in financial markets are usually referred to as financial assets. There are various classes of financial assets, namely, equity securities (e.g., common stocks), exchange-traded funds (ETFs), market indexes, commodities, exchanges rates, fixed-income securities, derivatives (e.g., options and futures), etc. A detailed description of each kind of asset is well documented, e.g., [27, 103]. For different kinds of assets, the key quantities of interest are not the same; for example, for equity securities the quantities of interest are the compounded returns or log-returns; for fixed-income securities they are the changes in yield to maturity; and for options they are changes in the rolling at-the-money forward implied volatility [143].

Roughly speaking, there are three families of investment philosophies: fundamental analysis, technical analysis, and quantitative analysis. Fundamental analysis uses financial and economical measures, such as earnings, dividend yields, expectations of future interest rates, and management, to determine the value of each share of the company's stocks and then recommends purchasing the stocks if the estimated value exceeds the current stock price [88, 89]. Warren Buffett of Berkshire Hathaway is probably the most famous practitioner of fundamental analysis [91]. Technical analysis, also known as "charting," is essentially the search for patterns in one dimensional charts of the prices of a stock. In a way, it pretends to be a scientific analysis of patterns (similar to machine learning) but generally implemented in an unscientific and anecdotal way with a low predictive power, as detailed in [132]. Quantitative analysis applies quantitative (namely scientific or mathematical) tools to discover the predictive patterns from financial data [128]. To put this in perspective with the previous approach, technical analysis is to quantitative analysis what astrology is to astronomy. The pioneer of the quantitative investment approach is Edward O. Thorp, who used

his knowledge of probability and statistics in the stock markets and has made a significant fortune since the late 1960s [193]. Quantitative analysis has become more and more widely used since advanced computer science technology has enabled practitioners to apply complex quantitative techniques to reap many more rewards more efficiently and more frequently in practice [4]. In fact, one could even go further to say that algorithmic trading has been one of the main driving forces in the technological advancement of computers. Some institutional hedge fund firms that rely on quantitative analysis include Renaissance Technologies, AQR Capital, Winton Capital Management, and D. E. Shaw & Co., to name a few.

In this monograph, we will focus on the quantitative analysis of equity securities since they are the simplest and easiest accessible assets. As we will discover, many quantitative techniques employed in signal processing methods may be applicable in quantitative investment. Nevertheless, the discussion in this monograph can be easily extended to some other tradeable assets such as commodities, ETFs, and futures.

Thus, to explore the multiple connections between quantitative investment in financial engineering and areas in signal processing and communications, we will show how to capitalize on existing mathematical tools and methodologies that have been developed and are widely applied in the context of signal processing applications to solve problems in the field of portfolio optimization and investment management in quantitative finance. In particular, we will explore financial engineering in several respects: i) we will provide the fundamentals of market data modeling and asset return predictability, as well as outline state-of-the-art methodologies for the estimation and forecasting of portfolio design parameters in realistic, non-frictionless financial markets; ii) we will present the problem of optimal portfolio construction, elaborate on advanced optimization issues, and make the connections between portfolio optimization and filter/beamforming design in signal processing; iii) we will reveal the theoretical mechanisms underlying the design and evaluation of statistical arbitrage trading strategies from a signal processing perspective based on multivariate data analysis and time series modeling; and iv) we will discuss the optimal order execution

and compare it with network scheduling in sensor networks and power allocation in communication systems.

We hope this monograph can provide more straightforward and systematic access to financial engineering for researchers in signal processing and communication societies[1] so that they can understand problems in financial engineering more easily and may even apply signal processing techniques to handle financial problems.

In the following content of this introduction, we first introduce financial engineering from a signal processing perspective and then make connections between problems arising in financial engineering and those arising in different areas of signal processing and communication systems. At the end, the outline of the monograph is detailed.

1.1 A Signal Processing Perspective on Financial Engineering

Figure 1.1 summarizes the procedure of quantitative investment. Roughly speaking and oversimplifying, there are three main steps (shown in Figure 1.1):

- financial modeling: modeling a very noisy financial time series to decompose it into trend and noise components;
- portfolio design: designing quantitative investment strategies based on the estimated financial models to optimize some preferred criterion; and
- order execution: properly executing the orders to establish or unwind positions of the designed portfolio in an optimal way.

In the following, we will further elaborate the above three steps from a signal processing perspective.

[1]There have been some initiatives in Signal Processing journals on the financial engineering topic, namely, the 2011 IEEE Signal Processing Magazine - Special Issue on Signal Processing for Financial Applications, the 2012 IEEE Journal of Selected Topics in Sginal Processing - Special Issue on Signal Processing Methods in Finance and Electronic Trading, and the 2016 IEEE Journal of Selected Topics in Signal Processing - Special Issue on Financial Signal Processing and Machine Learning for Electronic Trading.

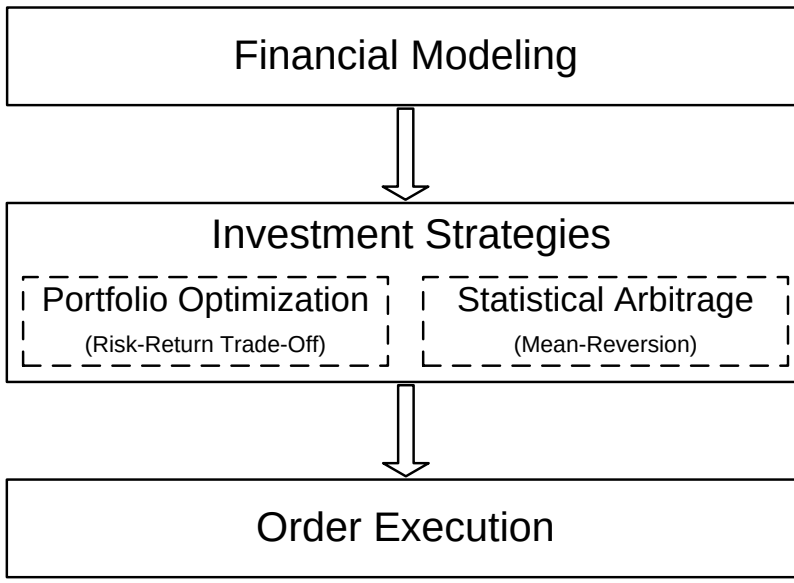

Figure 1.1: Block diagram of quantitative investment in financial engineering.

1.1.1 Financial Modeling

For equity securities, the log-prices (i.e., the logarithm of the prices) and the compounded returns or log-returns (i.e., the differences of the log-prices) are the quantities of interest. From a signal processing perspective, a log-price sequence can be decomposed into two parts: trend and noise components, which are also referred to as market and idiosyncratic components, respectively. The purpose of financial modeling or signal modeling is to decompose the trend components from the noisy financial series. Then based on the constructed financial models, one can properly design some quantitative investment strategies for future benefits [196, 129, 143].

For instance, a simple and popular financial model of the log-price series is the following random walk with drift:

$$y_t = \mu + y_{t-1} + w_t, \tag{1.1}$$

where y_t is the log-price at discrete-time t, $\{w_t\}$ is a zero-mean white noise series, and the constant term μ represents the time trend of the

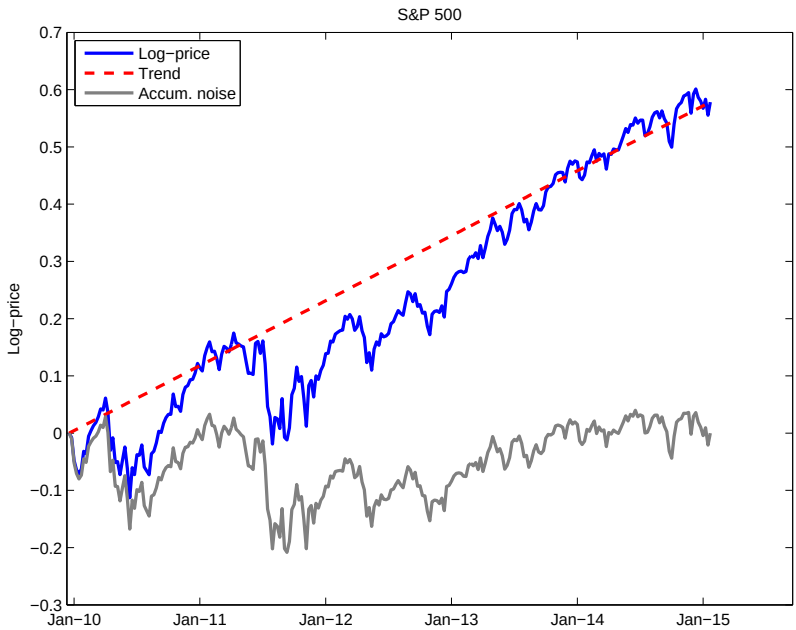

Figure 1.2: The decomposition of the log-price sequence of the S&P 500 Index into time trend component, and the component without time trend (i.e., the accumulative noise).

log-price y_t since $\mathsf{E}\left[y_t - y_{t-1}\right] = \mu$, which is usually referred to as drift.

Based on model (1.1), we can see the trend signal and noise components in the log-prices more clearly by rewriting y_t as follows:

$$y_t = \mu t + y_0 + \sum_{i=1}^{t} w_i, \tag{1.2}$$

where the term μt denotes the trend (e.g., uptrend if $\mu > 0$, downtrend if $\mu < 0$, or no trend if $\mu = 0$), and the term $\sum_{i=1}^{t} w_i$ denotes the accumulative noise as time evolves.

Figure 1.2 shows the weekly log-prices of the S&P 500 index from 04-Jan-2010 to 04-Feb-2015 (the log-prices are shifted down so that the initial log-price is zero, i.e., $y_0 = 0$), where the estimated drift is $\mu = 0.0022$. Obviously, we observe two patterns: first, there exists a significant uptrend since 2010 in the US market (see the dashed red

line μt); and second, the accumulative noise in the log-prices is not steady and looks like a random walk (see the solid gray line for the accumulative noise $\sum_{i=1}^{t} w_i = y_t - \mu t$).

1.1.2 Quantitative Investment

Once the specific financial model is calibrated from the financial time series, the next question is how to utilize such a calibrated financial model to invest. As mentioned before, one widely employed approach is to apply quantitative techniques to design the investment strategies, i.e., the quantitative investment [65, 128, 64, 143].

Figure 1.2 shows that there are two main components in a financial series: trend and noise. Correspondingly, there are two main types of quantitative investment strategies based on the two components: a trend-based approach, termed risk-return trade-off investment; and a noise-based approach, termed mean-reversion investment.

The trend-based risk-return trade-off investment tends to maximize the expected portfolio return while keeping the risk low; however, this is easier said than done because of the sensitivity to the imperfect estimation of the drift component and the covariance matrix of the noise component of multiple assets. In practice, one needs to consider the parameter estimation errors in the problem formulation to design the portfolio in a robust way. Traditionally, the variance of the portfolio return is taken as a measure of risk, and the method is thus referred to as "mean-variance portfolio optimization" in the financial literature [135, 137, 138]. From the signal processing perspective, interestingly, the design of a mean-variance portfolio is mathematically identical to the design of a filter in signal processing or the design of beamforming in wireless multi-antenna communication systems [123, 149, 213].

The noise-based mean-reversion investment aims at seeking profitability based on the noise component. For clarity of presentation, let us use a simple example of only two stocks to illustrate the rough idea. Suppose the log-price sequences of the two stocks are cointegrated (i.e., they share the same stochastic drift), at some point in time if one stock moves up while the other moves down, then people can short-sell the first overperforming stock and long/buy the second underperforming

stock[2], betting that the deviation between the two stocks will eventually diminish. This idea can be generalized from only two stocks to a larger number of stocks to create more profitable opportunities. This type of quantitative investment is often referred to as "pairs trading", or more generally, "statistical arbitrage" in the literature [160, 203].

1.1.3 Order Execution

Ideally, after one has made a prediction and designed a portfolio, the execution should be a seamless part of the process. However, in practice, the process of executing the orders affects the original predictions in the wrong way, i.e., the achieved prices of the executed orders will be worse than what they should have been. This detrimental effect is called market impact. Since it has been shown that smaller orders have a much smaller market impact, a natural idea to execute a large order is to partition it into many small pieces and then execute them sequentially [8, 18, 78, 146].

Interestingly, the order execution problem is close to many other scheduling and optimization problems in signal processing and communication systems. From a dynamic control point of view, the order execution problem is quite similar to sensor scheduling in dynamic wireless sensor networks [180, 181, 208]. From an optimization point of view, distributing a large order into many smaller sized orders over a certain time window [8, 79] corresponds to allocating total power over different communication channels in broadcasting networks [198] or wireless sensor networks [214].

1.2 Connections between Financial Engineering and Areas in Signal Processing and Communication Systems

We have already briefly introduced the main components of financial engineering from a signal processing perspective. In the following we make several specific connections between financial engineering and areas in signal processing and communication systems.

[2]In financial engineering, to "long" means simply to buy financial instruments, to "short-sell" (or simply, to "short") means to sell financial instruments that are not currently owned.

Modeling. One of the most popular models used in financial engineering is the autoregressive moving average (ARMA) model. It models the current observation (e.g., today's return) as the weighted summation of a linear combination of previous observations (e.g., several previous days' returns) and a moving average of the current and several previous noise components [196]. Actually, this model is also widely used in signal processing and it is referred to as a rational model because its z-transform is a rational function, or as a pole-zero model because the roots of the numerator polynomial of the z-transform are known as zeros and the roots of the denominator polynomial of the z-transform are known as poles [133].

Robust Covariance Matrix Estimation. After a specific model has been selected, the next step is to estimate or calibrate its parameters from the empirical data. In general, a critical parameter to be estimated is the covariance matrix of the returns of multiple stocks. Usually the empirical data contains noise and some robust estimation methods are needed in practice. One popular idea in financial engineering is to shrink the sample covariance matrix to the identity matrix as the robust covariance matrix estimator [120]. Interestingly, this is mathematically the same as the diagonal loading matrix (i.e., the addition of a scaled identity matrix to the sample interference-plus-noise covariance matrix) derived more than thirty years ago for robust adaptive beamforming in signal processing and communication systems [1, 38, 45]. For large-dimensional data, the asymptotic performance of the covariance matrix estimators is important. The mathematical tool for the asymptotic analysis is referred to as general asymptotics or large-dimensional general asymptotics in financial engineering [121, 122], or as random matrix theory (RMT) in information theory and communications [199].

Portfolio Optimization vs Filter/Beamforming Design. One popular portfolio optimization problem is the minimum variance problem:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \end{array} \tag{1.3}$$

where $\mathbf{w} \in \mathbb{R}^N$ is the portfolio vector variable representing the normalized dollars invested in N stocks, $\mathbf{w}^T\mathbf{1} = 1$ is the capital budget constraint, and $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$ is the (estimated in advance) positive definite covariance matrix of the stock returns.

The above problem (1.3) is really mathematically identical to the filter/beamforming design problem in signal processing [149]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^H\mathbf{R}\mathbf{w} \\ \text{subject to} & \mathbf{w}^H\mathbf{a} = 1, \end{array} \tag{1.4}$$

where $\mathbf{w} \in \mathbb{C}^N$ is the complex beamforming vector variable denoting the weights of N array observations and $\mathbf{a} \in \mathbb{C}^N$ and $\mathbf{R} \in \mathbb{C}^{N\times N}$ (estimated in advance) are the signal steering vector (also known as the transmission channel) and the positive definite interference-plus-noise covariance matrix, respectively. The similarity between problems (1.3) and (1.4) shows some potential connections between portfolio optimization and filter/beamforming design, and we will explore more related formulations in detail later in the monograph.

Index Tracking vs Sparse Signal Recovery. Index tracing is a widely used quantitative investment that aims at mimicking the market index but with much fewer stocks. That is, suppose that a benchmark index is composed of N stocks and let $\mathbf{r}^b = [r_1^b, \ldots, r_T^b]^T \in \mathbb{R}^T$ and $\mathbf{X} = [\mathbf{r}_1, \ldots, \mathbf{r}_T]^T \in \mathbb{R}^{T\times N}$ denote the returns of the benchmark index and the N stocks in the past T days, respectively, index tracking intends to find a sparse portfolio $\mathbf{w}$ to minimize the tracking error between the tracking portfolio and benchmark index [106]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \frac{1}{T}\|\mathbf{X}\mathbf{w} - \mathbf{r}^b\|_2^2 + \lambda\|\mathbf{w}\|_0 \\ \text{subject to} & \mathbf{1}^T\mathbf{w} = 1, \quad \mathbf{w} \geq \mathbf{0}, \end{array} \tag{1.5}$$

where $\lambda \geq 0$ is a predefined trade-off parameter.

Mathematically speaking, the above problem (1.5) is identical to the sparse signal recovery problem [37] and compressive sensing [51] in signal processing:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \frac{1}{T}\|\mathbf{\Phi}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_0 \tag{1.6}$$

Table 1.1: Connections between financial engineering and signal processing.

	Financial Engineering	Signal Processing
Modeling	ARMA model [196]	rational or pole-zero model [133]
Covariance Matrix Estimation	shrinkage sample covariance matrix estimator [120]	diagonal loading in beamforming [1, 38, 45]
Asymptotic Analysis	(large-dimensional) general asymptotics [121, 122]	random matrix theory [199]
Optimization	portfolio optimization [135, 137, 179, 213]	filter/beamforming design [149, 213]
Sparsity	index tracking [106]	sparse signal recovery [37, 51]

where $\lambda \geq 0$ is a predefined trade-off parameter, $\mathbf{\Phi} \in \mathbb{R}^{T \times N}$ is a dictionary matrix with $T \ll N$, $\mathbf{y} \in \mathbb{R}^T$ is a measurement vector, and $\mathbf{w} \in \mathbb{R}^N$ is a sparse signal to be recovered. Again, the similarity between the two problems (1.5) and (1.6) shows that the quantitative techniques dealing with sparsity may be useful for both index tracking and sparse signal recovery.

Table 1.1 summarizes the above comparisons in a more compact way and it is interesting to see so many similarities and connections between financial engineering and signal processing.

1.3 Outline

The abbreviations and notations used throughout the monograph are provided on pages 211 and 213, respectively.

Figure 1.3 shows the outline of the monograph and provides the recommended reading order for the reader's convenience. The detailed organization is as follows.

Part I mainly focuses on financial modeling (Chapters 2 and 3) and order execution (Chapter 4).

Chapter 2 starts with some basic financial concepts and then introduces several models, such as the i.i.d. model, factor model, ARMA model, autoregressive conditional heteroskedasticity (ARCH) model, generalized ARCH (GARCH) model, and vector error correction model (VECM), which will be used in the later chapters. Thus, this chapter provides a foundation for the following chapters in the monograph.

Chapter 3 deals with the model parameter estimation issues. In particular, it focuses on the estimation of the mean vector and the covariance matrix of the returns of multiple stocks. Usually, these two parameters are not easy to estimate in practice, especially under two scenarios: when the number of samples is small, and when there exists outliers. This chapter reviews the start-of-the-art robust estimation of the mean vector and the covariance matrix from both financial engineering and signal processing.

Chapter 4 formulates the order execution as optimization problems and presents the efficient solving approaches.

Once financial modeling and order execution have been introduced in Part I, we move to the design of quantitative investment strategies. As shown in Figure 1.1 there are two main types of investment strategies, namely risk-return trade-off investment strategies and mean-reversion investment strategies, which are documented in Parts II and III, respectively.

Part II entitled "Portfolio Optimization" focuses on the risk-return trade-off investment. It contains Chapters 5-9 and is organized as follows.

Chapter 5 reviews the most basic Markowitz mean-variance portfolio framework, that is, the objective is to optimize a trade-off between the mean and the variance of the portfolio return. However, this framework is not practical due to two reasons: first, the optimized strategy is extremely sensitive to the estimated mean vector and covariance matrix of the stock returns; and second, the variance is not an appropriate risk measurement in financial engineering. To overcome the second drawback, some more practical single side risk measurements,

e.g., Value-at-Risk (VaR) and Conditional VaR (CVaR), are introduced as the alternatives to the variance.

Chapter 6 presents the robust portfolio optimization to deal with parameter estimation errors. The idea is to employ different uncertainty sets to characterize different estimation errors and then derive the corresponding worst-case robust formulations.

Chapter 7, different from previous Chapters 5 and 6 that consider each portfolio individually, designs multiple portfolios corresponding to different clients jointly via a game theoretic approach by modeling a financial market as a game and each portfolio as a player in the game. This approach is important in practice because multiple investment decisions may affect each other.

Chapter 8 considers a passive quantitative investment method named index tracking. It aims at designing a portfolio that mimics a preferred benchmark index as closely as possible but with much fewer instruments.

Chapter 9 considers a newly developed approach to the portfolio design aiming at diversifying the risk, instead of diversifying the capital as usually done, among the available assets, which is called a "risk parity portfolio" in the literature.

Part III, containing Chapter 10, explores the mean-reversion investment that utilizes the noise component in the log-price sequences of multiple assets.

Chapter 10 introduces the idea of constructing a pair of two stocks via cointegration and optimizes the threshold for trading to achieve a preferred criterion. Then it extends further from pairs trading based on only two stocks to statistical arbitrage for multiple stocks.

After covering the main content of the three parts, Chapter 11 concludes the monograph.

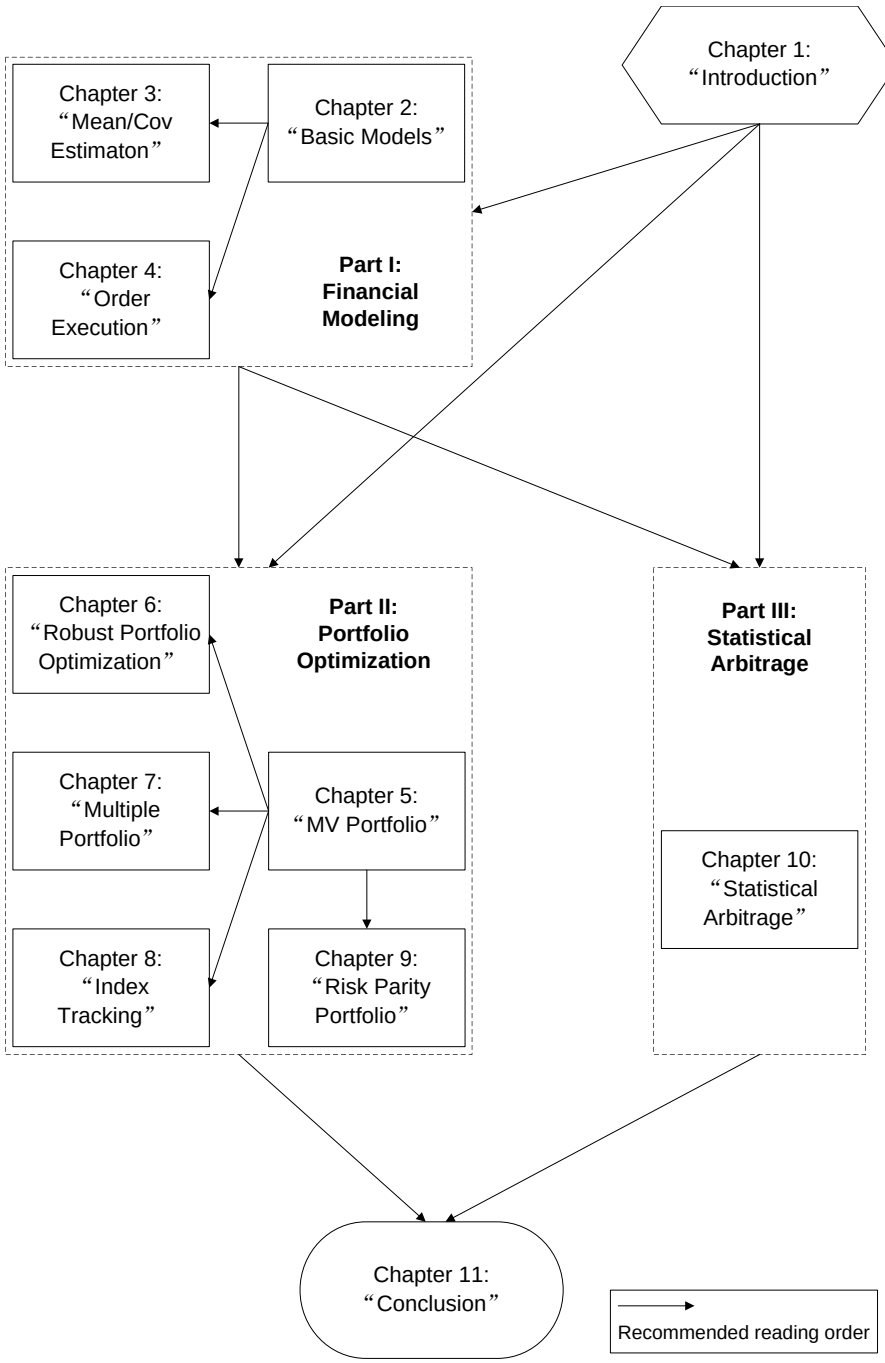

Figure 1.3: Outline of the monograph.

Part I

Financial Modeling & Order Execution

2

Modeling of Financial Time Series

Modeling of financial time series provides the quantitative tools to extract useful (or predictable) information for future investments. There are two main philosophies of modeling like then are in signal processing and control theory [98]: continuous-time and discrete-time systems. Continuous-time modeling, using the Black-Scholes model, for example, involves stochastic calculus and concepts like the Brownian motion that are at the core of many fundamental results. For computational purposes, however, discrete-time modeling is more convenient. In addition, practical investment strategies are usually naturally discretized, i.e., daily or monthly investments.

Therefore, this chapter focuses on discrete-time modeling of financial time series, i.e., the interested time series quantities (mainly the log-returns) of some interested assets (say N assets) given the past information (i.e., the past log-returns of the N assets).

The detailed organization is as follows. Section 2.1 starts with some basic financial concepts, i.e., prices and returns. Then Section 2.2 introduces the general structure of modeling and Sections 2.3-2.7 explain several specific models, such as the i.i.d. model, factor model, vector autoregressive moving average (VARMA) model, vector error correction

model (VECM), autoregressive conditional heteroskedasticity (ARCH) model, generalized ARCH (GARCH) model, and multivariate ARCH and GARCH models, which will be used in the later chapters. At the end, Section 2.8 summarizes all the models briefly.

This chapter focuses on the models themselves but leaves the fitting of the models with real data or parameter estimation to Chapter 3. All the models are introduced in their vector/multivariate cases.

2.1 Asset Returns

For simplicity, let us focus on a single asset. Let p_t be the price of an asset at (discrete) time index t.

2.1.1 Returns Based on Prices

Suppose the asset pays no dividends[1], the simple return (a.k.a. linear return or net return) over one interval from time $t-1$ to t is

$$R_t \triangleq \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1. \tag{2.1}$$

The numerator $p_t - p_{t-1}$ is the profit (or the loss in case of a negative profit) during the holding period and the denominator p_{t-1} is the initial investment at time $t-1$. Thus the simple return can be regarded as the profit rate.

Then the quantity

$$R_t + 1 = \frac{p_t}{p_{t-1}} \tag{2.2}$$

denotes the ratio between the end capital and the initial investment, thus it is referred to as total return or gross return.

Based on the above definitions for only one investment period, the gross return on the most recent k periods is the product of the past k single period gross returns

$$\begin{aligned} 1 + R_t(k) = \frac{p_t}{p_{t-k}} &= \frac{p_t}{p_{t-1}} \times \frac{p_{t-1}}{p_{t-2}} \times \cdots \times \frac{p_{t-k+1}}{p_{t-k}} \\ &= (1 + R_t) \times \cdots \times (1 + R_{t-k+1}), \end{aligned} \tag{2.3}$$

[1] If there exists dividend d_t at time t, then the simple return in (2.1) can be adjusted as $R_t = \frac{p_t - p_{t-1} + d_t}{p_{t-1}}$.

and the corresponding net return is

$$R_t(k) = \frac{p_t}{p_{t-k}} - 1. \tag{2.4}$$

2.1.2 Returns Based on Log-prices

The log-return (a.k.a. continuously compounded return) at time t is defined as follows:

$$r_t \triangleq \log(1 + R_t) = \log \frac{p_t}{p_{t-1}} = y_t - y_{t-1}, \tag{2.5}$$

where $y_t \triangleq \log p_t$ is the log-price and log denotes the natural logarithm.

Since the function $f(x) = \log(1 + x)$ has the first order Taylor approximation $f(x) = \log(1 + x) \approx x$ at point 0, we can see $r_t = \log(1 + R_t)$ is approximately equal to the net return R_t in (2.1), i.e., $r_t \approx R_t$, especially when R_t is small around zero (which is the case for the usual intervals).

The log-return on the most recent k periods is

$$\begin{aligned} r_t(k) &\triangleq \log(1 + R_t(k)) = \log[(1 + R_t) \times \cdots \times (1 + R_{t-k+1})] \\ &= \log(1 + R_t) + \log(1 + R_{t-1}) + \cdots + \log(1 + R_{t-k+1}) \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1}, \end{aligned} \tag{2.6}$$

which has a nice additive property over periods (recall that the linear multi-period net return $R_t(k)$ in (2.4) does not have such a property).

2.1.3 Portfolio Returns

For a portfolio composing of N assets, let $\mathbf{w} \in \mathbb{R}^N$ be a vector with w_i denoting normalized capital invested into the i-th asset. Then the net return of the portfolio over a single period t is $R_t^p = \sum_{i=1}^N w_i R_{it}$ where R_{it} is the net return of the i-th asset.

The log-return of a portfolio, however, does not have the above additivity property. If the simple returns R_{it} are all small in magnitude, they can be approximated by the log-returns r_{it} and the portfolio net return can be approximated as $R_t^p = \sum_{i=1}^N w_i R_{it} \approx \sum_{i=1}^N w_i r_{it}$. However, when some R_{it} are significantly different from zero, using $\sum_{i=1}^N w_i r_{it}$ to approximate $\sum_{i=1}^N w_i R_{it}$ may introduce some serious errors [144].

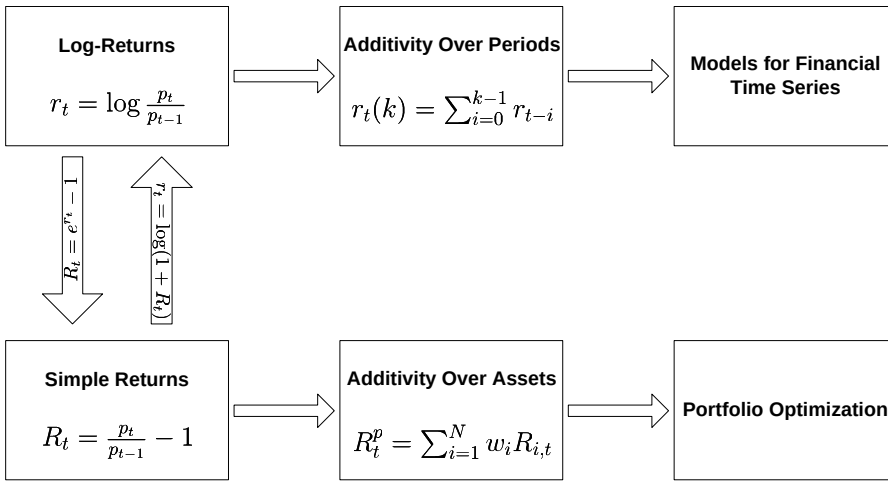

Figure 2.1: Simple returns versus log-returns.

2.1.4 Comparisons: Simple Returns versus Log-returns

Figure 2.1 provides a summary of the comparisons between simple returns and log-returns.

First, the simple returns have the advantage of additivity over assets. Because of that, it is the simple returns that will be used in portfolio optimization later in Part II.

Second, the log-returns have the advantage of additivity over assets periods. This makes the distribution of the log-returns in the future easier to compute and predict.

Third, the statistical properties of the log-returns are relatively more tractable. For example, from (2.1) we can see that simple returns are highly asymmetric because they are bounded below by -1 and unbounded above. Instead, the log-returns are relatively more symmetric and this makes the corresponding distributions easier to model.

It is the additivity over periods and statistical simplicity that are needed for modeling purposes and thus we focus on the log-returns in this chapter. However, as shown in Figure 2.1, either simple returns or log-returns should be used depending on the investor's specific goal.

2.2 General Structure of a Model

Most of the existing financial time series models aim at modeling the log-returns of N assets jointly denoted by $\mathbf{r}_t \in \mathbb{R}^N$. In particular, they model the log-returns at time t based on the previous historical data denoted by $\mathcal{F}_{t-1}$. However, modeling an N-dimensional random variable may be a daunting task not just because of the estimation aspect but also the storage issue. For this reason, most models simplify the task by modeling only the mean and covariance matrix.

Conditional on $\mathcal{F}_{t-1}$, we can decompose $\mathbf{r}_t \in \mathbb{R}^N$ as follows:

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{w}_t, \tag{2.7}$$

where $\boldsymbol{\mu}_t$ is the conditional mean

$$\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t|\mathcal{F}_{t-1}] \tag{2.8}$$

and $\mathbf{w}_t$ is a white noise with zero mean and conditional covariance

$$\boldsymbol{\Sigma}_t = \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T|\mathcal{F}_{t-1}]. \tag{2.9}$$

Here, $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ (or equivalently $\boldsymbol{\Sigma}_t^{1/2}$) are the two main components to be modeled, and they are usually referred to as conditional mean and conditional covariance matrix (or more often conditional volatility for $\boldsymbol{\Sigma}_t^{1/2}$), respectively, in the literature.

In the literature, the underlying distribution $\mathbf{w}_t$ is always assumed to be Gaussian (or sometimes a more general elliptical distribution) for mathematical simplicity even though reality does not fit the thin tails of the Gaussian distribution [143].

In the following, we first provide general models for both $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ and then explore several different types of specific models. Sections 2.3 and 2.4 model both conditional mean and covariance as constants, Sections 2.5 and 2.6 explore various models of the conditional mean but leave the conditional covariance matrix as a constant, and Section 2.7 focuses on modeling the conditional covariance matrix only. All the specific models can be regarded as special cases of the general models, and we summarize them in Section 2.8.

2.2.1 General Model for Conditional Mean $\boldsymbol{\mu}_t$

For most log-return series, the following model is enough to model the conditional mean $\boldsymbol{\mu}_t$:

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i} - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}, \tag{2.10}$$

where $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ denotes a constant vector, $\mathbf{x}_t \in \mathbb{R}^K$ denotes a vector of exogenous variables, $\boldsymbol{\Pi} \in \mathbb{R}^{N\times K}$ is a loading matrix, p and q are nonnegative integers, $\boldsymbol{\Phi}_i, \boldsymbol{\Theta}_j \in \mathbb{R}^{N\times N}$ are matrix parameters, and $\mathbf{r}_{t-i}$ and $\mathbf{w}_{t-j}$ are past log-returns and temporally white noise.

2.2.2 General Model for Conditional Covariance Matrix $\boldsymbol{\Sigma}_t$

For a multivariate case, there exist many different models of the conditional covariance matrix $\boldsymbol{\Sigma}_t$, and, in general, there does not exist a general model formulation that captures all the existing ones as special cases, e.g., see [16, 182, 196, 129]. Nevertheless, for the consistency of presentation, let us introduce the following model [62]:

$$\boldsymbol{\Sigma}_t = \mathbf{A}_0\mathbf{A}_0^T + \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T)\mathbf{A}_i^T + \sum_{j=1}^{s} \mathbf{B}_j \boldsymbol{\Sigma}_{t-j}\mathbf{B}_j^T, \tag{2.11}$$

where m and s are nonnegative integers and $\mathbf{A}_i, \mathbf{B}_j \in \mathbb{R}^{N\times N}$ are parameters. This model ensures a positive definite matrix provided that $\mathbf{A}_0\mathbf{A}_0^T$ is positive definite. The above model is referred to as the Baba-Engle-Kraft-Kroner (BEKK) model in the literature.

In practice, most models simply assume a constant covariance matrix $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w$, i.e., a special case of (2.11) with $m = 0$ and $s = 0$.

2.3 I.I.D. Model

Perhaps the simplest model for $\mathbf{r}_t$ is that it follows an i.i.d. distribution with fixed mean and covariance matrix, i.e.,

$$\mathbf{r}_t = \boldsymbol{\mu} + \mathbf{w}_t, \tag{2.12}$$

where $\mathbf{w}_t \in \mathbb{R}^N$ is a white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$.

Comparing the i.i.d. model (2.12) with the general model (2.7)-(2.11), obviously we can see it is the simplest special case with $\boldsymbol{\mu} = \boldsymbol{\phi}_0$, $\boldsymbol{\Pi} = \mathbf{0}$, $p = 0$, $q = 0$, $\boldsymbol{\Sigma}_w = \mathbf{A}_0\mathbf{A}_0^T$, $m = 0$, and $s = 0$. And the conditional mean and covariance matrix are both constant:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}, \tag{2.13}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w. \tag{2.14}$$

This i.i.d. model assumption may look simple, however, it is one of the most fundamental assumptions for many important works. One example is the Nobel prize-winning Markowitz portfolio theory [135, 136, 137, 138, 179] that will be covered in Chapter 5.

2.4 Factor Model

If we look at (2.12) carefully, we may think that the dimension of the market always equals the number of assets N. However, this may not be true in practice. In general, the market is composed of a large number of assets (i.e., N is large), but it is usually observed that its dimension is relatively small, that is, the market is only driven by a limited number of factors, say K factors with $K \ll N$.

The general factor model is

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \mathbf{h}(\mathbf{f}_t) + \mathbf{w}_t, \tag{2.15}$$

where $\boldsymbol{\phi}_0$ denotes a constant vector; $\mathbf{f}_t \in \mathbb{R}^K$ with $K \ll N$ is a vector of a few factors that are responsible for most of the randomness in the market, the vector function $\mathbf{h} : \mathbb{R}^K \mapsto \mathbb{R}^N$ denotes how the low dimensional factors affect the higher dimensional market; and a residual vector $\mathbf{w}_t$ of (possibly independent) perturbations that has only a marginal effect. In general, the function $\mathbf{h}$ is assumed to be linear.

This approach of modeling enjoys a wide popularity; refer to [42, 66, 67, 68, 69, 70, 118] for some typical references.

In the following, we consider two specific models of (2.15) with either explicit or hidden factors.

2.4.1 Explicit Factors

The explicit factor model is

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{f}_t + \mathbf{w}_t, \tag{2.16}$$

which is a specific case of (2.15) with $\mathbf{h}(\mathbf{f}_t) = \boldsymbol{\Pi}\mathbf{f}_t$, $\mathbf{f}_t \in \mathbb{R}^K$ being explicitly observable market variables, and $\boldsymbol{\Pi} \in \mathbb{R}^{N\times K}$ being the factor loading matrix.

Some popular explicit factors include returns on the market portfolio[2], growth rate of the GDP, interest rate on short term Treasury bills, inflation rate, unemployment, etc. [171].

Obviously, the factor model with explicit factors (2.16) is a special case of the general model (2.7)-(2.11) with exogenous input being the factors $\mathbf{x}_t = \mathbf{f}_t$, $p = 0$, and $q = 0$.

In general, it is assumed that $\mathbf{f}_t$ follows an i.i.d. distribution with constant mean $\boldsymbol{\mu}_f$ and constant covariance matrix $\boldsymbol{\Sigma}_f$, $\mathbf{w}_t$ follows an i.i.d. distribution with zero mean and (possibly diagonal) constant covariance matrix $\boldsymbol{\Sigma}_w$, and $\mathbf{f}_t$ and $\mathbf{w}_t$ are uncorrelated. Then the conditional mean and covariance matrix are both constant and can be computed as follows:

$$\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t|\mathcal{F}_{t-1}] = \mathsf{E}[\mathbf{r}_t] = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\boldsymbol{\mu}_f \tag{2.17}$$

$$\begin{aligned}\boldsymbol{\Sigma}_t &= \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T|\mathcal{F}_{t-1}], \\ &= \boldsymbol{\Pi}\boldsymbol{\Sigma}_f\boldsymbol{\Pi}^T + \boldsymbol{\Sigma}_w.\end{aligned} \tag{2.18}$$

Capital Asset Pricing Model (CAPM)

One of the most popular factor models is the CAPM with the returns on the market portfolio being the only factor [70]. The i-th stock return at time t is

$$r_{i,t} - r_f = \beta_i(r_{M,t} - r_f) + w_{i,t}, \tag{2.19}$$

where r_f is the risk-free rate, $r_{M,t}$ is the return of the market portfolio, and $w_{i,t}$ is a stock-specific white noise with zero mean and constant variance.

[2]The market portfolio is a portfolio consisting of all equities with the normalized portfolio weights being proportional to the market values of the equities.

Taking the expectation on both sides of (2.19) results in the so-called CAPM:

$$\mathsf{E}\left[r_{i,t}\right] - r_f = \beta_i(\mathsf{E}\left[r_{M,t}\right] - r_f). \tag{2.20}$$

Based on (2.20)

- $\mathsf{E}\left[r_{M,t}\right] - r_f$ measures the difference between the expected market return and risk-free rate, which is known as the market premium;
- $\mathsf{E}\left[r_{i,t}\right] - r_f$ measures the difference between the expected stock return and risk-free rate, which is known as the risk premium; and
- β_i in general is given by

$$\beta_i = \frac{\mathsf{Cov}(r_{i,t}, r_{M,t})}{\mathsf{Var}(r_{M,t})} \tag{2.21}$$

  which measures how sensitive the risk premium is to the market premium, that is, the risk premium equals the market premium times β_i.

Note that the conditional mean $\mathsf{E}\left[r_{i,t}|\mathcal{F}_{t-1}\right]$ is the same as the unconditional mean $\mathsf{E}\left[r_{i,t}\right] = r_f + \beta_i(\mathsf{E}\left[r_{M,t}\right] - r_f)$.

Taking the variance on both sides of (2.19) gives us the following relationship:

$$\mathsf{Var}\left[r_{i,t}\right] = \beta_i^2 \mathsf{Var}\left[r_{M,t}\right] + \mathsf{Var}\left[w_{i,t}\right], \tag{2.22}$$

which is decomposed into two parts:

- $\beta_i^2 \mathsf{Var}\left[r_{M,t}\right]$ measures the risk associated with the market and it is referred to as systematic risk, and
- $\mathsf{Var}\left[w_{i,t}\right]$ is specific to each stock and it is called nonsystematic risk.

Also, the conditional variance $\mathsf{Var}\left[r_{i,t}|\mathcal{F}_{t-1}\right]$ equals the unconditional variance $\mathsf{Var}\left[r_{i,t}\right]$.

2.4.2 Hidden Factors

The assumption of a linear model of (2.15) with hidden factors is that the factors are not explicit market variables but are functions of $\mathbf{r}_t$ that summarize as much information as possible.

One method is to define the hidden factors as affine transformations of $\mathbf{r}_t$ as follows:

$$\mathbf{f}_t = \mathbf{d} + \boldsymbol{\Upsilon}^T \mathbf{r}_t, \tag{2.23}$$

where $\mathbf{d} \in \mathbb{R}^K$ and $\boldsymbol{\Upsilon} \in \mathbb{R}^{N \times K}$ are parameters to be estimated.

Then the hidden factor model can be expressed as follows:

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}(\mathbf{d} + \boldsymbol{\Upsilon}^T \mathbf{r}_t) + \mathbf{w}_t, \tag{2.24}$$

which is a specific case of (2.15) with $\mathbf{h}(\mathbf{f}_t) = \boldsymbol{\Pi}\mathbf{f}_t$, $\mathbf{f}_t \in \mathbb{R}^K$ being the hidden variables defined in (2.23); $\boldsymbol{\Pi} \in \mathbb{R}^{N \times K}$ being the factor loading matrix; and $\mathbf{w}_t$ follows an i.i.d. distribution with zero mean and a (possibly diagonal) constant covariance matrix $\boldsymbol{\Sigma}_w$.

The model (2.24) can be further simplified as follows:

$$\mathbf{r}_t = \mathbf{m} + \boldsymbol{\Pi}\boldsymbol{\Upsilon}^T \mathbf{r}_t + \mathbf{w}_t, \tag{2.25}$$

where $\mathbf{m} = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{d}$ is an newly defined parameter.

The parameters $\mathbf{m}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Upsilon}$ can be estimated by the following nonlinear least-square (LS) regression:

$$\underset{\mathbf{m},\, \boldsymbol{\Pi},\, \boldsymbol{\Upsilon}}{\text{minimize}} \quad \mathsf{E}\left\| \mathbf{r}_t - \mathbf{m} - \boldsymbol{\Pi}\boldsymbol{\Upsilon}^T \mathbf{r}_t \right\|_2^2. \tag{2.26}$$

Recall that $\boldsymbol{\Pi}, \boldsymbol{\Upsilon} \in \mathbb{R}^{N \times K}$, then $\boldsymbol{\Pi}\boldsymbol{\Upsilon}^T \in \mathbb{R}^{N \times N}$ with $\text{rank}(\boldsymbol{\Pi}\boldsymbol{\Upsilon}^T) \leq K \ll N$, then intuitively problem (2.26) is projecting $\mathbf{r}_t$ onto a lower K-dimensional subspace with variations being captured as much as possible. Indeed, this technique is usually referred to as principal component analysis (PCA) [109] in the literature, the optimal solution of which can be stated in closed-form as follows [143]:

$$\boldsymbol{\Pi} = \boldsymbol{\Upsilon} = \mathbf{E}_K, \tag{2.27}$$

$$\mathbf{m} = \left(\mathbf{I} - \mathbf{E}_K \mathbf{E}_K^T\right) \mathsf{E}[\mathbf{r}_t], \tag{2.28}$$

where $\mathbf{E}_K \in \mathbb{R}^{N \times K}$ with the k-th column vector being the k-th largest eigenvector of the covariance matrix $\mathsf{Cov}[\mathbf{r}_t]$, $k = 1, \ldots, K.$, and it can be shown that the white noise $\mathbf{w}_t$ is uncorrelated of the hidden factors.

Then combining (2.25), (2.27) and (2.28) together, we can find the conditional mean and covariance matrix as follows:

$$\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t|\mathcal{F}_{t-1}] = \mathsf{E}[\mathbf{r}_t], \tag{2.29}$$

$$\begin{aligned}\boldsymbol{\Sigma}_t &= \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T|\mathcal{F}_{t-1}] \\ &= \mathbf{E}_K \boldsymbol{\Lambda}_K \mathbf{E}_K^T + \boldsymbol{\Sigma}_w,\end{aligned} \tag{2.30}$$

where $\boldsymbol{\Lambda}_K = \text{Diag}([\lambda_1, \ldots, \lambda_K])$ is a K-by-K diagonal matrix with λ_k being the k-th largest eigenvalue of $\mathsf{Cov}[\mathbf{r}_t]$, and we can see both the conditional mean and covariance matrix are constant and independent of time.

2.4.3 Comparisons: Explicit Factors versus Hidden Factors

Based on (2.17)-(2.18) or (2.29)-(2.30), we can see that the factor models, i.e., (2.16) and (2.25), decompose the conditional covariance $\boldsymbol{\Sigma}_t$ into two parts: low dimensional factors and marginal noise. The key is the way to choose or construct the factors, and the comparisons between the explicit and hidden factor models are as follows:

- The explicit factor model tends to explain the log-returns with a smaller number of fundamental or macroeconomic variables and thus it is easier to interpret. However, in general there is no systematic method to choose the right factors.

- The hidden factor model employs PCA to explore the structure of the covariance matrix and locate a low-dimensional subspace that captures most of the variation in the log-returns. It is a more systematical approach and thus it may provide a better explanatory power. One drawback of the hidden factors compared with the explicit factors is that they do not have explicit econometric interpretations.

2.5 VARMA Model

The previous i.i.d. and factor models, while commonly employed, do not incorporate any time-dependency in the model for $\mathbf{r}_t$. In other words, the conditional mean and covariance matrix are constant and

past information is not explicitly used (it can still be used implicitly via the estimation of the parameters).

The VARMA model can incorporate the past information into the model of conditional mean, although still not in the conditional covariance matrix.

Stationarity is an important characteristic for time series analysis which describes the time-invariant behavior of a time series. A multivariate time series $\mathbf{r}_t$ is said to be weakly stationary if its first and second moments are time-invariant. In general, a stationary time series is much easier to model, estimate, and analyze.

2.5.1 VAR(1) Model

Let us start with the vector autoregressive (VAR) model of order 1, denoted as VAR(1), as follows:

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \mathbf{r}_{t-1} + \mathbf{w}_t, \tag{2.31}$$

where $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ is a constant vector, $\boldsymbol{\Phi}_1 \in \mathbb{R}^{N \times N}$ is a matrix parameter, and $\mathbf{w}_t$ denotes a serially uncorrelated noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$. We can see that the term $\boldsymbol{\Phi}_1 \mathbf{r}_{t-1}$ models the serial correlation of the time series $\mathbf{r}_t$.

Also, compared with the general model (2.7)-(2.11), the VAR(1) model (2.31) is a special case with $\boldsymbol{\Pi} = \mathbf{0}$, $p = 1$, $q = m = s = 0$, and $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w$, and it is straightforward to obtain the conditional mean and covariance matrix based on (2.31) as follows:

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \mathbf{r}_{t-1}, \tag{2.32}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w. \tag{2.33}$$

Obviously, the conditional covariance matrix $\boldsymbol{\Sigma}_t$ is constant.

2.5.2 VAR(p) Model

The p-th order autoregressive process, denoted as VAR(p), extends the VAR(1) model by including more previous observations into the model as follows:

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i} + \mathbf{w}_t, \tag{2.34}$$

where p is a nonnegative integer, $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ is a constant vector, $\boldsymbol{\Phi}_i \in \mathbb{R}^{N\times N}$ are matrix parameters, and $\mathbf{w}_t$ denotes a serially uncorrelated white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$.

Clearly we can see that the time series $\mathbf{r}_t$ is serially correlated via the term $\sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i}$ which contains more previous observations than the AR(1) model (2.31). Similar to (2.32) and (2.33), the conditional mean and covariance matrix based on (2.34) are

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i}, \tag{2.35}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w, \tag{2.36}$$

where the conditional covariance matrix is constant.

2.5.3 VMA(q) Model

Even though the VAR model models the serial correlations, it imposes such correlations with all the past observations. We can observe this easily by substituting the VAR(1) model (2.31) recursively and we have that $\mathbf{r}_t$ is serially correlated to all the past observations $\mathbf{r}_0, \ldots, \mathbf{r}_{t-1}$, especially when the eigenvalues of $\boldsymbol{\Psi}_1$ are close to 1.

For some realistic cases, the time series $\mathbf{r}_t$ should only have serial correlation up to a small lag q such that $\mathbf{r}_t$ is serially uncorrelated to $\mathbf{r}_{t-\ell}$ for all $\ell > q$. Unfortunately, the VAR model does not have this property.

A useful alternative to the VAR model is a vector moving average (VMA) model. The VMA model of order q, denoted as VMA(q), is

$$\mathbf{r}_t = \boldsymbol{\mu} + \mathbf{w}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}, \tag{2.37}$$

where q is a nonnegative integer, $\boldsymbol{\mu} \in \mathbb{R}^N$ is a constant vector, $\boldsymbol{\Theta}_j \in \mathbb{R}^{N\times N}$ are matrix parameters, and $\mathbf{w}_t$ denotes a serially uncorrelated white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$.

Based on (2.37), it is easy to check that $\mathbf{r}_t$ is serially uncorrelated to $\mathbf{r}_{t-\ell}$ for all $\ell > q$. Also, the VMA(q) model (2.37) is a special case of the general model (2.7)-(2.11) with $\boldsymbol{\Pi} = \mathbf{0}$ and $p = m = s = 0$, and

we have the conditional mean and covariance matrix as follows:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu} - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}, \tag{2.38}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w, \tag{2.39}$$

where the conditional covariance matrix is constant.

2.5.4 VARMA Model

Sometimes, using simply a VAR model or a VMA model only is not enough to fit the data and it is helpful to combine them together. The combination of VAR(p) and VMA(q), referred to as VARMA(p, q), is given by

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i} + \mathbf{w}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}, \tag{2.40}$$

where p and q are nonnegative integers, $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ is a constant vector, the matrices $\boldsymbol{\Phi}_i, \boldsymbol{\Theta}_j \in \mathbb{R}^{N \times N}$ are parameters, and $\mathbf{w}_t$ is a white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$. Directly, the conditional mean and covariance matrix based on (2.40) are

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i} - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}, \tag{2.41}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w, \tag{2.42}$$

where the conditional covariance matrix is still constant.

Remark 2.1. The VARMA model is a powerful model of conditional mean, however, it also has some drawbacks that need to be dealt with carefully.

The identifiability issue, i.e., two VARMA(p, q) models with different coefficient matrices can be rewritten as the same VMA(∞) model, is one of the most important ones. This issue is important because the likelihood function of the VARMA(p, q) model may not be uniquely defined and thus the parameters cannot be estimated. To overcome this drawback, some model structural specifications are needed. There

are two main approaches namely the Kronecker index, and the scalar component model in the literature [197].

Another issue is that, for a causal and invertible VARMA model, the conditional maximum likelihood estimation may not result in a causal and invertible estimated VARMA model, especially when the number of samples is small [129, 197]. The solving approach is to either add more constraints in the conditional maximum likelihood estimation [169] or switch to the unconditional maximum likelihood estimation [197]. However, both of them require more intensive computation. ■

2.6 VECM

Until now we have focused on modeling directly the log-return series $\mathbf{r}_t$ instead of the log-price series $\mathbf{y}_t$ (recall that $\mathbf{r}_t = \Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$). This is because in general the log-price series $\mathbf{y}_t$ is not weakly stationary (think for example of Apple stock whose log-prices keep increasing) and thus is not easy to model, while its difference series, i.e., the log-return series $\mathbf{r}_t$, is weakly stationary and is easier to model and analyze.

However, it turns out that differencing may destroy part of the relationship among the log-prices which may be invaluable for a proper modeling with forecast power. It is therefore also important to analyze the original (probably non-stationary) time series directly [129].

Interestingly, it turns out that in fact a (probably non-stationary) VAR model may be enough. For example, one can always fit the log-price series $\mathbf{y}_t$ with a VAR model, say, the following VAR(p):

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \cdots + \boldsymbol{\Phi}_{t-p}\mathbf{y}_{t-p} + \mathbf{w}_t, \tag{2.43}$$

where p is a nonnegative integer, $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ is a constant vector, $\boldsymbol{\Phi}_i \in \mathbb{R}^{N\times N}$ are matrix parameters, and $\mathbf{w}_t$ denotes a serially uncorrelated white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$.

Here (2.43) models the log-price series and $\mathbf{y}_t$ is not necessarily stationary. The standard results for a stationary VAR model may not be useful.

In the literature, a time series is called integrated of order p, denoted as $I(p)$, if the time series obtained by differencing the time series p times is weakly stationary, while by differencing the time series $p-1$ times is

not weakly stationary [196, 129]. A multivariate time series is said to be cointegrated if it has at least one linear combination being integrated of a lower order. To illustrate the concepts visually, we consider a slightly modified example from [196] with only two dimensions as follows.

Example 2.1. Suppose the log-price series $\mathbf{y}_t$ follows

$$\mathbf{y}_t = \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{w}_t, \tag{2.44}$$

where $\mathbf{\Phi}_1 = \begin{bmatrix} 0.5 & -1 \\ -0.25 & 0.5 \end{bmatrix}$, and $\mathbf{w}_t$ follows an i.i.d. distribution with zero mean and constant covariance matrix $\mathbf{\Sigma}_w$. The model (2.44) (or $\mathbf{y}_t$) is not stationary because the eigenvalues of $\mathbf{\Phi}_1$ are 0 and 1 (recall for stationarity the modulus of the eigenvalues need to be less than one).

To check the integration order of $\mathbf{y}_t$, rewriting (2.44) as

$$\begin{bmatrix} 1 - 0.5B & B \\ 0.25B & 1 - 0.5B \end{bmatrix} \mathbf{y}_t = \mathbf{w}_t, \tag{2.45}$$

where B is the backshift operator, and premultiplying both sides of (2.45) by $\begin{bmatrix} 1 - 0.5B & -B \\ -0.25B & 1 - 0.5B \end{bmatrix}$ yields

$$\begin{bmatrix} 1 - B & 0 \\ 0 & 1 - B \end{bmatrix} \mathbf{y}_t = \begin{bmatrix} 1 - 0.5B & -B \\ -0.25B & 1 - 0.5B \end{bmatrix} \mathbf{w}_t. \tag{2.46}$$

Since the right hand side of (2.46) is stationary, so is the first order difference of $\mathbf{y}_t$ on the left hand side of (2.46). This implies that $\mathbf{y}_t$ is integrated of order one, i.e., it is $I(1)$.

To check whether $\mathbf{y}_t$ is cointegrated or not, we define $\mathbf{L} \triangleq \begin{bmatrix} 1 & -2 \\ 0.5 & 1 \end{bmatrix}$ and premultiply (2.44) by $\mathbf{L}$, then we have

$$\mathbf{L}\mathbf{y}_t = \mathbf{L}\mathbf{\Phi}_1 \mathbf{L}^{-1} \mathbf{L}\mathbf{y}_{t-1} + \mathbf{L}\mathbf{w}_t, \tag{2.47}$$

which can be rewritten more explicitly as

$$\begin{bmatrix} y_{1t} - 2y_{2t} \\ 0.5y_{1t} + y_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-1} - 2y_{2,t-1} \\ 0.5y_{1,t-1} + y_{2,t-1} \end{bmatrix} + \mathbf{L}\mathbf{w}_t. \tag{2.48}$$

Since $\mathbf{L}\mathbf{w}_t$ is always stationary, so is the linear combination $0.5y_{1t} + y_{2t}$, and thus $\mathbf{y}_t$ is cointegrated. This derived cointegration result in fact is very important and can be utilized to design very profitable quantitative trading strategies (which will be shown later in Part III).

Now we can observe that if we difference the log-price series directly and reach the model (2.46), we cannot obtain the cointegration result that $0.5y_{1t} + y_{2t}$ is stationary any more. Therefore, it is important to study the log-price series $\mathbf{y}_t$ directly as mentioned before. ■

The above Example 2.1 shows a specific example of cointegration. In practice, a systematic way to find the cointegrated components (if they exist) is via a vector error correction model (VECM) [61].

Let us assume the log-price series $\mathbf{y}_t$ is at most $I(1)$, that is, at least its difference series $\mathbf{r}_t$ or the log-return series is always weakly stationary. Using the relation $\mathbf{y}_t = \mathbf{y}_{t-1} + \mathbf{r}_t$, the VAR($p$) model (2.43) can always be rewritten as

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \tilde{\boldsymbol{\Phi}}_1\mathbf{r}_{t-1} + \cdots + \tilde{\boldsymbol{\Phi}}_{p-1}\mathbf{r}_{t-p+1} + \mathbf{w}_t, \tag{2.49}$$

where

$$\boldsymbol{\Pi} = -(\mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p) = -\boldsymbol{\Phi}(1) \tag{2.50}$$

$$\tilde{\boldsymbol{\Phi}}_j = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i, \quad j = 1, \ldots, p-1. \tag{2.51}$$

Interestingly, the above model (2.49) can also be regarded as a special case of the general model (2.10) with the exogenous variables being the previous log-prices, i.e., $\mathbf{x}_t = \mathbf{y}_{t-1}$. And the conditional mean and covariance matrix are

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \tilde{\boldsymbol{\Phi}}_i\mathbf{r}_{t-i}, \tag{2.52}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w, \tag{2.53}$$

where the conditional covariance matrix is constant.

Under the assumption that $\mathbf{y}_t$ is at most $I(1)$, it is straightforward to conclude that the term $\boldsymbol{\Pi}\mathbf{y}_{t-1}$ in the above model (2.49) is stationary, therefore, some linear combinations of $\mathbf{y}_t$ may be stationary. The term

$\mathbf{\Pi}\mathbf{y}_{t-1}$ is usually referred to as an error correction term and thus the model is called a VECM. There are three interesting cases of $\mathbf{\Pi}\mathbf{y}_{t-1}$:

1. $\text{rank}(\mathbf{\Pi}) = 0$. This implies $\mathbf{\Pi} = \mathbf{0}$ and $\mathbf{y}_t$ is not cointegrated since there is no linear combination of $\mathbf{y}_t$ being stationary. Then the VECM (2.49) reduces to a VAR$(p-1)$ for the log-return time series $\mathbf{r}_t$.

2. $\text{rank}(\mathbf{\Pi}) = N$. This implies $\mathbf{\Pi}$ is invertible. Then $\mathbf{y}_t$ must be stationary already since $\mathbf{r}_t$ and $\mathbf{w}_t$ are both stationary and $\mathbf{y}_t$ can be rewritten as a linear combination of $\mathbf{r}_t$ and $\mathbf{w}_t$ by left multiplying both sides of (2.49) by $\mathbf{\Pi}$ inverse. Thus, one can study $\mathbf{y}_t$ directly.

3. $0 < \text{rank}(\mathbf{\Pi}) = r < N$. This is the interesting case and $\mathbf{\Pi}$ can be decomposed as

$$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T, \tag{2.54}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{N\times r}$ with full column rank, i.e., $\text{rank}(\boldsymbol{\alpha}) = \text{rank}(\boldsymbol{\beta}) = r$. Then the VECM (2.49) becomes

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{y}_{t-1} + \tilde{\mathbf{\Phi}}_1\mathbf{r}_{t-1} + \cdots + \tilde{\mathbf{\Phi}}_{p-1}\mathbf{r}_{t-p+1} + \mathbf{w}_t. \tag{2.55}$$

This means that the log-price time series $\mathbf{y}_t$ has r linearly independent cointegrated components, i.e., $\boldsymbol{\beta}^T\mathbf{y}_t$. This interesting property can be used to design mean-reversion statistical arbitrage investment strategies, e.g., pairs trading strategies, as we will cover later in Part III.

2.7 Conditional Volatility Models

The previous models only model the conditional mean while always keeping the conditional volatility as a constant, e.g., see (2.14), (2.18), (2.30), (2.33), (2.36), (2.42), and (2.53). In the real market, usually time-varying rather than constant volatility is observed. For example, a well-known phenomenon is that high volatility is more likely followed by high volatility rather than low volatility and it is hence referred to as "volatility clustering". Let us illustrate the concept with the following Example 2.2.

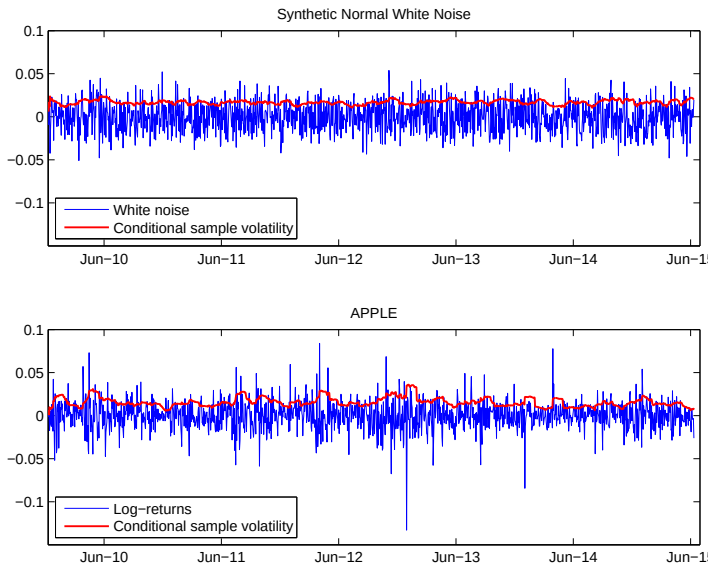

Figure 2.2: White noise versus APPLE log-returns. The conditional sample volatility is the sample standard deviation of the most recent 22 days observations (i.e., white noise observations or log-returns).

Example 2.2. We study the daily log-returns of Apple Inc. from 01-Jan-2010 to 08-Jul-2015. The sample volatility is $\sigma = 1.659 \times 10^{-2}$. We then synthetically simulate a Gaussian white noise series with zero mean and variance σ^2.

The top panel of Figure 2.2 shows a simulated realization of the Gaussian white noise series and the conditional sample volatility, and the bottom panel shows that of the log-returns series of Apple Inc. Here the conditional sample volatility is the sample standard deviation of the most recent 22 days observations (i.e., white noise observations or log-returns). Clearly, we can see that the synthetic Gaussian white noise series has quite stable conditional sample volatility while the log-return series of Apple Inc. has volatile conditional sample volatility and there exist some volatility clusters. ■

In this subsection, we mainly focus on reviewing the models of conditional volatility. Since there are many different multivariate models

of conditional volatility extending from the same univariate models, we will start with the univariate models first and then discuss the multivariate models.

2.7.1 Univariate ARCH Model

Recall that previously the white noise w_t in the general model (2.7) has always been modeled as a zero mean noise with constant variance. Since the conditional mean $\boldsymbol{\mu}_t$ in (2.7) has been well explored in the previous parts of this chapter, without loss of generality, we focus now on models for conditional volatility. The autoregressive conditional heteroskedasticity (ARCH) model is the first one that focuses on modeling the conditional volatility [59]. The ARCH(m) model is

$$w_t = \sigma_t z_t, \tag{2.56}$$

where $\{z_t\}$ is a white noise series with zero mean and unit variance and the conditional variance σ_t^2 is modeled by

$$\sigma_t^2 = \alpha + \sum_{i=1}^{m} \alpha_i w_{t-i}^2. \tag{2.57}$$

Here, m is a nonnegative integer, $\alpha > 0$, and $\alpha_i \geq 0$ for all $i > 0$. The coefficients α_i must satisfy some regularity conditions so that the unconditional variance of w_t is finite. Also, the white noise with zero mean and constant variance in model (2.7) can be regarded as a special case of (2.57) with $\alpha_i = 0$ for all $i > 0$. We can see that the past information is incorporated into the model by using $\sum_{i=1}^{m} \alpha_i w_{t-i}^2$ to model the variance (or equivalently, the square of the volatility).

Even though the ARCH model can model the conditional heteroskedasticity, it has several disadvantages [196]:

- positive and negative noise have the same effects on volatility because volatility modeled by (2.57) depends on the square of the previous noise; however, it is well known that they have different impacts on the financial assets;
- the ARCH model is too restrictive to capture some patterns, e.g., excess kurtosis;

- the ARCH model does not provide any new insight for understanding the source of variations and only provides a mechanical way to describe the behavior of the conditional variance; and
- ARCH models tend to overpredict the volatility because they respond slowly to large isolated noise to the return series.

2.7.2 Univariate GARCH Model

A limitation of the ARCH model is that the high volatility is not persistent enough and it often requires many parameters to describe the volatility process. An extension called Generalized ARCH (GARCH) was proposed to overcome this drawback [28]. The GARCH(m, s) model is

$$w_t = \sigma_t z_t, \tag{2.58}$$

where $\{z_t\}$ is a white noigse series with zero mean and constant unit variance, and the conditional variance σ_t^2 is modeled by

$$\sigma_t^2 = \alpha + \sum_{i=1}^{m} \alpha_i w_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2. \tag{2.59}$$

Here, m and s are nonnegative integers, $\alpha > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for all $i > 0$ and $j > 0$ and $\sum_{i=1}^{m} \alpha_i + \sum_{j=1}^{s} \beta_j \leq 1$.

We can see the GARCH model (2.59) in fact is obtained by adding the term $\sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$ to the previous ARCH model (2.59), therefore, the volatility is more persistent and the volatility clustering phenomenon can be modeled better. For illustrative purposes, a numerical example is provided as follows.

Example 2.3. We consider an ARCH(1) model with $\alpha = 0.01$ and $\alpha_1 = 0.2$, an ARCH(9) model with $\alpha = 0.01$ and $\alpha_1 = 0.2/2^{i-1}$, $i = 1, \ldots, 9$ and a GARCH(1,1) model with $\alpha = 0.01$, $\alpha_1 = 0.2$, and $\beta_1 = 0.7$.

Figure 2.3 shows the realization path and conditional volatilities of each model. The volatility clusters of the ARCH(1) are quite sharp and thus not persistent enough. The higher order ARCH(9) model overcomes the drawback to some degree however, it requires many more parameters (i.e., ten parameters compared to two of the ARCH(1)).

Comparatively, the GARCH(1,1) model captures the volatility clustering relatively more persistently and requires much less (i.e., only three) parameters. ■

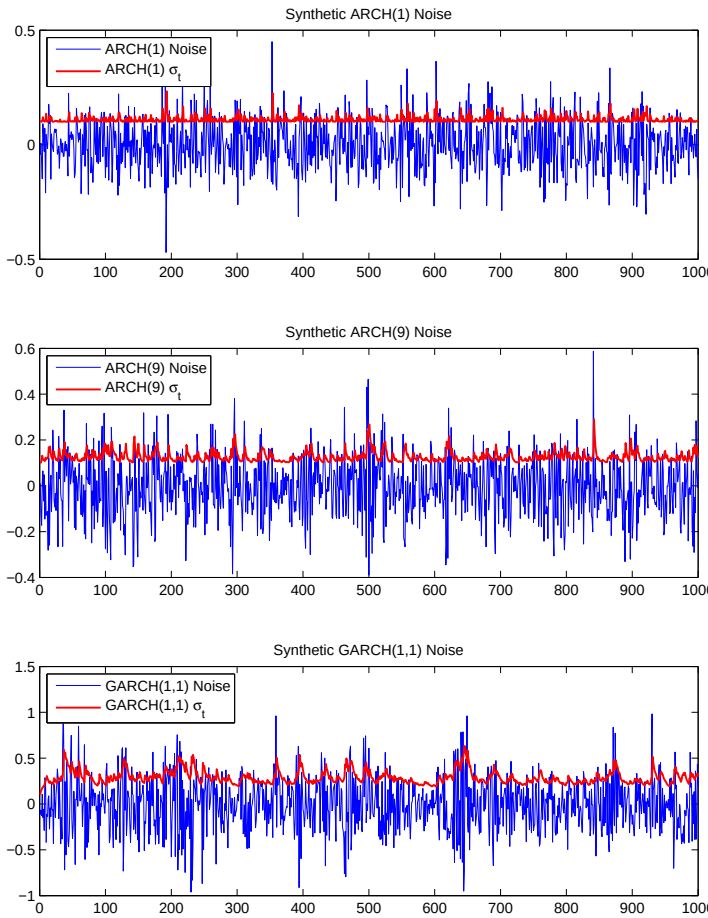

Figure 2.3: The conditional volatility of GARCH is more persistent.

2.7.3 Multivariate GARCH Model

The multivariate noise vector is modeled as

$$\mathbf{w}_t = \mathbf{\Sigma}_t^{1/2}\mathbf{z}_t, \tag{2.60}$$

where $\mathbf{z}_t \in \mathbb{R}^N$ is an i.i.d. white noise series with zero mean and constant covariance matrix $\mathbf{I}$. Then the key part is to model $\mathbf{\Sigma}_t$ conditional on the past information $\mathcal{F}_{t-1}$.

Since the ARCH model is a special case of the GARCH model, we focus on the GARCH model only. There are many different multivariate extensions of the univariate GARCH model, e.g., see [16, 182]. Here we focus on introducing several popular models.

VEC Model

One of the first extensions is the vector (VEC) GARCH model where the conditional covariance matrix linearly depends on some past conditional covariance matrices and the cross-products of some past noise as follows [30]:

$$\text{vech}(\mathbf{\Sigma}_t) = \mathbf{a}_0 + \sum_{i=1}^{m} \tilde{\mathbf{A}}_i \text{vech}(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T) + \sum_{j=1}^{s} \tilde{\mathbf{B}}_j \text{vech}(\mathbf{\Sigma}_{t-j}), \quad (2.61)$$

where m and s are nonnegative integers, the half-vectorization operator $\text{vech}(\cdot)$ denotes an $N(N+1)/2$ dimensional vector by vectorizing only the lower triangular part of its argument N-by-N square matrix, $\mathbf{a}_0$ is an $N(N+1)/2$ dimensional vector, and $\tilde{\mathbf{A}}_i$ and $\tilde{\mathbf{B}}_i$ are $N(N+1)/2$-by-$N(N+1)/2$ parameter matrices. This model is very flexible; however, in general it does not guarantee a positive definite covariance matrix $\mathbf{\Sigma}_t$ at each time and the number of parameters is very large unless N is small.

Diagonal VEC Model

A more parameter parsimonious model is to assume that $\tilde{\mathbf{A}}_i$ and $\tilde{\mathbf{B}}_i$ are diagonal, and the model is referred to as a diagonal VEC (DVEC) model [30], which can be simplified as

$$\mathbf{\Sigma}_t = \mathbf{A}_0 + \sum_{i=1}^{m} \mathbf{A}_i \odot (\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T) + \sum_{j=1}^{s} \mathbf{B}_j \odot \mathbf{\Sigma}_{t-j}, \quad (2.62)$$

where $\mathbf{A}_i, \mathbf{B}_j \in \mathbb{R}^{N\times N}$ are symmetric matrix parameters. Here, the operator $\odot$ denotes the Hadamard product, i.e., the element-wise product,

and $\mathbf{A}_i$ and $\mathbf{B}_j$ can be interpreted as moving weight matrices. However, the DVEC model still may not guarantee a positive definite covariance matrix $\mathbf{\Sigma}_t$ at each time.

BEKK Model

Later, the BEKK model is proposed to guarantee the conditional covariance matrix $\mathbf{\Sigma}_t$ to be positive definite [62]:

$$\mathbf{\Sigma}_t = \mathbf{A}_0\mathbf{A}_0^T + \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T)\mathbf{A}_i^T + \sum_{j=1}^{s} \mathbf{B}_j\mathbf{\Sigma}_{t-j}\mathbf{B}_j^T, \tag{2.63}$$

where m and s are nonnegative integers, $\mathbf{A}_i, \mathbf{B}_j \in \mathbb{R}^{N\times N}$ are matrix parameters, and $\mathbf{A}_0$ is lower triangular. Clearly, this model ensures a positive definite matrix $\mathbf{\Sigma}_t$ provided that $\mathbf{A}_0\mathbf{A}_0^T$ is positive definite; however, now the parameters $\mathbf{A}_i$ and $\mathbf{B}_j$ do not have direct interpretations.

CCC Model

Another model that restricts the number of model parameters and guarantees the positive definite conditional variance estimate is the constant conditional correlation (CCC) model [29]. The underlying idea is to assume that the conditional heteroskedasticity only exists in each asset and their correlations are constant. Mathematically, the conditional covariance matrix $\mathbf{\Sigma}_t$ is decomposed as follows:

$$\mathbf{\Sigma}_t = \mathbf{D}_t\mathbf{C}\mathbf{D}_t \tag{2.64}$$

where $\mathbf{D}_t = \mathrm{Diag}([\sigma_{1,t}, \ldots, \sigma_{N,t}])$ is the time-varying conditional volatilities of each stock and $\mathbf{C}$ is the CCC matrix of the standardized noise vector.

Then the conditional volatilities and correlations are modeled separately. For example, the conditional volatilities are modeled by N univariate GARCH models. Regarding the CCC matrix $\mathbf{C}$, it simply equals the conditional covariance of the following defined standardized noise vector:

$$\boldsymbol{\eta}_t \triangleq \mathbf{D}_t^{-1}\mathbf{w}_t, \tag{2.65}$$

that is,

$$\mathsf{E}\left[\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T|\mathcal{F}_{t-1}\right] = \mathbf{D}_t^{-1}\boldsymbol{\Sigma}_t\mathbf{D}_t^{-1} = \mathbf{C}. \tag{2.66}$$

In practice, the CCC matrix $\mathbf{C}$ is modeled as the covariance matrix of the estimated standardized noise $\hat{\boldsymbol{\eta}}_t \triangleq \hat{\mathbf{D}}_t^{-1}\mathbf{w}_t$ where $\hat{\mathbf{D}}_t$ is the estimated conditional volatilities of each asset [29].

DCC Model

The main limit of the CCC model is that the correlation is constant and there are no spillover and feedback effects across the conditional volatilities. To overcome this drawback, a dynamic conditional correlation (DCC) model is proposed [60]:

$$\boldsymbol{\Sigma}_t = \mathbf{D}_t\mathbf{C}_t\mathbf{D}_t. \tag{2.67}$$

Compared with the CCC model (2.64), the only difference is that now the conditional correlation matrix $\mathbf{C}_t$ is time-dependent.

To ensure that the estimate of the DCC matrix $\mathbf{C}_t$ is a matrix containing correlation coefficients, e.g., diagonal elements equal to 1, Engle [60] modeled it as follows. The ij-th element of DCC matrix is modeled as

$$\rho_{ij,t} = \frac{q_{ij,t}}{\sqrt{q_{ii,t}q_{jj,t}}} \tag{2.68}$$

and then each $q_{ij,t}$ is modeled by a simple GARCH$(1,1)$ model:

$$q_{ij,t} = \alpha(\eta_{i,t-1}\eta_{j,t-1}^T) + (1-\alpha)q_{ij,t-1}. \tag{2.69}$$

Model (2.69) admits a compact matrix notation as follows:

$$\mathbf{Q}_t = \alpha(\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^T) + (1-\alpha)\mathbf{Q}_{t-1} \tag{2.70}$$

and thus

$$\mathbf{C}_t = \mathrm{Diag}^{-1/2}(\mathbf{Q}_t)\mathbf{Q}_t\mathrm{Diag}^{-1/2}(\mathbf{Q}_t). \tag{2.71}$$

2.8 Summary of Different Models and Their Limitations

2.8.1 Summary

Until now we have briefly reviewed most of the basic models for the time series of financial markets, i.e., I.I.D. model, VARMA model, VECM, and multivariate ARCH/GARCH model.

Table 2.1 provides a compact summary of all the models. In practice, the models of conditional mean and covariance matrix can always be combined together, for example, VARMA and GARCH can be used to model the conditional mean and volatility jointly to fit the real financial data better.

2.8.2 Limitations

The previous covered models mainly work for daily, weekly, monthly, or yearly investments and they also have some limitations.

Not Valid for High-Frequency Trading

When the investment interval becomes very small, say several minutes, several seconds or even shorter, the previous models become invalid and one reaches like a "quantum regime" where things are not fluid anymore but quantized into a limit order book. The limit order book contains the list of all kinds of orders with the information of order sign (buy or sell), price, quantity, and timestamp at any given time point, and the records of the dynamics of the limit order book in general are referred to as high-frequency data or tick data. For investments based on high-frequency data, not only do the models (for high-frequency data) matter [97] but also the practical computer and internet communication technologies [4].

Fact 2.1. For high-frequency trading, the computer and internet communication technologies are extremely important. For example, high-frequency trading strategies require the execution of the orders with extremely low latency because high latency may push the price in the adverse direction and reduce the profitability significantly. To re-

Table 2.1: Summary of different financial models.

Model Structure	$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{w}_t$
General Cond. Mean	$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{x}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{r}_{t-i} - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \mathbf{w}_{t-j}$
General Cond. Volatility	$\boldsymbol{\Sigma}_t = \mathbf{A}_0\mathbf{A}_0^T + \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T)\mathbf{A}_i^T + \sum_{j=1}^{s} \mathbf{B}_j \boldsymbol{\Sigma}_{t-j}\mathbf{B}_j^T$

Models	Cond. Mean Model	Cond. Volatility Model
I.I.D. Model	const.: $\boldsymbol{\Pi} = \mathbf{0},\ p = q = 0$	const.: $m = s = 0$
Factor Models	$\mathbf{x}_t = \mathbf{f}_t,\ p = q = 0$	const.: $m = s = 0$
VAR Model	$\boldsymbol{\Pi} = \mathbf{0},\ q = 0$	const.: $m = s = 0$
VMA Model	$\boldsymbol{\Pi} = \mathbf{0},\ p = 0$	const.: $m = s = 0$
VARMA Model	$\boldsymbol{\Pi} = \mathbf{0}$	const.: $m = s = 0$
VECM	$\mathbf{x}_t = \mathbf{y}_{t-1}$	const.: $m = s = 0$
ARCH Model	const.: $\boldsymbol{\Pi} = \mathbf{0},\ p = q = 0$	$s = 0$
GARCH Model	const.: $\boldsymbol{\Pi} = \mathbf{0},\ p = q = 0$	General Cond. Volatility Model

duce such latency, nowadays many stock exchanges, e.g., NASDAQ[3], HKEx[4], etc., provide a "co-location" service that offers all customers the opportunity to co-locate their servers and equipment within the data centers of the stock exchanges. ■

Heavy Tails Issue

Another limitation is that most models implicitly assume a Gaussian distribution for mathematical simplicity [143]. However, for financial data it is known that the financial distributions have heavy tails in practice and the Gaussian assumption may totally fail simply because it predicts the large price changes much less likely than the actual case; see the following Fact 2.2 and the illustrative Example 2.4.

Fact 2.2. The Black-Scholes model [25] is a simple mathematical model and is famous for describing the option prices. However, it completely failed in practice because it assumed a Brownian model which translated into a Gaussian assumption, and underestimated the very possibility of a global crisis. In fact, the abuse of this model led to the crash in October 1987 during which the US market dropped 23% in a single day [31]. ■

Example 2.4. We study the daily log-returns of the S&P500 index from 04-Jan-2010 to 04-Feb-2015. The sample mean and variance are 4.5966×10^{-4} and 1.0199×10^{-4}, respectively.

Figure 2.4 shows the empirical quantiles the log-returns of the S&P500 index versus the theoretical quantiles of the Gaussian distribution $\mathcal{N}(4.5966 \times 10^{-4}, 1.0199 \times 10^{-4})$. The figure is plotted using the MATLAB function `qqplot` which uses symbol '+' to denote the sample data and superimposes a line joining the first and third quartiles of each distribution (this is a robust linear fit of the order statistics of the two samples). We can see that the empirical data has much heavier tails than the Gaussian distribution since the values of the small empirical quantiles are much smaller than the theoretical Gaussian ones

[3] http://www.nasdaqomx.com/transactions/technicalinformation/connectivity
[4] http://www.hkex.com.hk/eng/prod/hosting/hostingservices.htm

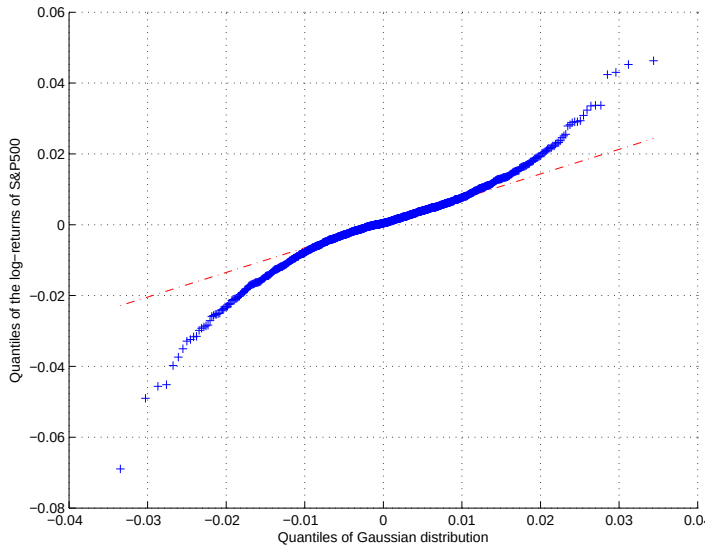

Figure 2.4: Quantile-Quantile plot of the daily log-returns of the S&P500 index versus the Gaussian distribution with the same mean and standard deviation.

and the values of the large empirical quantiles are much larger than the theoretical Gaussian ones. This practical issue is very important as it is very different from signal processing and communications where the noise is typically assumed to be Gaussian. ■

Part of this issue can be overcome by changing the Gaussian assumption to some other distribution with heavier tails, see parameter estimation in Chapter 3.

Lack of Stationarity of Real Data

The lack of stationarity of real data is also a critical limitation. Even if the models were accurate, the parameters defining them would change over time at a pace faster than one can properly estimate. Thus the calibrated models would always be prone to many estimation errors or, even worse, the regime of the market may change and the previously fitted models may be totally wrong [11].

Small Sample Regime

Another limitation, which arises in part from the lack of stationarity of data, is the lack of enough supply of historical data for fitting and estimation purposes, especially when the model dimension is large. This fits into the realm of a small sample regime for high-dimensional data that appears in some big data problems. Some methods to overcome this limitation will be discussed in parameter estimation in Chapter 3.

Other Practical Limitations

Apart from the above limitations, there are many other limitations due to small practical details. For example, some stocks may have a longer history than others, some stocks may not trade exactly the same days as others, and for the daily period it is not clear whether one should use the open price, close price, maximum price, or minimum price of each day. Some sophisticated methods involving different prices have been proposed [80, 84, 209].

Another practical issue the above models do not consider is the liquidity of the asset. This is important for exactly when to execute an order in the market and this will be covered in Chapter 4 order execution.

2.8.3 Concluding Remarks

Practical implementations are more complicated than the nice and clean mathematical models covered in this chapter; however, it is still meaningful to understand them because in principal "all models are false but some models are useful" [171]. It is always necessary to investigate various models with their limitations and thus one can pick up the most useful model for his/her own purposes of investment.

3

Modeling Fitting: Mean and Covariance Matrix Estimators

Models need to be fitted to real data before being used in practice. The previous chapter introduced various time series market models. This chapter focuses on the estimation of the model parameters, more specifically, the mean vector and covariance matrix.

Section 3.1 briefly introduces the practical fitting process and different types of estimation methods. Section 3.2 considers some specific examples for the large sample regime as a warm up, and it is followed by several practical challenges, i.e., the small sample regime in Section 3.3, the heavy tail issue in Section 3.4, and their combination in Section 3.5. At the end, Section 3.6 briefly summarizes all the estimation methods.

3.1 Fitting Process, Types of Estimators, and Main Focus

3.1.1 Fitting Process

Figure 3.1 shows the practical fitting process, roughly speaking, it can be decomposed into two parts: in-sample training and out-of-sample testing [95].

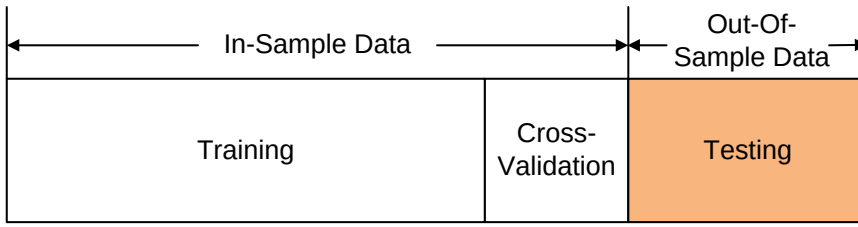

Figure 3.1: Fitting process.

A naive example of portfolio optimization is that, at the end of each month, one can always use the sample covariance matrix of the past one year daily returns of multiple stocks as the covariance matrix estimate and then compute the minimum-variance portfolio and hold it in the upcoming month to investigate the out-of-sample performance. The data of the past one year daily returns is the in-sample data and the data of the daily returns in the upcoming month is the out-of-sample data.

The above example is only an oversimplified example; in practice, to improve the out-of-sample testing results, there may exist some tuning parameters (e.g., see the shrinkage trade-off parameters in the shrinkage estimators in Section 3.3 later) in the estimators and the in-sample training can be further decomposed into two steps: i) split the in-sample data into training data and cross-validation data and fit the training data to the model with different (discretely sampled) tuning parameters, and ii) find out the parameter that gives the best cross-validation criterion of interest and then fit the in-sample data as a whole (i.e., the training and validation data together) to the model with the selected tuning parameter. Step i) of choosing the optimal tuning parameter is usually referred to as the cross-validation method in the literature [95]. After the in-sample training, one can conduct the out-of-sample testing.

3.1.2 Different Types of Estimators

In statistics, an estimator is simply a function of the current information (i.e., the observations) that computes a quantity of interest (e.g., the mean, the covariance matrix, or the other model parameters). The computed (or estimated) value is referred to as the estimate. Roughly speaking, there are three main types of estimators [143].

Non-parametric Estimators. Non-parametric estimators do not assume the observations follow any specific distribution but estimate the quantity of interest from the observations based on the law of large numbers. For example, the sample mean and the sample covariance matrix are two typical non-parametric estimators. In general, a large number of samples is required to ensure a low estimation error.

Maximum Likelihood Estimators (MLEs). In practice, the number of samples may not be large enough, and non-parametric methods may not provide satisfactory estimates. An alternative method is the parametric approach, that is, we first assume that the observations follow an underlying distribution with some unknown parameters and then define the estimates as the maximizer of the likelihood of the observations over the unknown parameters. For obvious reasons, these estimators are referred to as maximum likelihood estimators.

Shrinkage-Bayesian Estimators. For some applications, the number of samples may be too small compared to the data dimension and both the non-parametric and maximum likelihood (ML) estimators may not provide reliable estimates. In the literature, there are two (related) ways to improve the estimates. The first approach is to shrink the estimate to a given fixed target to get a new estimate. Some examples are the shrinkage covariance matrix in financial engineering and the diagonally loaded interference-plus-noise covariance matrix in signal processing. The second approach is to incorporate some Bayesian prior information into an estimator by adding a proper regularization term to the selected likelihood function (since combining the likelihood with the prior information simply results in a posterior distribution, this

method is also referred to as the maximum a posterior (MAP) estimation method). These two methods are closely related in the sense that a shrinkage estimator can usually be alternatively derived by adding a proper regularization term to a specific likelihood function.

Comparison on Different Types of Estimators

In general, when the number of samples is large enough, the non-parametric estimators already perform well due to the law of large numbers, the MLEs also work fine assuming the distribution of the observations is not far away from the underlying true one, and the Shrinkage-Bayesian may underperform if the observed data does not quite fit the assumed prior or shrinkage target. When the number of samples is medium, then the non-parametric estimators degenerate too much and MLEs are still relatively reliable. When the number of samples becomes too small, neither of the estimators are reliable and shrinking to some target or incorporating some prior information into the estimator usually improves the estimation quality to some degree.

3.2 Warm Up: Large Sample Regime

In this section we start with the large sample regime for the I.I.D. model under different distribution assumptions as a warm up and then we point out the real challenges faced in practice.

3.2.1 I.I.D. Model

Let us start with the simplest I.I.D. model, i.e.,

$$\mathbf{r}_t = \boldsymbol{\mu} + \mathbf{w}_t, \tag{3.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean and $\mathbf{w}_t \in \mathbb{R}^N$ is a white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}$.

Suppose we have T observations of the log-returns $\mathbf{r}_t$, $t = 1, \ldots, T$ and they are drawn according to (3.1). Then estimating the model simply means estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Sample Mean and Sample Covariance Matrix

Intuitively, the most straightforward estimators are the sample averages. The sample mean is

$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t, \tag{3.2}$$

and the sample covariance is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \hat{\boldsymbol{\mu}})(\mathbf{r}_t - \hat{\boldsymbol{\mu}})^T. \tag{3.3}$$

The popularity of such estimators comes from the law of large numbers (LLN) that under fairly general conditions the sample average estimates approximate the true expectations and the approximation accuracy increases as the number of samples increases, e.g., $\hat{\boldsymbol{\mu}} \to \boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}} \to \boldsymbol{\Sigma}$ as $T \to +\infty$ [101, 140].

Least-Square (LS) Estimation

We can first estimate the mean via minimizing the least-square error in the T observed i.i.d. samples, that is,

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{r}_t - \boldsymbol{\mu}\|_2^2\,. \tag{3.4}$$

Setting the derivative of the objective w.r.t. $\boldsymbol{\mu}$ yields as the optimal solution the same as the sample mean stated in (3.2). Then, the sample covariance matrix of the residuals coincides with the sample covariance matrix stated in (3.3).

Note that both the sample average and LS estimation methods do not assume the specific distribution of $\mathbf{r}_t$, thus they belong to the non-parametric approach.

ML Estimation

If we assume the underlying distribution is known, then the MLE can be employed. Here, we assume $\mathbf{r}_t$ are i.i.d. and follow an elliptical distribution [141, 140, 101]:

$$\mathbf{r}_t \sim \mathrm{EL}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g), \tag{3.5}$$

where $\boldsymbol{\mu} \in \mathbb{R}^N$ is a mean vector, $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ is a positive definite dispersion (or scatter) matrix and g is a probability density generator that mainly determines the thickness of the tails. The corresponding pdf function is given as follows:

$$f(\mathbf{r}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} g\left((\mathbf{r}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})\right). \tag{3.6}$$

The problem of model estimation consists of estimating the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the observations by maximizing the likelihood (3.6) as a function of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for given observations.

Given the T i.i.d. samples $\mathbf{r}_t$, $t = 1,\ldots,T$, the negative log-likelihood of such T samples is

$$\ell(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\log\prod_{t=1}^{T} f(\mathbf{r}_t) \tag{3.7}$$

$$= \frac{T}{2}\log|\boldsymbol{\Sigma}| - \sum_{t=1}^{T}\log\left(g\left((\mathbf{r}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{r}_t-\boldsymbol{\mu})\right)\right) \tag{3.8}$$

and then the estimates are the minimizer of $\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})$, i.e.,

$$(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \arg\min_{\boldsymbol{\mu},\,\boldsymbol{\Sigma}\succ\mathbf{0}} \ell(\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{3.9}$$

For clarity of presentation, first denote

$$d_t \triangleq (\mathbf{r}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{r}_t-\boldsymbol{\mu}). \tag{3.10}$$

Then finding the derivative of $\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ yields

$$\frac{\partial\ell}{\partial\boldsymbol{\mu}} = -\sum_{t=1}^{T}\frac{\partial\log\left(g(d_t)\right)}{\partial\boldsymbol{\mu}} = -\sum_{t=1}^{T}\frac{g'(d_t)}{g(d_t)}\frac{\partial d_t}{\partial\boldsymbol{\mu}} \tag{3.11}$$

$$= -\sum_{t=1}^{T} 2\frac{g'(d_t)}{g(d_t)}\boldsymbol{\Sigma}^{-1}\left(\mathbf{r}_t-\boldsymbol{\mu}\right) \tag{3.12}$$

$$\frac{\partial\ell}{\partial\boldsymbol{\Sigma}^{-1}} = -\frac{T}{2}\boldsymbol{\Sigma} - \sum_{t=1}^{T}\frac{g'(d_t)}{g(d_t)}\left(\mathbf{r}_t-\boldsymbol{\mu}\right)\left(\mathbf{r}_t-\boldsymbol{\mu}\right)^T \tag{3.13}$$

and setting both (3.12) and (3.13) to zero yields[1]

$$\boldsymbol{\mu} = \sum_{t=1}^{T} \frac{w(d_t)}{\sum_{i=1}^{T} w(d_i)} \mathbf{r}_t \tag{3.15}$$

$$\boldsymbol{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} w(d_t) \left(\mathbf{r}_t - \boldsymbol{\mu}\right) \left(\mathbf{r}_t - \boldsymbol{\mu}\right)^T \tag{3.16}$$

where

$$w(x) \triangleq -2 \frac{g'(x)}{g(x)} = \left(-2 \log g(x)\right)'. \tag{3.17}$$

Note that $\boldsymbol{\Sigma}$ satisfying (3.16) must be positive definite (with probability one) when T is large enough (i.e., $T \geq N+1$), thus the solutions of (3.15) and (3.16) are the minimizers of $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Gaussian Distribution. Note that the Gaussian distribution is a special case of the elliptical distribution with

$$g^{\mathrm{G}}(x) \triangleq \frac{e^{-x/2}}{(2\pi)^{N/2}}, \tag{3.18}$$

and from (3.17) we have

$$w^{\mathrm{G}}(x) = 1. \tag{3.19}$$

Interestingly, we can see that the relationships (3.15) and (3.16) reduce to the sample averages (3.2) and (3.3), respectively. Thus, when estimating mean and covariance, both the non-parametric least-square estimation and the parametric MLE under Gaussian assumption coincide with the sample average estimations.

Student-*t* Distribution. As mentioned before, the financial noise usually have heavier tails than the Gaussian assumption. In practice,

[1] An implicit expression of (3.15) is

$$\mathbf{0} = \sum_{t=1}^{T} w(d_t) \left(\mathbf{r}_t - \boldsymbol{\mu}\right). \tag{3.14}$$

this characteristic can be captured by an elliptical distribution called Student-t distribution and the density generator function reads

$$g^{\mathrm{S}}(x) \triangleq \frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{N/2}}(1+x/\nu)^{-\frac{1+N}{2}}, \tag{3.20}$$

and from (3.17) we have

$$w^{\mathrm{S}}(x) = \frac{\nu+N}{\nu+x}. \tag{3.21}$$

Here, the parameter $\nu > 0$ is the degree of freedom: the smaller ν is, the heavier the tails are. It can be shown that the Student-t converges to the Gaussian distribution, i.e., $g^{\mathrm{S}}(x) \to g^{\mathrm{G}}(x)$ as $\nu \to +\infty$ [141].

Cauchy Distribution. A special case of the Student-t distribution is $\nu = 1$ and the density generator function is

$$g^{\mathrm{C}}(x) \triangleq \frac{\Gamma\left(\frac{1+N}{2}\right)}{\Gamma\left(\frac{1}{2}\right)(\pi)^{N/2}}(1+x)^{-\frac{1+N}{2}}, \tag{3.22}$$

and from (3.17) we have

$$w^{\mathrm{C}}(x) = \frac{1+N}{1+x}. \tag{3.23}$$

Since $\nu = 1 < +\infty$, the Cauchy distribution usually has much heavier tails than the Gaussian distribution.

Gaussian versus Cauchy Distributions. Now we compare the MLE of Gaussian and Cauchy distributions. Figure 3.2 shows the one-dimensional pdfs of the standard Gaussian and the standard Cauchy distributions. We can see that the standard Gaussian distribution is thin-tailed and the standard Cauchy distribution is heavy-tailed. Then we can interpret their weights in (3.19) and (3.23) as follows.

The Gaussian weights $w^{\mathrm{G}}(d_t)$ are constant from (3.19), this is because the Gaussian distribution is very thin-tailed and the observations with large d_t (recall that $d_t = (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{r}_t - \boldsymbol{\mu})$ given in (3.10)) are relatively rare. So if an observation is far away from the mean position

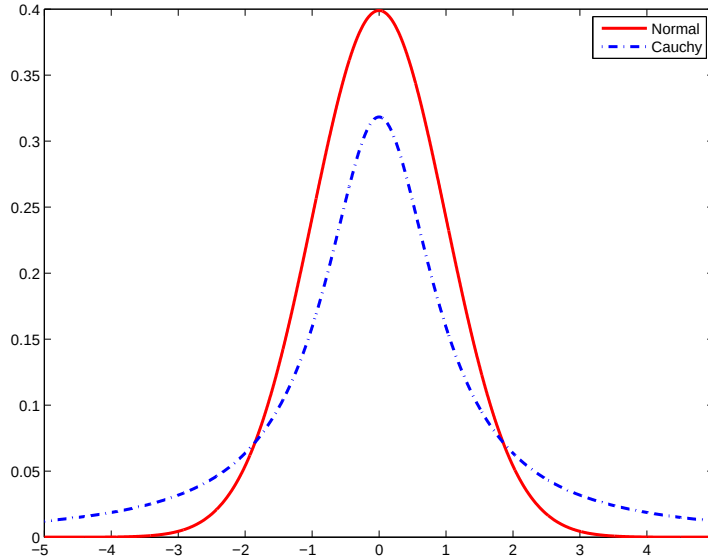

Figure 3.2: Comparison of one-dimensional Gaussian and Cauchy distributions.

(i.e., d_t is large), the only reason is that the dispersion is large and thus the Gaussian MLE assigns all the samples the same weights.

Compared with the Gaussian weights $w^{\mathrm{G}}(d_t)$, the Cauchy weights $w^{\mathrm{C}}(d_t)$ are smaller for the extreme events (i.e., the observations with large d_t) from (3.23). This is because the Cauchy distribution is heavy-tailed and the observations with large d_t are relatively more frequent and then the Cauchy MLE tends to give the extreme events smaller weights to diminish their negative effect that would otherwise distort the estimates.

From the above comparison, we can see that the MLE of the Gaussian distribution is more easily affected by extreme observations or outliers (because the outliers usually have large d_t as well) and the MLE of the Cauchy distribution is more robust to the extreme observations and outliers. This is an important property and will be explored in Sections 3.4 and 3.5 later. Nevertheless, here we still use an illustrative example to show the importance of the robust estimation.

Example 3.1. Suppose the dimension is $N = 2$ and we draw $T = 40$ i.i.d. samples $\mathbf{r}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, \tag{3.24}$$

and four i.i.d. outliers from $\mathcal{N}\left(\begin{bmatrix} -2 \\ 2 \end{bmatrix}, \mathbf{\Sigma}\right)$. We assume we know the underlying distribution is Gaussian with the mean known and we aim to estimate the covariance only.

Figure 3.3 shows the result of a specific realization. We can see that the Gaussian MLE is too sensitive to the outliers while the Cauchy MLE is more robust and provides much better estimation.

Gaussian MLE: The Gaussian MLE is computed from the sample covariance (3.3) with given mean $\boldsymbol{\mu} = \mathbf{0}$:

$$\hat{\mathbf{\Sigma}}^{\mathrm{G}} = \begin{bmatrix} 1.4407 & 0.4552 \\ 0.4552 & 1.1807 \end{bmatrix}. \tag{3.25}$$

Cauchy MLE: First we can find the solution of (3.16) with given mean $\boldsymbol{\mu} = \mathbf{0}$ and weight functions defined as (3.23), and we denote it as $\hat{\mathbf{\Sigma}}^{\mathrm{Shape}}$. Then the estimated covariance matrix of the underlying Gaussian distribution is $\hat{\mathbf{\Sigma}}^{\mathrm{C}} = \hat{\mathbf{\Sigma}}^{\mathrm{Shape}}/c$ where c is the size parameter defined as the solution of (3.55)[2]. We will see this procedure more clearly in Section 3.4.1. Here, we have

$$\hat{\mathbf{\Sigma}}^{\mathrm{C}} = \begin{bmatrix} 1.0351 & 0.7584 \\ 0.7584 & 1.0127 \end{bmatrix}. \tag{3.26}$$

Numerically, the Cauchy MLE $\hat{\mathbf{\Sigma}}^{\mathrm{C}}$ is much closer to the true covariance $\mathbf{\Sigma}$ than the Gaussian MLE $\hat{\mathbf{\Sigma}}^{\mathrm{G}}$. It verifies the result in Figure 3.3.

The MATLAB code of this example is included in Appendix A. ■

3.2.2 Other Models and Main Focus

The estimations of the other models more or less follow a similar procedure: first always rewrite the noise in terms of observations and pa-

[2]For this example, given $N = 2$, $w_2(x) = w^{\mathrm{C}}(x) = \frac{N+1}{1+x}$, solving (3.55), i.e., $\int_0^{+\infty} w^{\mathrm{C}}\left(\frac{x}{c}\right) \frac{x}{c} \chi_N^2(x) dx = N$, yields $c = 0.4944$.

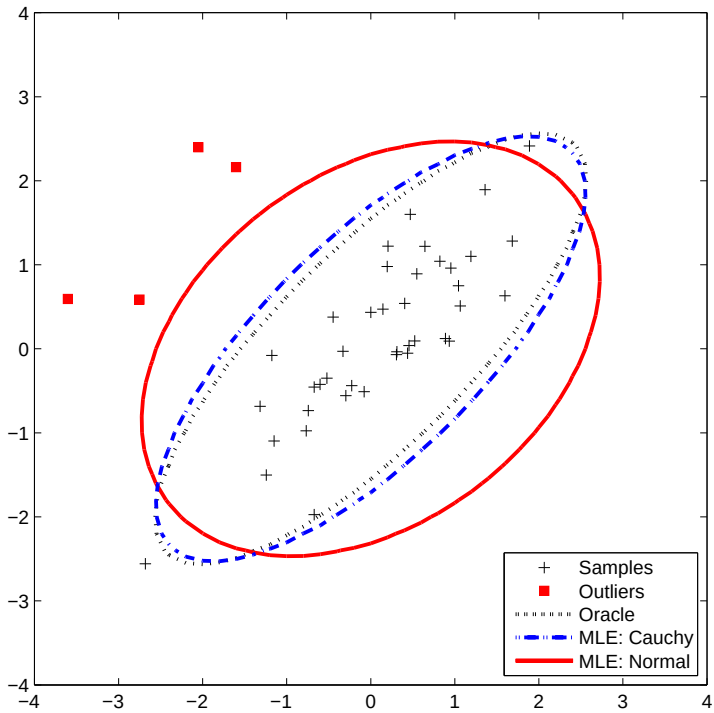

Figure 3.3: Comparison of Gaussian and Cauchy MLEs.

rameters, then if the pdf of the noise is known employ MLE, otherwise use a simple LS estimation. For example, the LS estimation method is used to estimate the parameters of an i.i.d. model, i.e., the conditional mean vector and covariance matrix. It is also widely used to estimate the linear coefficients of a VAR model and is based on which the conditional mean vector and covariance matrix can be computed directly [129, 197]. And interestingly, as shown before, the LS estimation methods coincides with the maximum likelihood estimation under the Gaussian noise assumption. For the estimation of different multivariate time series models, i.e., VAR, VARMA, VECM, and GARCH, the book [129] provides a good and comprehensive summary where the LS estimation and/or MLE of each model are explained in detail.

Also, to overcome the possible over-fitting or over-parameterization

issues due to outliers or lack of samples, the same idea of shrinkage-Bayesian is employed for the parameter estimation of different models, e.g., the i.i.d. model [53, 105, 187, 120, 121, 122], the factor model [40, 71], the VAR/VMA/VARMA model [15, 116, 126, 47, 186, 152, 129], the ARCH/GARCH model [72], etc.

Therefore, for clarity of presentation, the scope of this chapter is not to restate the existing well-developed estimation procedures for all the models but is to focus on the state-of-the-art estimation of the mean and covariance matrix that lies at the heart of all fitting methods. In fact, the estimation of the covariance matrix is of paramount importance in the financial engineering industry, as illustrated by the following fact.

Fact 3.1. Estimating the covariance or correlation between different assets is very important in practice. Accurate covariance enables one to make optimal portfolio optimization decisions and to control the risk better. In industry there are even some financial firms consulting on estimating the covariance. For example, see Studdridge International[3] which "is a high-end consulting firm specialized in estimating large-dimensional covariance matrices, and in exploiting the information they contain to make optimal decisions". ■

3.2.3 Real Challenges

So far we have introduced the sample average estimators and the MLE. Unfortunately, those two estimation methods are not reliable in practice due to the following real challenges.

- Small sample regime: when the number of samples is not enough compared to the dimension of the log-return vector (i.e., the number of stocks considered), the sample covariance may not even be invertible. Even if it is invertible, it may still not be well-conditioned and taking matrix inversion will amplify any tiny errors in the estimated covariance matrix. This becomes a practical challenge because of the rise of big data analysis in various applications, including financial engineering, signal processing,

[3] http://studdridge.com/what-we-do/

bioinformatics, etc. In our context, one manifestation of big data refers to the high dimensionality or large size of the universe of stocks.

- Heavy tails issue: another critical issue in financial engineering is that the distributions of the log-returns are always heavy-tailed (cf. Section 3.4). Thus, the widely used Gaussian assumption does not hold in practice and it does not lead to a proper fitting of real data. The traditional estimators based on the Gaussian assumption are too sensitive to extreme events and outliers, and as a consequence the estimates are distorted too much to be reliable.

In the following, we will explore and connect the recent different methods developed in both financial engineering and signal processing that deal with the above two issues.

3.3 Small Sample Regime: Shrinkage Estimators

When the number of samples is small compared with the data dimension, the total mean squared errors (MSEs) of the sample average estimators are mainly from the variances rather than the biases of the estimators [143]. It is well-known that lower MSEs can be achieved by allowing for some biases [56]. This can be implemented by shrinking the sample estimators to some known target values.

Generally speaking, the shrinkage estimator has the following form:

$$\tilde{\boldsymbol{\theta}} = \rho \mathbf{T} + (1 - \rho)\hat{\boldsymbol{\theta}}, \tag{3.27}$$

where $\hat{\boldsymbol{\theta}}$ is the sample average estimator (i.e., it can be either sample mean or sample covariance matrix), $\mathbf{T}$ is a target which can either be given or have a specific structure (e.g., it can be an identity matrix up to a scalar for the covariance matrix estimation), ρ is the shrinkage trade-off parameter, and $\tilde{\boldsymbol{\theta}}$ is the shrinkage estimator.

Now the critical problem is how to choose the shrinkage trade-off parameter ρ (and sometimes the target $\mathbf{T}$ as well) so that the MSE or some other criteria of interest is minimized. In general, there are two different approaches of finding the optimal shrinkage trade-off parameter: random matrix theory (RMT) and cross-validation.

RMT. This is a theoretical approach and the idea is to first assume the true parameter (to be estimated) is known and formulate a problem that minimizes the ideal criterion of interest. However, in practice the true parameter is never known and then under some technical assumptions and conditions the RMT is employed to either get the asymptotically optimal trade-off parameter in closed-form expression or derive an easy to solve numerical optimization problem. The advantage of this approach is that the (in-sample) asymptotically optimal trade-off parameter can be computed directly and efficiently. However, we need to point out that the (in-sample) asymptotically optimal trade-off parameter does not guarantee the best out-of-sample test result. Due to the mathematical simplicity the RMT has also found various applications in signal processing and wireless communication fields, e.g., see [199].

Cross-Validation. This is the numerical method introduced in Section 3.1.1. This approach tends to provide better out-of-sample results since the cross-validation is used to exhaustively search for the best trade-off parameter. However, it is also computationally more intensive since it requires to compute the estimates with many different shrinkage trade-off parameter values in the cross-validation step.

3.3.1 Shrinkage Mean

It is well-known from the central limit theorem that

$$\hat{\boldsymbol{\mu}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{T}\right) \tag{3.28}$$

and thus the MSE of $\hat{\boldsymbol{\mu}}$ is

$$\mathsf{E}\,\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = \frac{1}{T}\mathrm{Tr}(\boldsymbol{\Sigma}). \tag{3.29}$$

Sharing the same form as (3.27), the James-Stein shrinkage estimator [53, 105, 187] for the mean aims at shrinking the sample mean to a target $\mathbf{b}$:

$$\tilde{\boldsymbol{\mu}} = \rho\mathbf{b} + (1-\rho)\hat{\boldsymbol{\mu}}. \tag{3.30}$$

It is shown that a choice of ρ so that $\mathsf{E}\,\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \leq \mathsf{E}\,\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2$ is [143]

$$\tilde{\rho} = \frac{1}{T}\frac{N\tilde{\lambda} - 2\lambda_1}{\|\hat{\boldsymbol{\mu}} - \mathbf{b}\|_2^2} \tag{3.31}$$

if it is positive, otherwise it is zero. Here, $\tilde{\lambda}$ and λ_1 are the average and the largest value of the eigenvalues of $\boldsymbol{\Sigma}$, respectively. Intuitively, $\tilde{\rho}$ vanishes as T increases, and the shrinkage estimator gets closer to the sample mean.

Apart from any fixed $\mathbf{b}$ independent of the observations, some other examples of $\mathbf{b}$ are $\frac{\mathbf{1}^T\hat{\boldsymbol{\mu}}}{N}\mathbf{1}$, which is the scenario-dependent grand mean [143]; and $\frac{\mathbf{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}{\mathbf{1}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}\mathbf{1}$, which is a volatility-weighted grand mean where $\hat{\boldsymbol{\Sigma}}$ is an estimator of the covariance matrix [110].

Example 3.2. We set $N = 40$ and draw $T = 10, 20, \ldots, 80$ i.i.d. samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$. Suppose $\boldsymbol{\Sigma}$ is known exactly, we compare the sample mean with three shrinkage estimators: i) the constant target $\mathbf{b} = 0.2 \times \mathbf{1}$, ii) the scenario-dependent (SD) target $\mathbf{b} = \frac{\mathbf{1}^T\hat{\boldsymbol{\mu}}}{N}\mathbf{1}$, and iii) the volatility-weighted (VW) target $\mathbf{b} = \frac{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}\mathbf{1}$.

Figure 3.4 shows the numerical results where the MSE is averaged over 200 realizations. We can see that the sample mean has a numerical MSE close to the theoretical one $\frac{1}{T}\mathrm{Tr}(\boldsymbol{\Sigma})$, and all the shrinkage estimators outperform the sample mean. Among them, the shrinkage estimator with the scenario-dependent and volatility-weighted targets outperforms the one with the constant target. ■

3.3.2 Shrinkage Scatter Matrix Based on RMT

Now we assume the mean is known exactly and the goal is to estimate the scatter (or dispersion, or covariance if it exists) matrix. For the shrinkage scatter matrix, the identity matrix in general is selected as the target, and there exist many works aiming at selecting the shrinkage trade-off parameter according to different criteria.

MSE

For example, Ledoit and Wolf [121] aimed at finding the linear combination of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the identity matrix

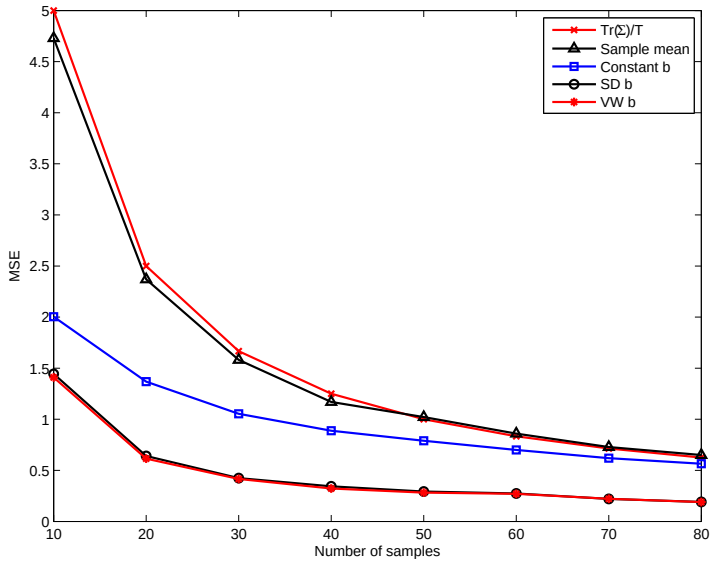

Figure 3.4: Shrinkage mean estimations.

such that the expected quadratic loss between the estimation and the (unknown) true covariance $\mathbf{\Sigma}$ was minimized:

$$\begin{aligned} \underset{\rho_1,\rho_2}{\text{minimize}} \quad & \mathsf{E}\left\|\tilde{\mathbf{\Sigma}}-\mathbf{\Sigma}\right\|_F^2 \\ \text{subject to} \quad & \tilde{\mathbf{\Sigma}}=\rho_1\mathbf{I}+\rho_2\hat{\mathbf{\Sigma}}. \end{aligned} \tag{3.32}$$

Suppose the true covariance matrix $\mathbf{\Sigma}$ is known, problem (3.32) is a quadratic programming and the variables are two scalars. The optimal solution admits a closed-form as follows.

Theorem 3.1 ([121, Theorem 2.1])**.** Problem (3.32) admits the optimal solution

$$\tilde{\mathbf{\Sigma}}^\star=\tilde{\rho}\tilde{\lambda}\mathbf{I}+(1-\tilde{\rho})\hat{\mathbf{\Sigma}}, \tag{3.33}$$

where

$$\tilde{\lambda}=\frac{\mathrm{Tr}(\mathbf{\Sigma})}{N} \quad \text{and} \quad \tilde{\rho}=\frac{\mathsf{E}\left\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\right\|_F^2}{\mathsf{E}\left\|\hat{\mathbf{\Sigma}}-\tilde{\lambda}\mathbf{I}\right\|_F^2}. \tag{3.34}$$

Unfortunately, the true covariance $\mathbf{\Sigma}$ is not known in practice, hence, $\tilde{\lambda}$ and $\tilde{\rho}$ are not directly computable. Ledoit and Wolf further proposed the consistent estimators of $\tilde{\lambda}$ and $\tilde{\rho}$ as follows:

$$\hat{\tilde{\lambda}} = \frac{\mathrm{Tr}(\hat{\mathbf{\Sigma}})}{N}, \tag{3.35}$$

$$\hat{\tilde{\rho}} = \min\left(\frac{1}{T}\frac{\frac{1}{T}\sum_{t=1}^{T}\mathrm{Tr}(\mathbf{r}_t\mathbf{r}_t^T - \hat{\mathbf{\Sigma}})^2}{\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \hat{\tilde{\lambda}}\mathbf{I})^2}, 1\right). \tag{3.36}$$

Then simply replacing $\tilde{\lambda}$ and $\tilde{\rho}$ in (3.33) with $\hat{\tilde{\lambda}}$ and $\hat{\tilde{\rho}}$ yields a consistent estimator of $\tilde{\mathbf{\Sigma}}^{\star}$. Intuitively, $\hat{\tilde{\rho}}$ vanishes as T increases, and the shrinkage estimator becomes closer to the sample covariance estimator. The results of (3.35) and (3.36) were derived based on the RMT, which is also a popular quantitative tool in signal processing and wireless communication, e.g., see [199].

Interestingly, the idea of shrinking the sample covariance matrix to the identity matrix has also been widely used in array signal processing and is referred to as diagonal loading, e.g., see [1, 45, 38]. However, the trade-off parameter is usually chosen in an ad hoc way, e.g., the diagonal loading matrix may be chosen as $\hat{\mathbf{\Sigma}}+10\sigma^2\mathbf{I}$ where σ^2 is the noise power in a single sensor [204]. Here, (3.33) provides a more sensible way to select the trade-off parameter.

Example 3.3. We use the same parameter settings as Example 3.2, but now we assume the mean is known and we want to estimate the covariance matrix. For the Gaussian case, it can be shown that [143, 121]

$$\mathsf{E}\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F^2 = \frac{1}{T}\left(\mathrm{Tr}(\mathbf{\Sigma}^2) + \left(1 - \frac{1}{T}\right)(\mathrm{Tr}(\mathbf{\Sigma}))^2\right), \tag{3.37}$$

$$\mathsf{E}\left\|\hat{\mathbf{\Sigma}} - \tilde{\lambda}\mathbf{I}\right\|_F^2 = \mathsf{E}\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F^2 + \left\|\mathbf{\Sigma} - \tilde{\lambda}\mathbf{I}\right\|_F^2. \tag{3.38}$$

If we knew $\mathbf{\Sigma}$, we could obtain $\tilde{\mathbf{\Sigma}}^{\star}$ in (3.33) directly. It is referred to as the "oracle" estimator since $\mathbf{\Sigma}$ is never known in practice. The practical estimator obtained by replacing $\tilde{\lambda}$ and $\tilde{\rho}$ in (3.33) with $\hat{\tilde{\lambda}}$ and $\hat{\tilde{\rho}}$ is referred to as the Ledoit-Wolf (LW) estimator.

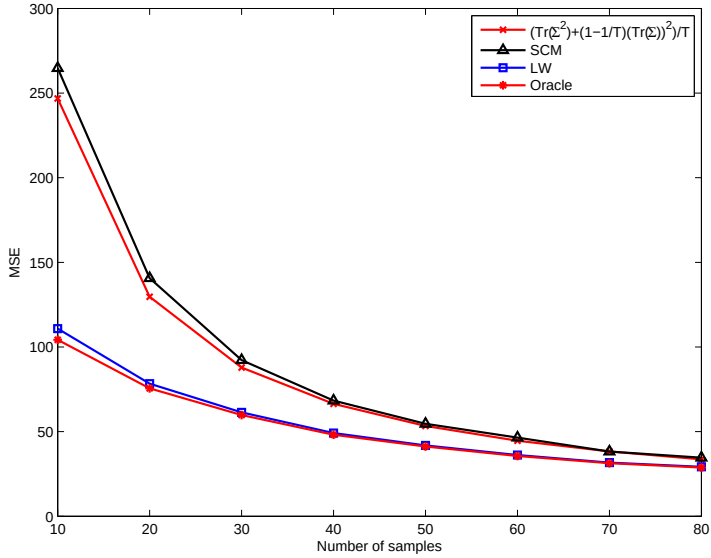

Figure 3.5: Shrinkage covariance estimation.

Figure 3.5 shows the numerical results where the MSEs are averaged over 200 realizations. We can see that the sample covariance has a numerical MSE close to the theoretical one, i.e., (3.37) and the shrinkage LW estimator outperforms the sample covariance. Interestingly, the LW estimator performs closely to the oracle estimator. ■

Quadratic Loss of Precision Matrix

For many cases, it is the precision matrix (i.e., the inverse of the dispersion or covariance matrix) that is used in practice, e.g., see the minimum variance (MV) portfolio

$$\mathbf{w}^{\mathrm{MV}} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}, \tag{3.39}$$

which is the optimal solution of (1.3) introduced before.

Since the inversion operation can dramatically amplify the estimation error, for applications similar to the minimum variance portfolio, it is more sensible to minimize the estimation error in the precision

matrix directly instead of minimizing the estimation error in the covariance matrix.

Based on the shrinkage structure, Zhang et al. [212] considered the problem of minimizing the quadratic loss of the precision matrix directly as follows:

$$\begin{array}{ll} \underset{\rho\geq 0,\, \mathbf{T}\in\mathcal{D}_+}{\text{minimize}} & \frac{1}{N}\left\|\tilde{\mathbf{\Sigma}}^{-1}-\mathbf{\Sigma}^{-1}\right\|_F^2 \\ \text{subject to} & \tilde{\mathbf{\Sigma}}=\rho\mathbf{I}+\frac{1}{T}\mathbf{R}\mathbf{T}\mathbf{R}^T, \end{array} \tag{3.40}$$

where $\mathbf{R}=[\mathbf{r}_1,\ldots,\mathbf{r}_T]\in\mathbb{R}^{N\times T}$ is the data matrix of T observations, and $\mathbf{T}\in\mathcal{D}_+$ is a T-by-T nonnegative and diagonal weight matrix.

Even if the true covariance matrix $\mathbf{\Sigma}$ is known, problem (3.40) is much harder than problem (3.32) because the objective of problem (3.40) cannot be explicitly computed anymore.

Under some technical conditions, Zhang et al. [212] showed that asymptotically there exists a global optimal solution of the form ($\rho, \mathbf{T}=\alpha\mathbf{I}$) and derived the following asymptotic problem:

$$\begin{array}{ll} \underset{\rho\geq 0,\,\alpha\geq 0}{\text{minimize}} & \frac{1}{N}\left\|\tilde{\mathbf{\Sigma}}^{-1}-\hat{\mathbf{\Sigma}}^{-1}\right\|_F^2 \\ & +\frac{2}{N}\mathrm{Tr}\left(\rho^{-1}\left(\hat{\tilde{\delta}}\tilde{\mathbf{\Sigma}}^{-1}-(1-c_N)\hat{\mathbf{\Sigma}}^{-1}\right)+\hat{\mathbf{\Sigma}}^{-1}\tilde{\mathbf{\Sigma}}^{-1}\right) \\ & -(2c_N-c_N^2)\frac{1}{N}\mathrm{Tr}\left(\hat{\mathbf{\Sigma}}^{-2}\right) \\ & -(c_N-c_N^2)\left(\frac{1}{N}\mathrm{Tr}\left(\hat{\mathbf{\Sigma}}^{-1}\right)\right)^2 \\ \text{subject to} & \tilde{\mathbf{\Sigma}}=\rho\mathbf{I}+\alpha\hat{\mathbf{\Sigma}}, \\ & \hat{\tilde{\delta}}=\alpha\left(1-\frac{1}{T}\mathrm{Tr}\left(\alpha\hat{\mathbf{\Sigma}}\tilde{\mathbf{\Sigma}}^{-1}\right)\right), \end{array} \tag{3.41}$$

where $c_N\triangleq\frac{N}{T}$ and $\hat{\tilde{\delta}}$ are intermediate parameters.

We can understand (3.41) as thus, it replaces the unknown true covariance matrix $\mathbf{\Sigma}$ with the explicitly computable sample covariance matrix $\hat{\mathbf{\Sigma}}$ and then adds some correction terms to increase the approximation accuracy.

Problem (3.41) is nonconvex but it can be solved via exhaustive search since there are only two scalar variables $\rho\geq 0$ and $\alpha\geq 0$.

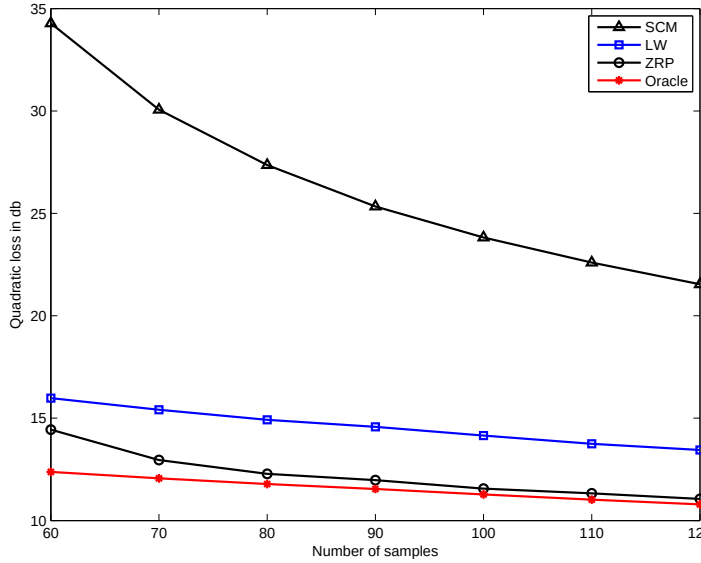

Figure 3.6: Shrinkage precision matrix estimations.

Example 3.4. Suppose $N = 40$ and the i.i.d. samples are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = 0.9^{|i-j|}$. The number of samples T varies as $60, 70, \ldots, 120$. We compare the following four estimators: i) the inverse of the sample covariance matrix, ii) the inverse of the LW covariance estimator, iii) the exhaustive search solution of problem (3.41) over $(\rho, \alpha) \geq \mathbf{0}$, which is referred to as the ZRP estimator, and iv) the exhaustive search solution of problem (3.40) with the structure $(\rho, \alpha\mathbf{I})$ over $(\rho, \alpha) \geq \mathbf{0}$ assuming that $\mathbf{\Sigma}$ is known, is referred to as the "oracle" estimator since $\mathbf{\Sigma}$ is never known in practice.

Figure 3.6 shows the numerical results where the quadratic loss is averaged over 200 realizations. We can see that estimating the precision matrix directly provides lower quadratic losses. ■

Remark 3.1. Note that problem (3.41) requires the sample covariance matrix to be invertible. For the singular case, Zhang et al. [212] studied an alternative loss function called Stein's loss. For simplicity, we have only included the results of quadratic loss here. ■

Sharpe Ratio

All the previous works focus on selecting the shrinkage trade-off parameter to improve the covariance (or precision) estimation accuracy, and recall that the target of an investor is always to achieve better out-of-sample result (e.g., higher realized Sharpe ratio). Even though an accurate covariance (or precision) estimator necessarily leads to a better out-of-sample result, a more sensible approach is to select the shrinkage trade-off parameter so that the out-of-sample criterion of interest is optimized directly.

Here we take the Sharpe ratio (with risk-free return being zero)

$$\mathrm{SR} = \frac{\mathbf{w}^T \boldsymbol{\mu}}{\sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}} \tag{3.42}$$

as the criterion of interest and an optimal solution is given by [65] (also see (5.15) later)

$$\mathbf{w}_{\mathrm{SR}}^{\star} \propto \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \tag{3.43}$$

In practice the true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are never known and the estimates from the training samples are used instead. In [213], the sample mean $\hat{\boldsymbol{\mu}}$ and the shrinkage covariance matrix

$$\tilde{\boldsymbol{\Sigma}} = \rho_1 \mathbf{I} + \rho_2 \hat{\boldsymbol{\Sigma}}, \tag{3.44}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, are used and the resulted portfolio is

$$\hat{\mathbf{w}}_{\mathrm{SR}} \propto \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}. \tag{3.45}$$

Note that the Sharpe ratio (3.42) is scale invariant in $\mathbf{w}$, thus ρ_2 can be arbitrarily set to 1 and the more sensible approach proposed in [213] is to find the shrinkage trade-off parameter ρ_1 such that the realized out-of-sample Sharpe ratio of the portfolio $\hat{\mathbf{w}}_{\mathrm{SR}}$ is maximized:

$$\begin{aligned} \underset{\rho_1 \geq 0}{\text{maximize}} \quad & \frac{\boldsymbol{\mu}^T \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}{\sqrt{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}} \\ \text{subject to} \quad & \tilde{\boldsymbol{\Sigma}} = \rho_1 \mathbf{I} + \hat{\boldsymbol{\Sigma}}. \end{aligned} \tag{3.46}$$

However, the objective of (3.46) is not computable since the true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. Under some technical conditions, the authors of [213] derived an asymptotically equivalent problem based on RMT as follows:

$$\begin{aligned} \underset{\rho_1 \geq 0}{\text{maximize}} \quad & \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} - \hat{\delta}}{\sqrt{b \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}} \\ \text{subject to} \quad & \tilde{\boldsymbol{\Sigma}} = \rho_1 \mathbf{I} + \hat{\boldsymbol{\Sigma}} \\ & D = \frac{1}{T} \text{Tr}\left(\hat{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\Sigma}}^{-1}\right) \\ & \hat{\delta} = D/(1-D) \\ & b = \frac{T}{\text{Tr}\left(\mathbf{W}(\mathbf{I} + \hat{\delta}\mathbf{W})^{-2}\right)} \end{aligned} \tag{3.47}$$

where $\mathbf{W} \triangleq \mathbf{I} - \frac{1}{T}\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{T \times T}$ is a predefined parameter, and D, $\hat{\delta}$, and b are intermediate parameters.

The problem (3.41) can be understood as follows: the unknown true mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are replaced by the explicitly computable sample mean $\hat{\boldsymbol{\mu}}$ and sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and then some correction terms, i.e., $\hat{\delta}$ in the numerator and b in the denominator, are incorporated to increase the approximation accuracy. This problem is still nonconvex but it can be solved via exhaustive search since there is only one scalar variable $\rho_1 \geq 0$.

To investigate the performance of (3.47), let us now consider a real experiment conducted in [213] as follows.

Example 3.5. Let us consider the daily returns of the 45 stocks under Hang Seng Index from 03-Jun-2009 to 31-Jul-2011. The portfolio is updated at each 10 days and the past $T = 75, 76, \ldots, 95$ observations are used to design the portfolios at each update period. The compared portfolios are: i) the method (3.47) based on RMT (referred to as RMT), ii) the portfolio (3.42) based on the Ledoit-Wolf (LW) estimator (referred to as LW), iii) the portfolio (3.42) based on the SCM (referred to as SCM), and iv) the uniform portfolio (referred to as Uniform).

Figure 3.7 shows the out-of-sample Sharpe ratio of the four compared methods. It can be observed that when T changes from 75 to 81,

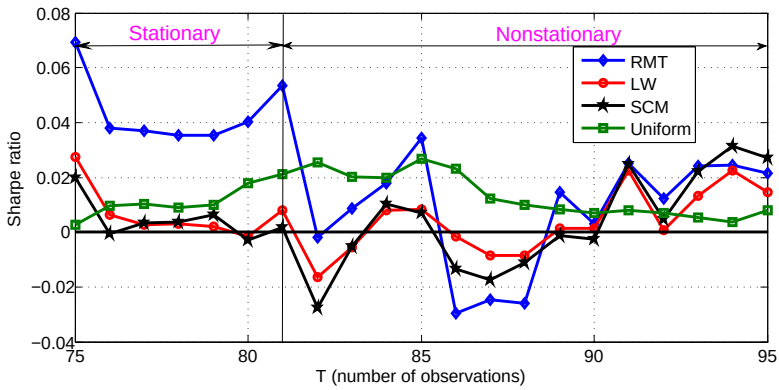

Figure 3.7: Out-of-sample Sharpe ratio of RMT, LW, SCM and Uniform portfolios.

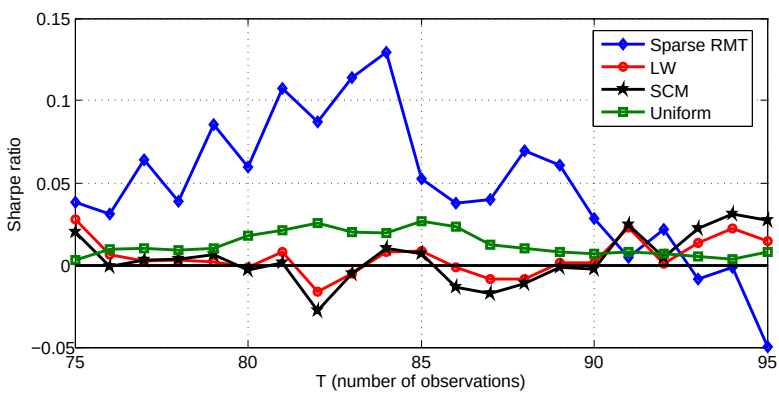

Figure 3.8: Out-of-sample Sharpe ratio of sparse RMT, LW, SCM and Uniform portfolios.

the RMT method outperforms the others, but when $T > 81$, its performance becomes unstable. This is mainly because the mean return and covariance matrix cannot be stationary in a long period (e.g., $T > 81$) [213]. Later an improved RMT portfolio was proposed by setting the weights whose absolute values are less than 5% of the summed absolute values of all the weights to zeros, and this portfolio is referred to as a sparse RMT portfolio. Figure 3.8 shows the out-of-sample Sharpe ratio of the different methods, and it can be seen that the sparse RMT portfolio outperforms all the other methods significantly when T changes from 75 to 90. ■

Remark 3.2. For simplicity we have only considered the Sharpe ratio here. Some other criteria are also studied in the literature in the content of both beamforming design and portfolio optimization, e.g., variance, MSE and SNR for beamforming design [213], and portfolio variance for portfolio optimization [170, 213]. ■

Remark 3.3. There now exist some recent works on including sparsity in the estimates, e.g., the covariance [20, 21, 117] or the precision matrix [82, 99, 211]. In general, some regularization terms are added to propose sparsity (or group sparsity). For example, one widely used regularization is ℓ_1-norm and the technique is usually referred to as LASSO (least absolute shrinkage and selection operator). The book [96] serves as a good summary reference on various topics related to sparsity. Apart from adding the sparsity in the estimation parameters, including sparsity in portfolio optimization is also of interest in some financial problems. Chapter 8 in the later part will demonstrate some widely used techniques to impose sparsity in portfolio optimization. ■

3.4 Heavy Tail Issue: Robust Estimators

From the previous Example 3.1 (see also Figure 3.3) we have already seen that the traditional sample average estimators (or equivalently, the MLE under Gaussian distribution assumption) are very sensitive to the extreme events and outliers; instead, the MLE under heavy-tail assumption (e.g., the Cauchy distribution) provides more robust estimations. In this part, we will explore more general robust estimators.

3.4.1 M-Estimators

Multivariate M-estimators can be defined as a generalization of the MLEs of elliptical distributions [140, 101]. Given the i.i.d. samples $\mathbf{r}_t$, $t = 1, \ldots, T$, the M-estimates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are defined as the solutions to the fixed-point equations

$$\mathbf{0} = \sum_{t=1}^{T} w_1(d_t)\,(\mathbf{r}_t - \boldsymbol{\mu}) \tag{3.48}$$

$$\boldsymbol{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} w_2(d_t)\,(\mathbf{r}_t - \boldsymbol{\mu})\,(\mathbf{r}_t - \boldsymbol{\mu})^T, \tag{3.49}$$

where $d_t = (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r}_t - \boldsymbol{\mu})$ and the weight functions $w_1(x)$ and $w_2(x)$ are both nonnegative, nonincreasing, and continuous functions in $x \in (0, +\infty)$, and they are not necessarily equal. The existence and uniqueness of solutions can be guaranteed under some technical conditions, and the uniqueness requires that $xw_2(x)$ is a strictly increasing function of $x \in (0, +\infty)$ [139, 192]. Suppose the solution to (3.48) and (3.49) exists and is unique, and let us denote it as $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$.

Observe that the elliptical MLE given by (3.15) and (3.16) can be regarded as a special case with $w_1(x) = w_2(x) = -2\frac{g'(x)}{g(x)}$, where g is a density generating function.

Asymptotics

Assume the i.i.d. samples $\mathbf{r}_t \sim \mathrm{EL}(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}}, g)$ where the superscript "o" stands for "oracle". Then as $T \to \infty$, the solution to (3.48) and (3.49), i.e., $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, will converge with probability one to the unique solution, denoted as $(\hat{\boldsymbol{\mu}}_\infty, \hat{\boldsymbol{\Sigma}}_\infty)$, to the fixed-point equations

$$\mathbf{0} = \mathsf{E}\left[w_1(d)\,(\mathbf{r} - \boldsymbol{\mu})\right] \tag{3.50}$$

$$\boldsymbol{\Sigma} = \mathsf{E}\left[w_2(d)\,(\mathbf{r} - \boldsymbol{\mu})\,(\mathbf{r} - \boldsymbol{\mu})^T\right], \tag{3.51}$$

where $d = (\mathbf{r} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r} - \boldsymbol{\mu})$, and the following relationships hold

$$\hat{\boldsymbol{\mu}}_\infty = \boldsymbol{\mu}^{\mathrm{o}} \tag{3.52}$$

$$\hat{\boldsymbol{\Sigma}}_\infty = c\boldsymbol{\Sigma}^{\mathrm{o}} \tag{3.53}$$

regardless of $w_1(d)$ and $w_2(d)$ as long as they satisfy some technical assumptions such that the solution to (3.48) and (3.49) exists and is unique [140].

Here, the size parameter $c > 0$ is given by

$$\mathsf{E}\left[w_2\left(\frac{\|\mathbf{x}\|_2^2}{c}\right)\frac{\|\mathbf{x}\|_2^2}{c}\right] = N \tag{3.54}$$

with $\mathbf{x} \sim \mathrm{EL}(\mathbf{0}, \mathbf{I}, g)$ sharing the same density generating function as $\mathbf{r}$. For the Gaussian case, since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ implies $\|\mathbf{x}\|_2^2 \sim \chi_N^2$, then the relationship (3.54) can be simplified as

$$\int_0^{+\infty} w_2\left(\frac{x}{c}\right)\frac{x}{c}\chi_N^2(x)dx = N. \tag{3.55}$$

Numerical Algorithm

Algorithm 1 is a numerical iterative method that converges to the unique solution (if it exists and is unique) and the initial values only affect the number of iterations [12].

Algorithm 1 M-Estimator

Input: any $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_0 \succ \mathbf{0}$.

Output: the solution to (3.48) and (3.49).

1: **repeat**
2: $\quad d_{kt} = 1 + (\mathbf{r}_t - \boldsymbol{\mu}_k)\,\boldsymbol{\Sigma}_k^{-1}\,(\mathbf{r}_t - \boldsymbol{\mu}_k)^T$
3: $\quad \boldsymbol{\mu}_{k+1} = \frac{\sum_{t=1}^T w_1(d_{kt})\mathbf{r}_t}{\sum_{t=1}^T w_1(d_{kt})}$
4: $\quad \boldsymbol{\Sigma}_{k+1} = \frac{1}{T}\sum_{t=1}^T w_2(d_{kt})\,(\mathbf{r}_t - \boldsymbol{\mu}_{k+1})\,(\mathbf{r}_t - \boldsymbol{\mu}_{k+1})^T$
5: $\quad k \leftarrow k + 1$
6: **until** convergence

3.4.2 Tyler's Estimator

Tyler's estimator was proposed to find the right balance between efficiency and robustness [201]. It assumes zero mean and focuses on estimating the scatter matrix only. Tyler's estimate is defined as the

solution to the-fixed point equation

$$\mathbf{\Sigma} = \frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\mathbf{\Sigma}^{-1}\mathbf{r}_t}. \tag{3.56}$$

Note that here $xw_2(x) = K$ is not strictly increasing and the results of the M-estimator do not apply. Tyler established the conditions (e.g., one condition is $T \geq N+1$) for existence and uniqueness (up to a positive scalar) of a solution to the fixed-point equation (3.56), and proposed the following iterative Algorithm 2 to achieve the unique trace normalized solution.

Algorithm 2 Tyler's Estimator

Input: $\mathbf{\Sigma}_0 \succ \mathbf{0}$

Output: solution to (3.56)

1: **repeat**
2: $\quad\tilde{\mathbf{\Sigma}}_{k+1} = \frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\mathbf{\Sigma}_k^{-1}\mathbf{r}_t}$
3: $\quad\mathbf{\Sigma}_{k+1} = \frac{\tilde{\mathbf{\Sigma}}_{k+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{k+1})}$
4: $\quad k \leftarrow k+1$
5: **until** convergence

We have previously seen that M-estimators can be regarded as generalized MLEs. Interestingly, Tyler's estimator can be derived from an MLE perspective as well.

It is known that if $\mathbf{r} \sim \mathrm{El}(\mathbf{0}, \mathbf{\Sigma}, g)$, then the normalized samples $\mathbf{s} = \frac{\mathbf{r}}{\|\mathbf{r}\|_2}$ follow [202, 113, 81]

$$f(\mathbf{s}) = \frac{\Gamma\left(\frac{N}{2}\right)}{2\pi^{N/2}}\frac{1}{\sqrt{|\mathbf{\Sigma}|}}\left(\mathbf{s}^T\mathbf{\Sigma}^{-1}\mathbf{s}\right)^{-N/2}, \tag{3.57}$$

which is independent of the density generating function g. Then the MLE of $\mathbf{\Sigma}$ can be obtained by minimizing the scale-invariant negative log-likelihood function

$$L(\mathbf{\Sigma}) = \frac{T}{2}\log|\mathbf{\Sigma}| + \sum_{t=1}^{T}\frac{N}{2}\log\left(\mathbf{s}_t^T\mathbf{\Sigma}^{-1}\mathbf{s}_t\right) \tag{3.58}$$

or, equivalently,

$$L^{\text{Tyler}}(\boldsymbol{\Sigma}) = \frac{T}{2}\log|\boldsymbol{\Sigma}| + \sum_{t=1}^{T}\frac{N}{2}\log\left(\mathbf{r}_t^T\boldsymbol{\Sigma}^{-1}\mathbf{r}_t\right). \tag{3.59}$$

Finally, setting the derivative of $L^{\text{Tyler}}(\boldsymbol{\Sigma})$ w.r.t. to $\boldsymbol{\Sigma}^{-1}$ to zero yields the fixed-point equation (3.56).

3.5 Small Sample Regime & Heavy Tail Issue: Regularized Robust Estimators

One regularity condition for the previous mentioned robust estimators is that the number of samples is at least $T \geq N + 1$. In practice, the universe of stocks may be large and the number available samples for the fitting may be scarce in comparison (e.g., $N = 500$ stocks of the S&P500 and less than two years of daily data, say, $T \approx 400$). Thus when $T \geq N + 1$ is violated and the ordinary robust estimators cannot be applied anymore, or even when it is satisfied, regularization still helps if T is not sufficiently large.

In this part, we mainly study the recent advances on robust estimators with regularizations so that the reliable statistical inference can still be conducted even when the data contains extreme events and/or outliers and the number of samples is limited compared to the data dimension.

3.5.1 Regularized Robust Estimation of Scatter Estimator

This subsection contains the most recent advances on the regularized Tyler's estimator.

Diagonally Loaded Estimator

Similar to the idea of shrinkage covariance or diagonal loading, the authors of [2, 39] proposed to shrink the Tyler update covariance matrix (i.e., step 2 of Algorithm 2) to the identity matrix.

Algorithm 3 summarizes the iterative computing procedure where $\alpha \geq 0$ is a scalar parameter.

Algorithm 3 Tyler's Estimator with shrinkage

Input: $\mathbf{\Sigma}_0 \succ \mathbf{0}$

Output: A unique positive definite matrix

1: **repeat**

2: $\quad \tilde{\mathbf{\Sigma}}_{k+1} = \frac{1}{1+\alpha}\frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\mathbf{\Sigma}_k^{-1}\mathbf{r}_t} + \frac{\alpha}{1+\alpha}\mathbf{I}$

3: $\quad \mathbf{\Sigma}_{k+1} = \frac{\tilde{\mathbf{\Sigma}}_{k+1}}{\mathrm{Tr}(\tilde{\mathbf{\Sigma}}_{k+1})}$

4: $\quad k \leftarrow k+1$

5: **until** convergence

Chen et al. [39] proved that for any $\alpha > 0$ Algorithm 3 converges to a unique point and they proposed a systematic way to select α.

Even though this estimator is widely used and performs well in practice, it is still considered to be heuristic and does not have an interpretation based on minimizing a cost function.

Kullback-Leibler Divergence Regularized Estimator

Interestingly, the heuristic regularization in Algorithm 3 can be formally interpreted as the solution to a Kullback-Leibler (KL) regularized Tyler's loss function (3.59) [190].

For two multivariate Gaussian distributions, e.g., $\mathcal{N}_\Sigma(\mathbf{0}, \mathbf{\Sigma})$ and $\mathcal{N}_T(\mathbf{0}, \mathbf{T})$, the KL divergence is defined as [44]

$$D_{KL}(\mathcal{N}_T||\mathcal{N}_\Sigma) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) - K - \log\left(\frac{|\mathbf{T}|}{|\mathbf{\Sigma}|}\right)\right), \tag{3.60}$$

where the positive definite matrix $\mathbf{T}$ can be interpreted as the target that represents some prior information.

Recall $L^{\mathrm{Tyler}}(\mathbf{\Sigma})$ in (3.59), then ignoring the constant terms results in the following KL divergence regularized $L^{\mathrm{Tyler}}(\mathbf{\Sigma})$:

$$\begin{aligned} L^{\mathrm{KL}}(\mathbf{\Sigma}) = &\log|\mathbf{\Sigma}| + \frac{N}{T}\sum_{t=1}^{T}\log(\mathbf{r}_t^T\mathbf{\Sigma}\mathbf{r}_t) \\ &+ \alpha\left(\mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) + \log|\mathbf{\Sigma}|\right), \end{aligned} \tag{3.61}$$

where $\alpha \geq 0$ is the regularize parameter.

Minimizing (3.61) leads to the following fixed-point equation:

$$\boldsymbol{\Sigma} = \frac{1}{1+\alpha}\frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\boldsymbol{\Sigma}^{-1}\mathbf{r}_t} + \frac{\alpha}{1+\alpha}\mathbf{T}. \tag{3.62}$$

Note that $\mathbf{T} = \mathbf{I}$ recovers the regularization in Algorithm 3.

Interestingly, almost at the same time, three independent works, i.e., [190], [158], and [157], achieved the same result as follows.

Theorem 3.2. Suppose $\mathbf{r}_t$ are drawn i.i.d. from a zero mean elliptical distribution, then the fixed-point equation (3.62) admits a unique solution if and only if $T > \frac{N}{1+\alpha}$. ■

The following Algorithm 4 computes the unique solution.

Algorithm 4 Tyler's Estimator with KL divergence penalty

Input: $\boldsymbol{\Sigma}_0 \succ \mathbf{0}$
Output: the unique solution to (3.62)

1: **repeat**
2: $\quad\boldsymbol{\Sigma}_{k+1} = \frac{1}{1+\alpha}\frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\boldsymbol{\Sigma}_k^{-1}\mathbf{r}_t} + \frac{\alpha}{1+\alpha}\mathbf{T}$
3: $\quad k \leftarrow k+1$
4: **until** convergence

Wiesel's Penalty. There also exist some other regularizations. One example is the Wiesel's penalty [207]:

$$h(\boldsymbol{\Sigma}) = K\log(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})) + \log|\boldsymbol{\Sigma}|, \tag{3.63}$$

and minimizing the Wiesel's penalty regularized Tyler's loss function results in solving the following fixed-point equation:

$$\boldsymbol{\Sigma} = \frac{1}{1+\alpha}\frac{N}{T}\sum_{t=1}^{T}\frac{\mathbf{r}_t\mathbf{r}_t^T}{\mathbf{r}_t^T\boldsymbol{\Sigma}^{-1}\mathbf{r}_t} + \frac{\alpha}{1+\alpha}\frac{N\mathbf{T}}{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})}, \tag{3.64}$$

where $\alpha > 0$ is the regularization parameter.

Example 3.6. Now we set $N = 39$ and draw i.i.d. samples from a Student-t distribution $t_\nu(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}})$ with $\nu = 3$, $\boldsymbol{\mu}^{\mathrm{o}} = \mathbf{0}$, and $\boldsymbol{\Sigma}_{ij}^{\mathrm{o}} =$

$0.8^{|i-j|}$. We assume the mean is known and focus on estimating the normalized scatter matrix only. Note that now the distribution is heavy-tailed. The number of samples $T = 20, 30, \ldots, 100$. The performance metric is the normalized MSE (NMSE) [207]:

$$\text{NMSE} = \frac{\mathsf{E}\left[\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^{\text{o}}\right\|_F^2\right]}{\|\boldsymbol{\Sigma}^{\text{o}}\|_F^2}, \tag{3.65}$$

where all matrices are normalized by their traces.

We simulate the following five estimators: i) the sample covariance matrix, ii) the LW covariance estimator, iii) the Tyler's estimator (i.e., Algorithm 2), and iv) two KL divergence regularized Tyler's estimators (i.e., Algorithm 4) with noninformative identity target $\mathbf{T} = \mathbf{I}$ and informative target $\mathbf{T}$ where $\mathbf{T}_{ij} = 0.7^{|i-j|}$. For tuning the parameter α of the KL regularized Tyler's estimators, a standard cross-validation method is in [207] and a method based on random matrix theory is in [43]. For simulation simplicity, we simulate α such that $\rho(\alpha) = \frac{\alpha}{1+\alpha}$ is each of the ten uniform grid points of the interval $(\max(0, 1 - T/N) + 0.01, 1)$ and we report the best result. Nevertheless, this experiment aims at providing an illustrative example to reveal the ideas behind different estimators and for more intensive numerical experiments, please refer to [39, 207, 190, 157, 158].

Figure 3.9 shows the numerical results where the NMSEs are averaged over 200 realizations. We have several interesting observations: i) the sample covariance matrix and the LW estimator (recall that the LW estimator also relies on the sample covariance matrix) both perform badly since the underlying distribution is heavy-tailed and the extreme events distort the estimation, ii) Tyler's estimator is robust since it uses the weights $\frac{N}{\mathbf{r}_t^T \boldsymbol{\Sigma}^{-1} \mathbf{r}_t}$ to eliminate the effect of the extreme events, however it only works when the number of samples is larger than the data dimension since $\boldsymbol{\Sigma}$ needs to be invertible, iv) the regularized Tyler's estimator with noninformative identity target improves the estimation quality, and v) the informative prior target furthermore improves the estimation quality. All these observations coincide with the ideas behind the different estimators. ■

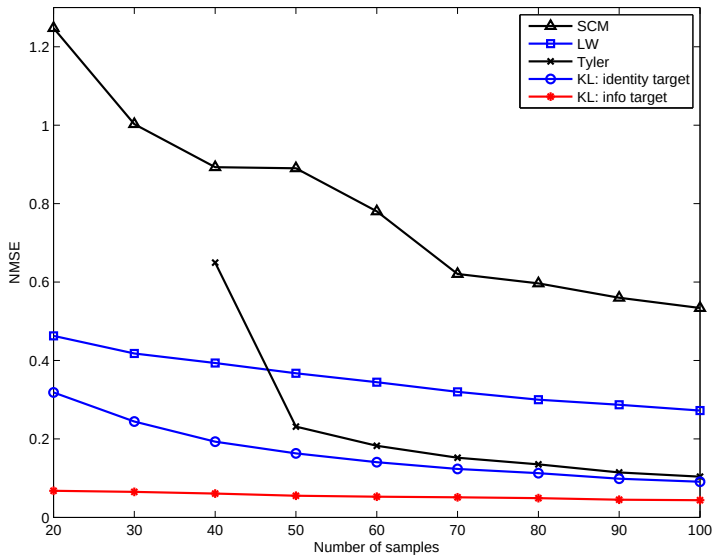

Figure 3.9: Regularized robust covariance estimations.

3.5.2 Regularized Robust Estimation of Mean and Covariance

Previously we reviewed the robust estimation of covariance matrix only assuming the mean was known. Now we consider the joint estimation of the mean and covariance matrix.

Suppose the samples $\mathbf{r}_t$, $t = 1, \ldots, T$ are drawn i.i.d. from an elliptical distribution $\mathrm{El}(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}}, g)$ where the density generating function g is assumed unknown. Since the specific elliptical distribution is assumed unknown, we will do the fitting under a conservative heavy-tail distribution (note that the estimation will work for any elliptical distribution). In particular, it is convenient to use the Cauchy distribution since it has very heavy tails and still the scatter matrix exists (the covariance matrix does not exist).

Consider the Cauchy MLE, that is, the estimates of the mean and the scatter are the minimizers of the following negative log-likelihood:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{T}{2} \log |\boldsymbol{\Sigma}| + \frac{N+1}{2} \sum_{t=1}^{T} \log \left(1 + (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r}_t - \boldsymbol{\mu})\right), \tag{3.66}$$

or, equivalently, the solutions of the following fixed-point equations:

$$\mathbf{0} = \frac{N+1}{T} \sum_{t=1}^{T} \frac{\mathbf{r}_t - \boldsymbol{\mu}}{1 + (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r}_t - \boldsymbol{\mu})}, \tag{3.67}$$

$$\boldsymbol{\Sigma} = \frac{N+1}{T} \sum_{t=1}^{T} \frac{(\mathbf{r}_t - \boldsymbol{\mu})(\mathbf{r}_t - \boldsymbol{\mu})^T}{1 + (\mathrm{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathrm{r}_t - \boldsymbol{\mu})}. \tag{3.68}$$

Asymptotics

Recall the M-estimators in Section 3.4.1. Note that (3.67) and (3.68) are the same as (3.15) and (3.16) with weights given by $w^{\mathrm{C}}(x) = \frac{N+1}{1+x}$ in (3.23), and they are a special case of the M-estimators (3.48) and (3.49) with $w_1(x) = w_2(x) = w^{\mathrm{C}}(x)$.

This implies the asymptotics of the M-estimators applies. That is, under some technical conditions, as $T \to \infty$ the asymptotic solution of (3.67) and (3.68) converges to a unique point, denoted as $(\hat{\boldsymbol{\mu}}_\infty, \hat{\boldsymbol{\Sigma}}_\infty)$. Similar to (3.52) and (3.53), we have $(\hat{\boldsymbol{\mu}}_\infty, \hat{\boldsymbol{\Sigma}}_\infty) = (\boldsymbol{\mu}^{\mathrm{o}}, c\boldsymbol{\Sigma}^{\mathrm{o}})$ where the size parameter c is unknown since now the density generating function g is unknown.

In other words, asymptotically the Cauchy MLE can estimate the mean and the shape well. Therefore, it is more sensible to regularize only the shape but not the size if a regularization is necessary.

Small Sample Regime

When the number of samples is limited, say less than the data dimension, the regularized Cauchy MLE is more reliable. Sun et al. [191] proposed the following penalty function:

$$h(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \alpha \left(N \log(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})) + \log|\boldsymbol{\Sigma}| \right) + \gamma \log\left(1 + (\mathbf{t}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right), \tag{3.69}$$

where $\alpha \geq 0$ and $\gamma \geq 0$ are regularization parameters. It is easy to verify that the minimizer of (3.69) is $(\mathbf{t}, c\mathbf{T})$ for any $c > 0$ [191]. That is, (3.69) shrinks the mean to $\mathbf{t}$ and scatter to the shape of $\mathbf{T}$ only. This justifies that $h(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a proper penalty function since it only penalizes the shape but not the size of the scatter matrix.

Then the regularized Cauchy MLE problem is

$$\begin{aligned} \underset{\boldsymbol{\mu},\, \boldsymbol{\Sigma} \succ \mathbf{0}}{\text{minimize}} \quad & \frac{N+1}{2} \sum_{t=1}^{T} \log\left(1 + (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{r}_t - \boldsymbol{\mu})\right) \\ & + \frac{T}{2} \log|\boldsymbol{\Sigma}| + \alpha \left(N \log(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})) + \log|\boldsymbol{\Sigma}| \right) \\ & + \gamma \log\left(1 + (\mathbf{t}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right). \end{aligned} \tag{3.70}$$

Setting the derivatives of the objective w.r.t $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ to zeros yields the following fixed-point equations:

$$\boldsymbol{\mu} = \frac{(N+1)\sum_{t=1}^{T} w_t(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbf{r}_t + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbf{t}}{(N+1)\sum_{t=1}^{T} w_t(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \tag{3.71}$$

$$\begin{aligned} \boldsymbol{\Sigma} = & \frac{N+1}{T+2\alpha} \sum_{t=1}^{T} w_t(\boldsymbol{\mu}, \boldsymbol{\Sigma})(\mathbf{r}_t - \boldsymbol{\mu})(\mathbf{r}_t - \boldsymbol{\mu})^T \\ & + \frac{2\gamma}{T+2\alpha} w_{\mathbf{t}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})(\mathbf{t} - \boldsymbol{\mu})(\mathbf{t} - \boldsymbol{\mu})^T \\ & + \frac{2\alpha N}{T+2\alpha} \frac{\mathbf{T}}{\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})}, \end{aligned} \tag{3.72}$$

where

$$w_t(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{1 + (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{r}_t - \boldsymbol{\mu})}, \tag{3.73}$$

$$w_{\mathbf{t}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{1 + (\mathbf{t} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{t} - \boldsymbol{\mu})}. \tag{3.74}$$

The properties of the problem (3.70) are as follows.

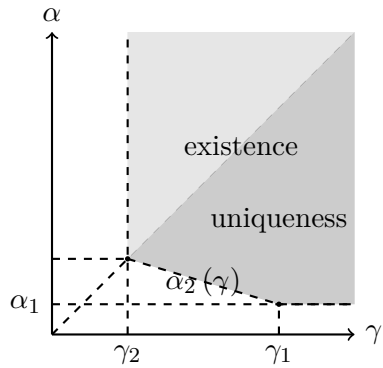

Figure 3.10: Values that the regularization parameters α and γ can take for the existence and uniqueness of the regularized Cauchy MLE.

Theorem 3.3. Assume the underlying distribution of the samples is continuous, $\mathbf{T} \succ \mathbf{0}$, $T > 1$, $\alpha \geq 0$ and $\gamma \geq 0$, then we have
Existence: Problem (3.70) has a minimizer if either of the following conditions are satisfied:
(i) $\gamma > \gamma_1$ and $\alpha > \alpha_1$
(ii) $\gamma_2 < \gamma \leq \gamma_1$ and $\alpha > \alpha_2(\gamma)$
where $\gamma_1 = N/2$, $\gamma_2 = (N + 1 - T)/2$, $\alpha_1 = (N - T)/2$, and

$$\alpha_2(\gamma) = \frac{1}{2}\left(N + 1 - T - \frac{2\gamma + T - N - 1}{T - 1}\right).$$

Uniqueness: The solution is unique if $\gamma \geq \alpha$. ■

Proof. See [191, Theorem 2, Corollary 3, and Theorem 4]. □

Figure 3.10 shows the regions of regularization parameter values for the existences and uniqueness of the regularized Cauchy MLE (3.70).

To compute the unique solution, Sun et al. [191] also proposed several iterative algorithms, including the following Algorithm 5, with convergence guaranteed based on the majorization-minimization (MM) theory.

Algorithm 5 Iterative Regularized Cauchy MLE

Input: $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0 \succ \mathbf{0}$

Output: The unique solution of (3.70)

1: **repeat**

2: $$\boldsymbol{\mu}_{k+1} = \frac{(N+1)\sum_{t=1}^{T} w_t(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\mathbf{r}_t + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\mathbf{t}}{(N+1)\sum_{t=1}^{T} w_t(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + 2\gamma w_{\mathbf{t}}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

3: $$\begin{aligned}\boldsymbol{\Sigma} =& \frac{N+1}{T+2\alpha}\sum_{t=1}^{T} w_t(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)(\mathbf{r}_t - \boldsymbol{\mu}_{k+1})(\mathbf{r}_t - \boldsymbol{\mu}_{k+1})^T \\ &+ \frac{2\gamma}{T+2\alpha} w_{\mathbf{t}}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)(\mathbf{t} - \boldsymbol{\mu}_{k+1})(\mathbf{t} - \boldsymbol{\mu}_{k+1})^T \\ &+ \frac{2\alpha N}{T+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\mathbf{T})}\end{aligned}$$

4: $k \leftarrow k+1$

5: **until** convergence

Example 3.7. In this example, we study the robustness of different outliers. We fix $N = 100$ and draw i.i.d. samples from $\mathcal{N}(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}})$ with $\boldsymbol{\mu}^{\mathrm{o}} = \mathbf{1}$, and $\boldsymbol{\Sigma}^{\mathrm{o}}_{ij} = 0.8^{|i-j|}$. Then we draw outliers as $\mathbf{r}_{\mathrm{outlier}} \sim \boldsymbol{\mu} + r\mathbf{s}$ where $\mathbf{s}$ is uniformly distributed on a sphere such that $\|\mathbf{s}\|_2 = 1$ and $r \sim \mathrm{Uniform}[2l, 2l+1]$ where $l \triangleq \max_t\{\|\mathbf{r}_t\|_2\}$. The total number of samples is $T = 120$ and the fraction of outliers varies as $0.02, 0.05, \ldots, 0.2$. Since we are now estimating both the mean and the covariance matrix we need a combined measure of performance. We use the symmetric KL divergence distance [191]:

$$\begin{aligned}\text{KL distance} = \mathsf{E}\big[&D_{KL}(\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})||\mathcal{N}(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}})) \\ &+ D_{KL}(\mathcal{N}(\boldsymbol{\mu}^{\mathrm{o}}, \boldsymbol{\Sigma}^{\mathrm{o}})||\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}))\big],\end{aligned} \tag{3.75}$$

where all matrices are normalized by their traces.

We simulate the following six estimators: i) the sample covariance matrix, ii) the LW covariance estimator, iii) the Cauchy MLE estimator, and iv) three regularized Cauchy MLE estimators (i.e., Algorithm 5) with: iv-a) $\mathbf{t}$ being the sample mean and noninformative identity covariance target $\mathbf{T} = \mathbf{I}$, iv-b) $\mathbf{t}$ being the sample mean and informative covariance target $\mathbf{T}$ where $\mathbf{T}_{ij} = 0.7^{|i-j|}$, and iv-c) informative mean target $\mathbf{t} = 0.9\boldsymbol{\mu}^{\mathrm{o}}$ and informative covariance target with $\mathbf{T}_{ij} = 0.7^{|i-j|}$.

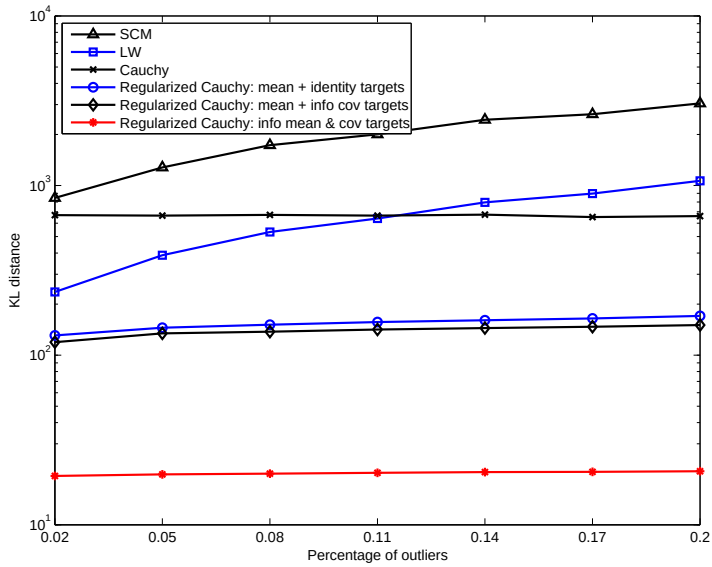

Figure 3.11: Regularized robust mean and covariance estimations.

For the regularized Cauchy MLEs, for simulation simplicity we set $\alpha = \gamma$ and simulate α such that $\rho(\alpha) = \frac{T}{T+2\alpha} \in \{0.1, 0.2, \ldots, 1\}$ and report the best result. A more practical but complicated way is cross-validation [207].

Figure 3.11 shows the numerical results where the KL distances are averaged over 200 realizations. We can see similar observations to that of Example 3.6. Briefly speaking, the regularized Cauchy MLE (even with a noninformative target) does improve the estimation quality in a small sample regime and the improvement becomes more significant when the percentage of outliers increases. For more intensive numerical experiments, the interested reader can refer to [191]. ■

3.6 Summary of Different Estimators

In this chapter we have reviewed different types of estimators: non-parametric estimators (e.g., sample mean/covariance, LS estimator), ML estimators, and shrinkage/regularized estimators mainly based on the I.I.D. model. Table 3.1 provides a brief and compact summary.

Table 3.1: Summary of different estimators.

Type	Name	Fixed-Point Eqs. or Expression or Problem	Scenario
Non-parametric	Sample averages	(3.2) and (3.3)	Large sample; Same as Gaussian MLE
	Least square		
MLE	Elliptical	(3.15) and (3.16) with weight (3.17)	Large or medium sample
	Gaussian	(3.15) and (3.16) with weight (3.19)	
	Cauchy	(3.15) and (3.16) with weight (3.23)	
Regularized or Shrinkage	Mean	(3.30)-(3.31)	Small sample without extreme events or outliers
	Covariance	(3.33), (3.35)-(3.36)	
	Precision	Problem (3.41)	
Robust	M-estimator	(3.48)-(3.49)	Generalized MLE for large or medium sample with extreme events or outliers
	Tyler	(3.56)	
Regularized Robust	KL regularized Tyler	Minimizing (3.61) or solving (3.62)	A combination of shrinkage idea and robust estimators; Small sample with extreme events and/or outliers
	Wiesel regularized Tyler	Solving (3.64)	
	Regularized Cauchy	Minimizing (3.70) or solving (3.71)-(3.72)	

4

Order Execution

Order execution bridges a desired ideal target and the real world: once a portfolio has been designed, it needs to be executed in the real markets. This chapter studies the order execution problem and how to optimally execute such orders.

Section 4.1 briefly reviews the limit order book system and introduces the concept of market impact. Section 4.2 further presents the price model and execution cost. Section 4.3 focuses on minimizing the expected execution cost, and Section 4.4 considers an extension of minimizing the mean-variance trade-off of execution cost. Finally, Section 4.5 considers minimizing a more practical criterion, i.e., the Conditional Value-at-Risk (CVaR), of the execution cost.

4.1 Limit Order Book and Market Impact

4.1.1 Limit Order Book

Once a buy (respectively, sell) order has been submitted, it will not be executed immediately. Instead, it will be checked for whether it can be matched by the previously submitted sell (respectively, buy) orders. A limit order book at a specific time is the snapshot of all the active outstanding orders at that time [87].

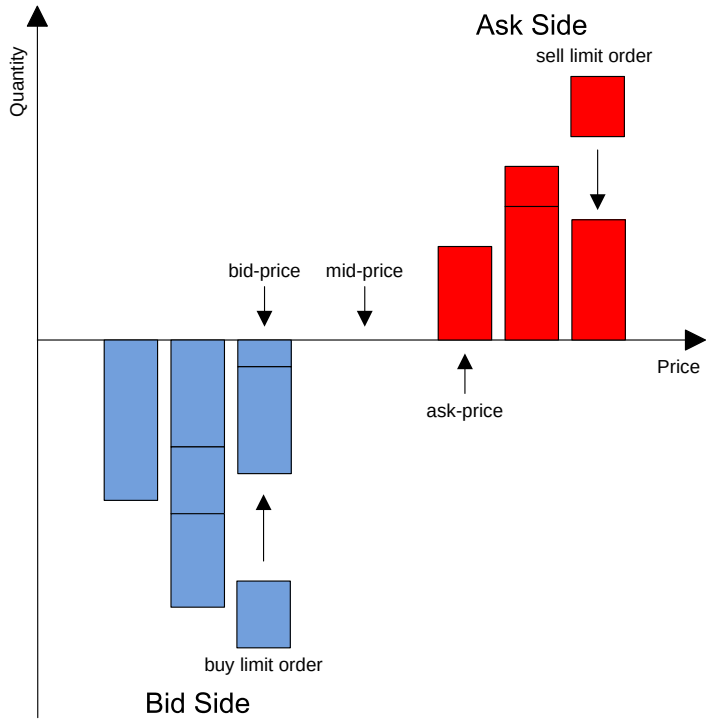

Figure 4.1: Limit order book and two limit orders.

Orders that do not cause an immediate matching upon submission but become active orders in a limit order book are known as limit orders. Figure 4.1 shows an illustrative example of a limit order book and two new limit orders. The limit orders that ask for sell are called sell limit orders. They may have different ask prices and the lowest ask price is referred to as ask-price. Respectively, the limit orders that bid for buy are called buy limit orders. They have different bid prices and the highest bid price is referred to as bid-price. The average of the bid-price and the ask-price is referred to as mid-price and the difference between the ask-price and the bid-price is called a bid-ask spread.

Orders that cross the bid-ask spread cause an immediate matching upon submission. This can happen, for example, when a small buy order is submitted with a bid price equal to the current ask-price (and

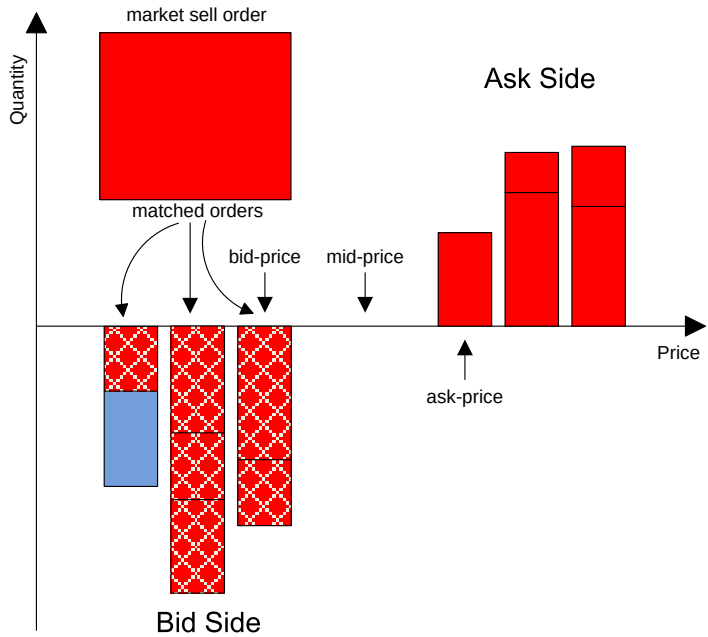

Figure 4.2: A new market sell order and matched trades.

the amount of that order can be absorbed). Also, there is a type of execution order, called a market order, that does not have an associated price, instead it will be matched to the best existing price in the order book and executed immediately. However, the execution of large orders follows a different pattern.

Figure 4.2 shows a submission of a large market sell order that its owner simply wants to sell at whatever price the limit order book can provide immediately. Once the large market sell order has been submitted, it matches some limit buy orders in the limit order book (in order from high to low price). Figure 4.2 also shows the matched trades with different quantities at different prices.

Right after the trades have been executed, they are eliminated from the limit order book. Figure 4.3 shows the updated limit order book. Clearly, we can see that the overall average trade price is lower than the initial bid-price and the new bid-price also becomes much lower.

Figure 4.3: Market impact: the large market sell order moves prices in the opposite direction.

That is, a large market sell order moves the prices in the opposite direction, i.e., when one wants to sell, he/she actually sells lower. Similarly, when one wants to buy, he/she actually buys higher. Such an effect is known as market impact and will be explained next.

4.1.2 Market Impact

In the practice of quantitative investment, portfolio allocation decisions and trading strategies are realized through the execution of buy and sell orders in organized exchanges via brokers through the limit order book systems. Due to practical limitations concerning market liquidity, i.e., availability of required volume levels matching the size of an outstanding order for a specific asset, executing transactions in the market has an effect on the prices of assets: buying pushes the prices upward and selling pushes the prices downward, as shown in the previous illustrative Figures 4.2 and 4.3. This market impact is reflected on the cost incurred when implementing trades [94, 156].

Figure 4.4[1] illustratively shows how the price evolves versus time when a large market sell order, say s shares, is executed directly. This market order must be matched immediately and represents a demand for liquidity. Similar to Figure 4.2, selling this large amount of s

[1]Figures 4.4 and 4.5 are reproduced based on [65, Figures 12.1 and 12.2].

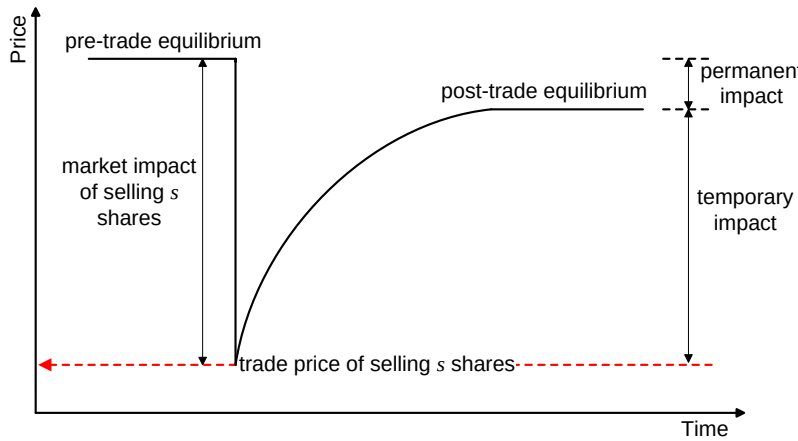

Figure 4.4: Market impact of a single large order.

shares incurs significant market impact and the executed price is much lower than the pre-trade equilibrium price. As time goes by, liquidity providers replenish the bid side and the limit order book reaches a post-trade equilibrium; however, the price is still lower than the pre-trade equilibrium. The difference between the pre-trade and post-trade equilibrium is due to the information that an investor has decided to sell s shares and it is referred to as permanent impact. The remaining impact is called temporary and it is because the investor wants to sell the order immediately regardless of price. In practice, the temporary market impact is more significant than the permanent one [115]. Another observation is that the permanent impact propagates with time while the temporary impact diminishes after some time period.

Small orders in general have much smaller market impacts. Intuitively, a large order can be partitioned into many small orders to be executed sequentially to reduce the overall market impact. Figure 4.5 shows the example of partitioning a large order of s shares into two equal small orders executed sequentially. We can observe several things: 1) the market impact of selling $s/2$ shares is much smaller than that of selling s directly and 2) executing them sequentially helps to reduce the overall market impact since the price may be recovered from the

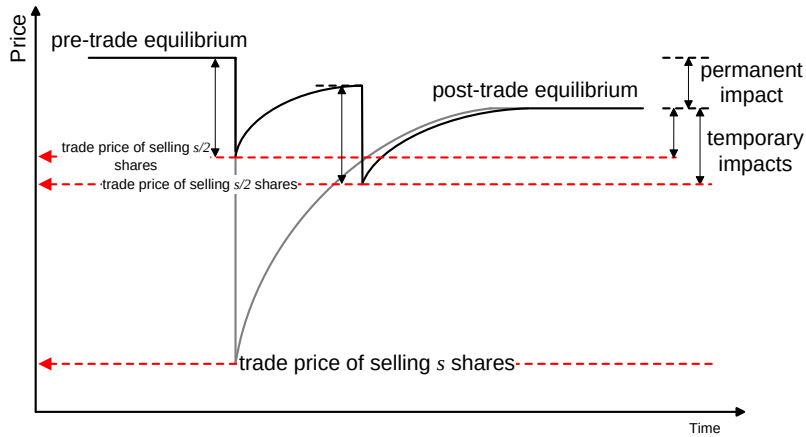

Figure 4.5: Market impact of two sequential small orders.

temporary impact caused by the first trade before the second trade happens. Obviously, the average of the two trade prices of selling $s/2$ is much higher than that of selling s shares at once directly. That is, overall, the average trade price of the total s shares achieved by executing small orders sequentially is much higher than that of executing the large order once.

Naturally, the idea of optimal order execution is to partition a large order into many small pieces and execute them sequentially. The minimization of the execution cost through optimal order execution algorithms is crucial for preserving in practice the profit structure of theoretically sound investment processes [65]. Otherwise, one may expect to make a certain profit with a carefully designed portfolio that will vanish or even become negative. Interestingly, this order execution problem is close to many other scheduling and optimization problems in signal processing. From a dynamic control point of view, the order execution problem of finding an optimal order execution strategy to minimize the mean-variance trade-off of the execution cost [18] is quite similar to the problem of finding an optimal sensor scheduling strategy to minimize the state estimation error in dynamic wireless sensor networks [208, 180, 181]. From an optimization point of view, distributing

a large order into smaller size orders over a certain time window to minimize the execution cost [8, 79] is similar to allocating total power over different channels to achieve the capacity region for parallel Gaussian broadcast channels [198], or to minimize the J-divergence between the distributions of the detection statistic in wireless sensor networks [214].

Fact 4.1. One usually focuses on minimizing the market impact and therefore reducing execution cost. However, every coin has two sides and in practice it is possible to use the market impact to make money as well [83]. Perhaps the most famous example is "Black Wednesday". In 1992, there was a devaluation trend of pound sterling and Geogre Soros' Quantum fund began to massively short-sell pounds on Tuesday, September 15, 1992 and triggered a more intensive trend of devaluation of the pound. On Wednesday, September 16, 1992, the Bank of England was not able to protect the pound anymore and the British Conservative government was forced to withdraw the pound sterling from the European Exchange Rate Mechanism. During that period, Soros first held a total of US$10 billion short positions on GBP and later closed the position at a lower value so that he made US$1 billion. Because of that, he has since been known as "The Man Who Broke the Bank of England". ■

4.2 Price Model and Execution Cost

4.2.1 Price Model

Before introducing a price model, let us define the notation. The buy and the sell problems are similar to each other and for the sake of notation we focus on the sell problem. Assume we hold N stocks with an initial price $\mathbf{p}_0 \triangleq [p_{10}, \ldots, , p_{N0}]^T$ with the number of shares to sell denoted by $\mathbf{s} \triangleq [s_1, \ldots, s_N]^T$ and we want to completely execute them before time T. Assume there is no short-selling, and denote the number of shares for the N stocks executed over the t-th period as $\mathbf{n}_t \triangleq [n_{1t}, \ldots, n_{Nt}]^T \geq \mathbf{0}$, $t = 1, \ldots, T$. We write the order execution sequence $\{\mathbf{n}_1, \ldots, \mathbf{n}_T\}$ as an N-by-T matrix $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_T]$, such that $\mathbf{N1} = \mathbf{s}$, where $\mathbf{1}$ is a T-dimensional vector having all entries equal

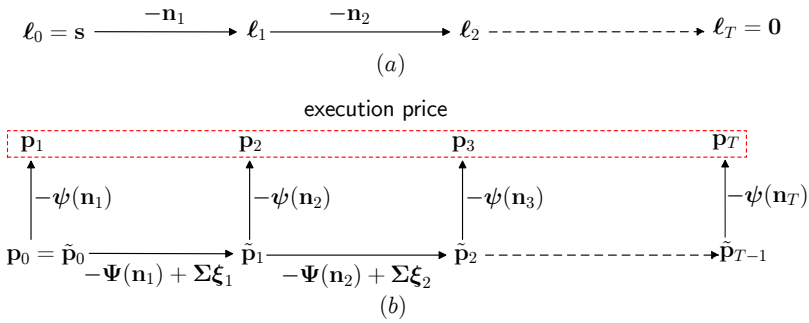

Figure 4.6: Illustrations of (a) trading trajectory and (b) price model.

to 1. We define such a matrix $\mathbf{N}$ as an execution strategy. The number of remaining shares after the t-th period is $\boldsymbol{\ell}_t$, with $\boldsymbol{\ell}_t = \boldsymbol{\ell}_{t-1} - \mathbf{n}_t$, initial condition $\boldsymbol{\ell}_0 = \mathbf{s}$, and end condition $\boldsymbol{\ell}_T = 0$. Then, $\mathbf{L} = [\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T]$ is a trading trajectory, and it evolves as shown in Figure 4.6a.

We then consider a model for the price dynamics taking into account the market impact incurred when executing the order. Specifically, the execution price in the t-th period $\mathbf{p}_t(\mathbf{n}_1, \ldots, \mathbf{n}_t)$ is a random variable depending on past executions and also the current execution. In order to characterize how it evolves over time, a number of different price models have been proposed [102, 8, 18]. Here, we consider the quite general price model in [8] with both linear permanent and temporary market impact components. More specifically, prices evolve, for $t = 1, \ldots, T$, as

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1} - \boldsymbol{\Psi}(\mathbf{n}_t) + \boldsymbol{\Sigma}\boldsymbol{\xi}_t, \tag{4.1}$$

$$\mathbf{p}_t = \tilde{\mathbf{p}}_{t-1} - \boldsymbol{\psi}(\mathbf{n}_t), \tag{4.2}$$

where $\tilde{\mathbf{p}}_t \triangleq [\tilde{p}_{1t}, \ldots, \tilde{p}_{Nt}]^T$ is a hidden variable denoting the permanent impact prices with initial value $\tilde{\mathbf{p}}_0 = \mathbf{p}_0$, $\mathbf{p}_t \triangleq [p_{1t}, \ldots, p_{Nt}]^T$ is the actual execution price, $\boldsymbol{\xi}_t \triangleq [\xi_{1t}, \ldots, \xi_{rt}]^T$ is the random noise with all the elements being i.i.d. random variables with zero mean and unit variance, $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times r}$ is the volatility matrix, and $\boldsymbol{\Psi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ are permanent and temporary, respectively, linear market impact functions that take the form

$$\boldsymbol{\Psi}(\mathbf{n}_t) = \boldsymbol{\Theta}\mathbf{n}_t, \tag{4.3}$$

$$\boldsymbol{\psi}(\mathbf{n}_t) = \boldsymbol{\Omega}\mathbf{n}_t. \tag{4.4}$$

In the above price model, the parameters $\boldsymbol{\Sigma}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Omega}$ represent the linear coefficient matrices of noise, permanent market impact, and temporary market impact. They are usually fixed and calibrated in advance by using data on the bid-ask spread, the volatility, and the daily trading volume. Similar to [8], we assume that $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ is positive definite[2], and for simplicity, we assume that the matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$ are both symmetric. Figure 4.6(b) summarizes price model.

4.2.2 Execution Cost

Let $\mathbf{P} \triangleq [\mathbf{p}_1, \ldots, \mathbf{p}_T]$ and $\boldsymbol{\Xi} \triangleq [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T]$. The ideal value in the absence of market impact and market noise would be $\mathbf{p}_0^T\mathbf{s}$ but in practice it becomes $\sum_{t=1}^{T} \mathbf{n}_t^T\mathbf{p}_t = \mathrm{Tr}\left(\mathbf{P}^T\mathbf{N}\right)$, and the gap between them is defined as the execution cost (i.e., the implementation shortfall) as follows [159]:

$$X(\mathbf{N}) = \begin{cases} \mathbf{p}_0^T\mathbf{s} - \mathrm{Tr}\left(\mathbf{P}^T\mathbf{N}\right), & \text{sell program} \\ \mathrm{Tr}\left(\mathbf{P}^T\mathbf{N}\right) - \mathbf{p}_0^T\mathbf{s}, & \text{buy program.} \end{cases} \tag{4.5}$$

Based on the above price model for a sell program, plugging (4.1)-(4.4) into (4.5) and after some mathematical manipulations, we obtain

$$X(\mathbf{N}) = \frac{1}{2}\mathbf{s}^T\boldsymbol{\Theta}\mathbf{s} + \mathrm{Tr}\left(\mathbf{N}^T\tilde{\boldsymbol{\Omega}}\mathbf{N}\right) - \mathrm{Tr}\left(\mathbf{L}^T\boldsymbol{\Sigma}\boldsymbol{\Xi}\right), \tag{4.6}$$

where $\tilde{\boldsymbol{\Omega}} \triangleq \boldsymbol{\Omega} - \frac{1}{2}\boldsymbol{\Theta}$, and the mean and the variance are

$$\mathsf{E}\left[X(\mathbf{N})\right] = \frac{1}{2}\mathbf{s}^T\boldsymbol{\Theta}\mathbf{s} + \mathrm{Tr}\left(\mathbf{N}^T\tilde{\boldsymbol{\Omega}}\mathbf{N}\right) \tag{4.7}$$

$$\mathsf{Var}\left[X(\mathbf{N})\right] = \mathrm{Tr}\left(\mathbf{L}^T\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\mathbf{L}\right). \tag{4.8}$$

Recall that $\mathbf{L}$ is a function of $\mathbf{N}$: $\boldsymbol{\ell}_t = \boldsymbol{\ell}_{t-1} - \mathbf{n}_t$ and $\boldsymbol{\ell}_0 = \mathbf{s}$.

In practice, it is also assumed that $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \frac{1}{2}\boldsymbol{\Theta} \succ \mathbf{0}$, as in [145]. This makes sense because usually the temporary market impact is much higher than the permanent market impact in financial markets. Indeed,

[2]Because if $\mathbf{n}_t^T\boldsymbol{\Omega}\mathbf{n}_t \leq 0$, it would mean the temporary market impact of executing $\mathbf{n}_t$ in fact would benefit the trading or at least would lose nothing, which would go against the goal of reducing the market impact.

the amount caused by the permanent impact is a relatively small percentage of the pure cost component, and Kissell et. al. estimate it to be 5% [115, pp. 182]. Then both the mean (4.7) and the variance (4.8) are quadratic convex in $\mathbf{N}$.

4.3 Minimizing Expected Execution Cost

The first problem formulation was proposed in [18] for the single asset case, and it aims at minimizing the expected execution cost

$$\begin{array}{ll} \underset{\mathbf{N}}{\text{minimize}} & \mathsf{E}\left[X(\mathbf{N})\right] \\ \text{subject to} & \mathbf{N}\mathbf{1} = \mathbf{s},\ \mathbf{N} \geq \mathbf{0}. \end{array} \tag{4.9}$$

Since $\tilde{\mathbf{\Omega}} \succ \mathbf{0}$, the problem is already quadratic convex and thus can be efficiently and numerically solved. When $\tilde{\mathbf{\Omega}}$ is diagonal, it is not hard to show that the problem (4.9) of N assets can be decomposed into N small problems of a single asset, and following the derivation in [18], the optimal execution strategy is to uniformly distribute the large order among the T execution periods, that is, $\mathbf{N} = \frac{1}{T}\mathbf{s}\mathbf{1}^T$.

4.4 Minimizing Mean-Variance Trade-off of Execution Cost

An obvious disadvantage of the problem (4.9) is that it does not consider the risk of the execution cost. By taking the variance as the risk measurement, Almgren and Chriss [8] extended (4.9) by minimizing a mean-variance trade-off of the execution cost as follows:

$$\begin{array}{ll} \underset{\mathbf{N}}{\text{minimize}} & \mathsf{E}\left[X(\mathbf{N})\right] + \lambda \mathsf{Var}\left[X(\mathbf{N})\right] \\ \text{subject to} & \mathbf{N}\mathbf{1} = \mathbf{s},\ \mathbf{N} \geq \mathbf{0}, \end{array} \tag{4.10}$$

where $\lambda \geq 0$ is a fixed parameter modeling an investor's risk aversion level. The larger the value of λ, the more risk averse the investor ($\lambda = 0$ means the investor is risk neutral and it corresponds to problem (4.9)). For obvious reasons, such an approach is commonly referred to in the literature as the mean-variance optimization approach. Note that since $\tilde{\mathbf{\Omega}} \succ \mathbf{0}$, the problem is already convex.

4.5 Minimizing CVaR of Execution Cost

However, it is well known that the variance used in (4.9) is not an appropriate risk measure when dealing with financial returns from non-normal, negatively skewed, and leptokurtic distributions [141]. In order to overcome the inadequacy of variance, CVaR (also known in the literature as Expected Shortfall, Expected Tail Loss, Tail Conditional Expectation, and Tail VaR) has been proposed as a single side alternative risk measurement [166] and it has been employed significantly in financial engineering, see [7, 63, 172, 100], for portfolio or risk management. Interestingly, such a single side risk measurement technique has also found some applications in signal processing recently, see [124, 183], for chance constrained communication systems.

4.5.1 CVaR and Problem Formulation

The CVaR is defined as the conditional mean value of a random variable exceeding a particular percentile. This precisely measures the risky realizations, as opposed to the variance that simply measures how spread the distribution is and mixes together both tails.

For illustrative purposes, Figure 4.7 shows the definition of the CVaR of a random variable. Mathematically, given an random variable Z, the CVaR of the execution cost at the $1-\varepsilon$ confidence level can be expressed as

$$\mathrm{CVaR}_{1-\varepsilon}(Z) = \mathsf{E}\left[Z \middle| Z > \mathrm{VaR}_{1-\varepsilon}(Z)\right], \tag{4.11}$$

where the Value-at-Risk of the execution cost at the $1-\varepsilon$ confidence level, denoted as $\mathrm{VaR}_{1-\varepsilon}(Z)$, is the $(1-\varepsilon)$-quantile of Z:

$$\mathrm{VaR}_{1-\varepsilon}(Z) = \inf_{\zeta\in\mathbb{R}} \{\zeta | P(Z > \zeta) \leq \varepsilon\}. \tag{4.12}$$

Note that given an execution strategy $\mathbf{N}$, the execution cost $X(\mathbf{N})$ is a random variable and the problem of minimizing the CVaR of the execution cost turns out to be [79, 77, 78]:

$$\begin{array}{ll} \underset{\mathbf{N}}{\mathsf{minimize}} & \mathrm{CVaR}_{1-\varepsilon}(X(\mathbf{N})) \\ \mathsf{subject\ to} & \mathbf{N1} = \mathbf{s},\ \mathbf{N} \geq \mathbf{0}. \end{array} \tag{4.13}$$

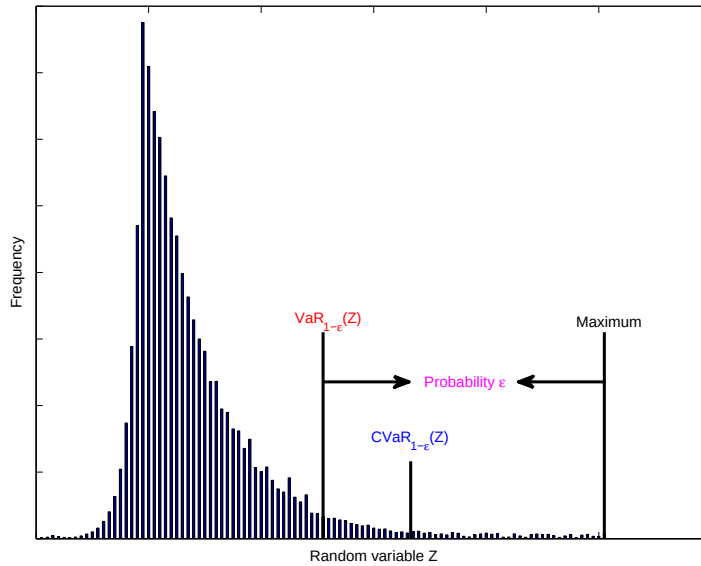

Figure 4.7: The VaR and CVaR of a random variable.

At first glance, $\mathrm{CVaR}_{1-\varepsilon}\left(X\left(\mathbf{N}\right)\right)$ is hard to deal with because it contains a conditional expectation exceeding a threshold that is not fixed. To proceed, we will make use of the following auxiliary function.

Auxiliary Function. Following the approach in [166], we can define an auxiliary function of $\mathrm{CVaR}_{1-\varepsilon}(X(\mathbf{N}))$ as follows:

$$F_{\varepsilon}(\mathbf{N},\zeta) = \zeta + \varepsilon^{-1}\mathsf{E}\left[X(\mathbf{N}) - \zeta\right]^{+} \tag{4.14}$$

where $[x]^{+} = \max\left(x, 0\right)$. Observe that (4.14) is convex w.r.t. both ζ and $\mathbf{N}$, since $X(\mathbf{N})$ is convex quadratic in $\mathbf{N}$, and additionally, we further have [166]:

$$\mathrm{CVaR}_{1-\varepsilon}(X(\mathbf{N})) = \min_{\zeta} F_{\varepsilon}(\mathbf{N},\zeta). \tag{4.15}$$

Then, the original problem (4.13) can be more efficiently optimized by using the property in (4.15). To that effect, notice that we need to compute the expectation $\mathsf{E}\left[X(\mathbf{N}) - \zeta\right]^{+}$.

4.5.2 Sample Average Approximation

The first idea is to use the sample average approximation (SAA) to approximate $\mathsf{E}\left[X(\mathbf{N})-\zeta\right]^{+}$, and the CVaR problem (4.13) is approximated by

$$\begin{array}{ll}\underset{\mathbf{N},\mathbf{z},\zeta}{\text{minimize}} & \zeta+\varepsilon^{-1}M^{-1}\sum_{i=1}^{M}z_i\\ \text{subject to} & 0\leq z_i\geq \frac{1}{2}\mathbf{s}^T\mathbf{\Theta}\mathbf{s}+\text{Tr}\left(\mathbf{N}^T\tilde{\mathbf{\Omega}}\mathbf{N}\right)-\text{Tr}\left(\mathbf{L}^T\mathbf{\Sigma}\mathbf{\Xi}^i\right)-\zeta,\\ & \qquad\forall i=1,\ldots,M\\ & \mathbf{N1}=\mathbf{s},\ \mathbf{N}\geq\mathbf{0},\end{array}\tag{4.16}$$

where $\mathbf{\Xi}^i$ is the i-th realization of noise sampled from the distribution of ξ_{it}'s, and M is the number of noise realizations. As pointed out in [142] although the SAA method can provide an accurate execution strategy for a very large number of realizations, such a method is impaired by large storage requirements and high computational complexity, especially when M is large.

4.5.3 Analytical Approach

To overcome the drawback of the SAA method, an analytical approach to handling $\text{CVaR}_{1-\varepsilon}\left(X\left(\mathbf{N}\right)\right)$ and solving (4.13) for both Gaussian and Non-Gaussian noise was proposed in [79, 77, 78]. The idea is to either find the explicit expression of $\mathsf{E}\left[X(\mathbf{N})-\zeta\right]^{+}$ for the Gaussian cases or construct a save convex approximation of $\mathsf{E}\left[X(\mathbf{N})-\zeta\right]^{+}$ for the general non-Gaussian noise.

Gaussian Noise

For the Gaussian case, the following analytical equivalent formulation of the problem (4.13) was derived in [78].

Lemma 4.1. If all the ξ_{it} are i.i.d. and $\xi_{it}\sim\mathcal{N}\left(0,1\right)$, and $\tilde{\mathbf{\Omega}}\succ\mathbf{0}$ in the price model (4.1)-(4.2), we have that (4.13) is equivalent to the convex problem:

$$\begin{array}{ll}\underset{\mathbf{N}}{\text{minimize}} & \frac{1}{2}\mathbf{s}^T\mathbf{\Theta}\mathbf{s}+\text{Tr}\left(\mathbf{N}^T\tilde{\mathbf{\Omega}}\mathbf{N}\right)+\kappa\left(\varepsilon\right)\left\|\mathbf{\Sigma}^T\mathbf{L}\right\|_F\\ \text{subject to} & \mathbf{N1}=\mathbf{s},\ \mathbf{N}\geq\mathbf{0},\end{array}\tag{4.17}$$

where $\kappa(\varepsilon) = \exp\left(-\left(Q^{-1}(\varepsilon)\right)^2/2\right)/\sqrt{2\pi}\varepsilon$ and $Q^{-1}(x)$ is the inverse Q-function.[3]. ■

Interestingly, the results in Lemma 4.1 can be extended to more general elliptical distributions [78].

General Non-Gaussian Noise

The elliptical distributions are only appropriate in situations where returns are symmetric or not strongly asymmetric, but fail to successfully model highly asymmetric or, equivalently, skewed returns [141]. For such cases, $\mathrm{CVaR}_{1-\varepsilon}(X(\mathbf{N}))$ admits no explicit expression, and an alternative way to approach the CVaR execution problem is to solve a safe tractable convex approximation of $\mathrm{CVaR}_{1-\varepsilon}(X(\mathbf{N}))$ instead. The following technical assumption is needed.

Assumption 4.1. The moment generating function of the random variable ξ_{it}, i.e., $M_{it}(z) = \mathsf{E}\left[e^{z\xi_{it}}\right]$, is finite-valued for all $z \in \mathbb{R}$ and can be computed efficiently. ■

Then one can have the following result [78].

Proposition 4.1 (Bernstein's Approximation)**.** If all the ξ_{it} are i.i.d. satisfying Assumption 4.1, and $\tilde{\mathbf{\Omega}} \succ \mathbf{0}$ in the price model (4.1)-(4.2), a safe tractable convex approximation of (4.13) is

$$\begin{array}{ll} \underset{\mathbf{N},\, z>0}{\mathsf{minimize}} & \frac{1}{2}\mathbf{s}^T\mathbf{\Theta}\mathbf{s} + \mathrm{Tr}\left(\mathbf{N}^T\tilde{\mathbf{\Omega}}\mathbf{N}\right) \\ & + \sum_{t=1}^{T}\sum_{i=1}^{r} z\log M_{it}\left(z^{-1}g_{it}(\mathbf{N})\right) - z\log\varepsilon \\ \mathsf{subject\ to} & \mathbf{N}\mathbf{1} = \mathbf{s},\ \mathbf{N} \geq \mathbf{0}, \end{array} \tag{4.19}$$

where $g_{it}(\mathbf{N}) = -\sum_{j=1}^{m}\ell_{jt}\mathbf{\Sigma}_{ji}$. ■

Note that $z\log M_{it}\left(z^{-1}g_{it}(\mathbf{N})\right)$ with $z > 0$ is the perspective function of the convex log-sum-exp function $\log M_{it}(g_{it}(\mathbf{N}))$ and thus is jointly convex in $(\mathbf{N}, z)$ [32]. Given that $\tilde{\mathbf{\Omega}} \succ \mathbf{0}$, we are able to conclude that problem (4.19) is convex.

[3]The Q-function is defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{u^2}{2}}du. \tag{4.18}$$

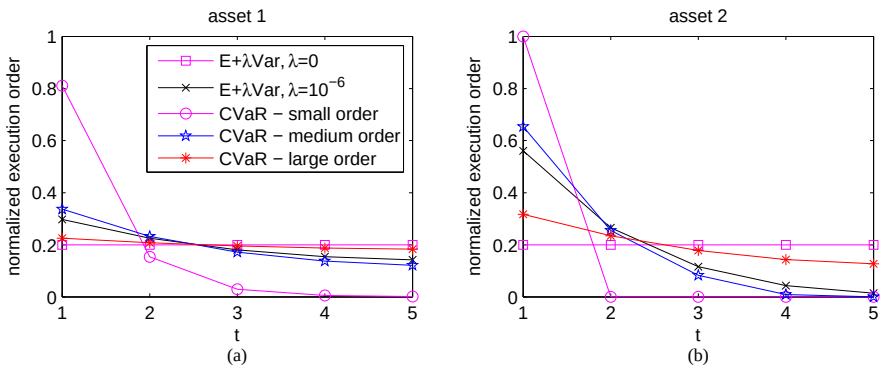

Figure 4.8: Order execution with $T = 5$: small order $= 0.2 \times \mathbf{s}$, medium order $\mathbf{s} = \left[10^6, 10^6\right]^T$, and large order $= 5 \times \mathbf{s}$. (a) asset 1. (b) asset 2.

Let us consider an illustrative example to understand different methods of order execution based on the Gaussian noise.

Example 4.1. Suppose there are $N = 2$ assets, $r = 2$ noise sources, and i.i.d. noise $\xi_{it} \sim \mathcal{N}(0, 1)$. The parameter matrices are

$$\mathbf{\Omega} = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} \times 10^{-6}, \quad \mathbf{\Theta} = \begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix} \times 10^{-7}, \tag{4.20}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.6191 & 0.1292 \\ 0.1292 & 0.6191 \end{bmatrix}. \tag{4.21}$$

We consider three kinds of sizes of the initial order: the medium initial order size is $\mathbf{s} = \left[10^6, 10^6\right]^T$, small order $= 0.2 \times \mathbf{s}$ and large order $= 5 \times \mathbf{s}$. We simulate three different methods: i) the problem (4.9) of minimizing the expected execution cost (or equivalently, the problem (4.10) with $\lambda = 0$), ii) the problem of minimizing the mean-variance trade-off of the execution cost with $\lambda = 10^{-6}$, and iii) the closed-form CVaR formulation of (4.17) for the Gaussian case with $\varepsilon = 0.05$.

Figure 4.8 shows the normalized order execution strategies of Example 4.1 with different initial order sizes. First, we find that minimizing the expected execution cost always distributes the order uniformly among the execution periods, which verifies the results in Section 4.3. Second, the mean-variance approach with fixed $\lambda = 10^{-6}$ adjusts the

execution strategies according to the variance of the execution cost; however, it always gives the same normalized execution order strategy no matter what the initial order size is. While the CVaR approach executes the small initial order faster to reduce the risk of not completing the execution, it spreads the large initial order more to avoid the huge market impact caused by one single large order. Another interesting observation is that asset 2 is executed faster than asset 1 because it has a smaller market impact (see (4.20)) and thus is more liquid. The results show that the CVaR approach can adjust the execution strategies depending on the initial order size but the mean-variance approach (including the case $\lambda = 0$) cannot. Thus, the CVaR approach is more appropriate for the order execution problem. ■

Remark 4.1. Apart from the above reviewed non-robust cases, there are also some other related works appearing simultaneously and independently. For example, for the robust mean-variance order execution problem see [146, 77, 78], for a numerical Monte Carlo simulation based CVaR formulation of the order execution see [147], and for the robust CVaR formulation of the order execution see [77, 78]. ■

Part II

Portfolio Optimization (Risk-Return Trade-off)

5

Portfolio Optimization with Known Parameters

Modeling of time series (overviewed in Part I) is at the core of and is a preliminary step in quantitative investment. The design of investment strategies is the natural next step and will be explored in the form of portfolio optimization (in Part II) and statistical arbitrage (in Part III).

As a start, Part II, this chapter introduces the most basic framework of Markowitz portfolio optimization under the assumption that the model parameters, i.e., the expected return $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the asset net returns, are perfectly known. We need to point out that in practice $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ need to be estimated from the past observations as discussed in the previous Chapter 3.

The organization of this chapter is as follows. Section 5.1 reviews the Markowitz mean-variance portfolio optimization. Section 5.2 points out two serious drawbacks of the Markowitz framework: variance as a risk measurement is not appropriate, and the mean-variance framework is very sensitive to parameter estimation errors. To overcome the first drawback, Section 5.2.1 covers the works on a single side risk measurement instead of variance. The literature results dealing with the second drawback are left to the next chapter, robust portfolio optimization.

5.1 Markowitz Mean-Variance Portfolio Optimization

The Markowitz mean-variance framework, introduced by Harry Markowitz [135] in 1952, provides a first quantitative approach to construct portfolios, which is the foundation of the nowadays Modern Portfolio Theory (for a comprehensive review, see [58]). Because of this fundamental contribution, Harry Markowitz shared the Nobel prize with another two researchers, Merton Miller and William Sharpe, in 1990.

The idea of the Markowitz framework is to find a trade-off between the expected return and the risk of the portfolio measured by the variance. Given that the expected return $\boldsymbol{\mu}$ and the positive definite covariance matrix $\boldsymbol{\Sigma}$ of the assets are perfectly known, the expected return and variance of a portfolio $\mathbf{w}$ are $\mathbf{w}^T\boldsymbol{\mu}$ and $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$, respectively.

Remark 5.1. Recall from Section 2.1.4 that it is the mean vector and covariance matrix for simple returns that are used for portfolio optimization. However, Part I mainly focuses on modeling log-returns since its statistical properties are more tractable. The good thing is that one can have the mean vector and covariance matrix for simple returns based on that for log-returns directly under the Gaussian assumption. That is, suppose the log-returns of N assets follow a multivariate Gaussian distribution $\mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, the mean vector and covariance matrix for the simple returns are

$$\boldsymbol{\mu} = e^{\bar{\boldsymbol{\mu}}+\bar{\boldsymbol{\sigma}}/2} - \mathbf{1} \tag{5.1}$$

$$\boldsymbol{\Sigma} = \left((\boldsymbol{\mu}+\mathbf{1})(\boldsymbol{\mu}+\mathbf{1})^T\right) \odot \left(e^{\bar{\boldsymbol{\Sigma}}} - \mathbf{1}_{N\times N}\right), \tag{5.2}$$

where $\bar{\boldsymbol{\sigma}} = [\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{NN}]^T$ is the vector of the variances of the N stocks, $\mathbf{1}$ is a N dimensional all one vector, $\mathbf{1}_{N\times N}$ is a N-by-N all one matrix, and $e^{\mathbf{X}}$ is an elementwise exponential operator, i.e., $[e^{\mathbf{X}}]_{ij} = e^{\mathbf{X}_{ij}}$. ■

5.1.1 Mean-Variance Trade-Off Optimization

There are three alternative but equivalent formulations, i.e., the risk minimization problem, return maximization problem, and risk-adjusted return maximization problem, and all of them are useful in practical applications.

Risk Minimization Problem

The risk minimization formulation aims at minimizing the portfolio variance with the expected portfolio return being above a given target:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T \boldsymbol{\mu} \geq \mu_0, \\ & \mathbf{w}^T \mathbf{1} = 1, \end{aligned} \tag{5.3}$$

where μ_0 is a expected return target parameter. The constraint $\mathbf{w}^T \mathbf{1} = 1$ is the capital budget constraint. Note that the above problem is convex given that $\boldsymbol{\Sigma}$ is positive definite and thus it can always be solved efficiently.

An interesting case of problem (5.3) that achieves the minimum variance regardless of the expected portfolio return is

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T \mathbf{1} = 1, \end{aligned} \tag{5.4}$$

which for obvious reasons is referred to as a global minimum variance portfolio (GMVP). Since the GMVP is a convex QP with only one linear equality constraint, solving the Karush-Kuhn-Tucker (KKT) optimality conditions [32] directly yields the closed-form solution expressed as follows:

$$\mathbf{w}_{\text{GMVP}} = \frac{1}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}} \boldsymbol{\Sigma}^{-1} \mathbf{1}. \tag{5.5}$$

Then the portfolio mean and variance of the GMVP are easily computed by

$$\mu_{\text{GMVP}} = \boldsymbol{\mu}^T \mathbf{w}_{\text{GMVP}} = \frac{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}}, \tag{5.6}$$

$$\sigma^2_{\text{GMVP}} = \mathbf{w}^T_{\text{GMVP}} \boldsymbol{\Sigma} \mathbf{w}_{\text{GMVP}} = \frac{1}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}}. \tag{5.7}$$

Return Maximization Problem

Instead of seeking the minimum variance, an alternative problem is to search for the maximum expected return with the variance under

control, say, less than a given target. This problem is referred to as a return maximization problem and has the following form:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \mathbf{w}^T\boldsymbol{\mu} \\ \text{subject to} & \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \leq \sigma_0^2, \\ & \mathbf{w}^T\mathbf{1} = 1, \end{array} \tag{5.8}$$

where σ_0^2 is the parameter that controls the variance target. Again, since the covariance matrix $\boldsymbol{\Sigma}$ is positive definite, the above problem has a linear objective with linear and convex quadratic constraints, and thus it is efficiently computable.

Risk-Adjusted Return Maximization Problem

The third problem formulation is to maximize a risk-adjusted return as follows:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \mathbf{w}^T\boldsymbol{\mu} - \lambda\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \end{array} \tag{5.9}$$

where $\lambda \geq 0$ is a given trade-off parameter between the portfolio expected return and variance. When $\lambda > 0$, it is a convex QP with only one linear constraint which admits a closed-form solution as follows:

$$\mathbf{w}^\star = \frac{1}{2\lambda}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} + \nu^\star\mathbf{1}\right), \tag{5.10}$$

where $\nu^\star$ is the optimal dual variable

$$\nu^\star = \frac{2\lambda - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}. \tag{5.11}$$

Efficient Frontier

Each of the above three problem formulations, i.e., (5.3), (5.8), and (5.9), has one controlling parameter and they are equivalent in the sense that when the parameters change (i.e., μ_0 changes from μ_{GMVP} to $+\infty$, σ_0^2 changes from σ_{GMVP}^2 to $+\infty$, and λ changes from 0 to $+\infty$), they result in the same mean-variance[1] trade-off curve (Pareto curve), which

[1] In the financial literature, it is standard deviation instead of variance that is used for illustrative purposes.

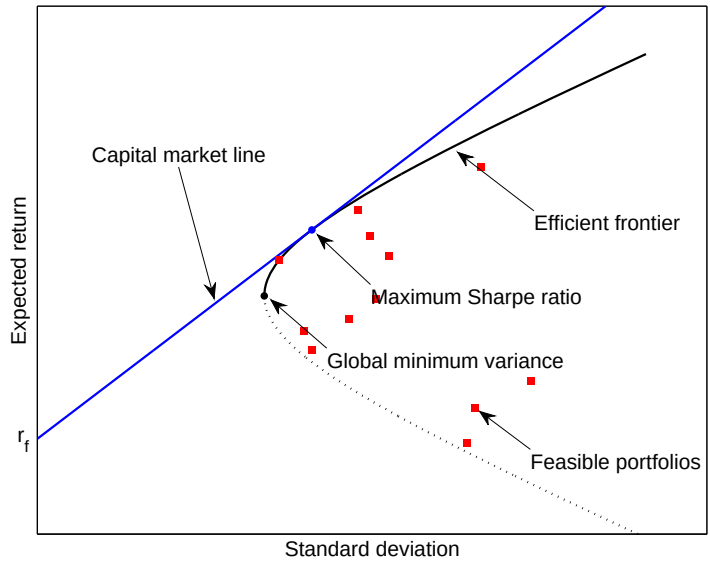

Figure 5.1: Illustration of the efficient frontier, capital market line, and global minimum variance and maximum Sharpe ratio portfolios.

is usually referred to as an efficient frontier in the financial literature, e.g., see [65, 58]. For example, when $\lambda \to +\infty$, the portfolio (5.10) goes to the GMVP (5.5).

Figure 5.1 shows the shape of an efficient frontier (see the black solid curve) and all the other feasible portfolios fall below the efficient frontier (see that all the red square points fall below the back solid curve). The GMVP is the leftmost point that has the minimum variance among all the feasible portfolios (see the black round dot). A simplified version code of Figure 5.1 is included in Appendix B.

5.1.2 Sharpe Ratio Optimization

All the portfolios on the efficient frontier are optimal depending on the investor's risk profile, that is, the choice of the parameters μ_0, σ_0, or λ. However, one may still ask which portfolio may be the most meaningful in practice. Precisely, Sharpe [179] first proposed the optimization of

the following problem:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \dfrac{\mathbf{w}^T\boldsymbol{\mu} - r_f}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \end{array} \tag{5.12}$$

where r_f is the return of a risk-free asset[2]. The objective of (5.12), is usually referred to as the Sharpe ratio, which measures the excess return (i.e., $\mathbf{w}^T\boldsymbol{\mu} - r_f$) normalized by the risk (i.e., $\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$), and the problem is thus called the Sharpe ratio maximization problem.

Since the Sharpe ratio is nonconcave, the Sharpe ratio maximization problem is not a convex problem. Fortunately, it can be reformulated in convex form as follows. First, note that $\mathbf{w}^T\mathbf{1} = 1$, then the problem (5.12) can be rewritten as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \dfrac{\mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1})}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1. \end{array} \tag{5.13}$$

Observe that the objective of (5.13) now is scale invariant w.r.t. $\mathbf{w}$, thus the constraint $\mathbf{w}^T\mathbf{1} = 1$ can be relaxed to $\mathbf{w}^T\mathbf{1} > 0$ and then one can arbitrarily set $\mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1}) = 1$ and minimize $\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ instead. Thus, the problem (5.13) can be further reformulated into a convex form:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} & \mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1}) = 1, \\ & \mathbf{w}^T\mathbf{1} > 0. \end{array} \tag{5.14}$$

Any normalized solution of (5.14) so that the summation of all the portfolio weight values being one is an optimal solution of (5.13).

The problem (5.14) without $\mathbf{w}^T\mathbf{1} > 0$ is a convex QP with only one linear equality constraint and thus admits a closed-form solution:

$$\mathbf{w}_{\text{SR}} = \frac{1}{(\boldsymbol{\mu} - r_f\mathbf{1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1})}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - r_f\mathbf{1}), \tag{5.15}$$

[2]Usually a risk-free asset is assumed to have zero risk or variance. In practice, for example, the US Treasuries, especially T-bills, are considered as risk-free assets because they are backed by the U.S. government.

then $\mathbf{w}_{\text{SR}}$ is also an optimal solution of the problem (5.14) if $\mathbf{w}_{\text{SR}}^T\mathbf{1} > 0$ (which is always observed in practice); otherwise, one can always find an optimal solution of (5.14) efficiently via a standard optimization solver since it is a convex QP.

Figure 5.1 shows the Sharpe ratio point on the efficient frontier (see the blue round point) that has the maximum Sharpe ratio (or equivalently, the maximum slope between the points on the efficient frontier and the risk-free point). If one is allowed to borrow or lend the risk-free asset, then he/she can have a portfolio that falls on the solid blue line, which is usually referred to as the capital market line in the financial literature [58].

Another interesting observation is that when $r_f = 0$ and all the assets have the same expected return, i.e., $\boldsymbol{\mu} = \alpha\mathbf{1}$ for some $\alpha > 0$, the Sharpe ratio solution (5.15) coincides with the GMVP in (5.5).

5.1.3 Connections between Portfolio and Beamforming

Let us first start with introducing the formulation of beamforming. The output of a narrowband beamformer is given by

$$y(t) = \mathbf{w}^H\mathbf{x}(t), \tag{5.16}$$

where t is the time index, $\mathbf{x}(t) \in \mathbb{C}^N$ is the complex vector of array observations (i.e., measurements at different antennas), $\mathbf{w} \in \mathbb{C}^N$ is the complex vector of beamformer weights, and N is the number of array sensors.

The observation vector is modeled as

$$\mathbf{x}(t) = \underbrace{s(t)\mathbf{a}}_{\triangleq \mathbf{s}(t)} + \mathbf{i}(t) + \mathbf{n}(t), \tag{5.17}$$

where $\mathbf{s}(t)$, $\mathbf{i}(t)$, and $\mathbf{n}(t)$ are the desired signal, interference, and noise components, respectively. The signal $s(t)$ is the temporal waveform and $\mathbf{a}$ is the spatial steering vector.

Then the goal of beamforming design is to design a weight vector or beamvector $\mathbf{w}$ that maximizes the SINR [149]:

$$\underset{\mathbf{w}}{\text{maximize}} \quad \frac{\sigma_s^2|\mathbf{w}^H\mathbf{a}|^2}{\mathbf{w}^H\mathbf{R}\mathbf{w}} \tag{5.18}$$

where σ_s^2 is the signal power, $|\cdot|$ denotes the magnitude of a complex number, and

$$\mathbf{R} = \mathsf{E}\left[(\mathbf{i}(t) + \mathbf{n}(t))(\mathbf{i}(t) + \mathbf{n}(t))^H\right] \tag{5.19}$$

is the $N \times N$ interference-plus-noise covariance matrix.

Note that the objective of (5.18) is invariant to the magnitude and the phase of $\mathbf{w}$, thus one can arbitrarily set the complex number $\mathbf{w}^H\mathbf{a}$ to be real and equal to one, i.e., $\mathbf{w}^H\mathbf{a} = 1$, and then the problem (5.18) can be reformulated as [149]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\mathsf{minimize}} & \mathbf{w}^H\mathbf{R}\mathbf{w} \\ \mathsf{subject\ to} & \mathbf{w}^H\mathbf{a} = 1, \end{array} \tag{5.20}$$

which is the problem (1.4) mentioned in the introduction of Chapter 1.

The solution is found in closed-form as

$$\mathbf{w} = \frac{1}{\mathbf{a}^H\mathbf{R}^{-1}\mathbf{a}}\mathbf{R}^{-1}\mathbf{a}, \tag{5.21}$$

which shares the same mathematical form as the GMVP in (5.5) with the real-valued net returns covariance matrix $\mathbf{\Sigma}$ being replaced by the complex-valued interference-plus-noise covariance matrix and the constant vector $\mathbf{1}$ being replaced by the complex-valued signal steering vector $\mathbf{a}$.

5.1.4 Practical Constraints

In practice, the optimization problems are not as clean as stated above and there are always some additional constraints due to market regularizations, capital budgets, investors' preferences, etc. (some of which are not even convex) [65, 63].

Long-Only Constraints

This is the most natural constraint and models the fact that one cannot sell what one does not have:

$$\mathbf{w} \geq 0. \tag{5.22}$$

This is a usual constraint since many funds and institutional investors are not allowed to short-sell in the market, which means selling what one does not have, and would translate into a negative weight (since that value is owed rather than owned).

Turnover Constraints

If we denote the current portfolio as $\mathbf{w}_0$, and the target portfolio to be designed as $\mathbf{w}$, then $\Delta\mathbf{w} \triangleq \mathbf{w} - \mathbf{w}_0$ denotes the turnover, i.e., the capital to be traded. Usually, the smaller the turnover is, the lower the transaction cost is. Thus, we can limit the turnover either on each asset

$$|\Delta w_i| \leq U_i \tag{5.23}$$

or on the whole portfolio:

$$\|\Delta\mathbf{w}\|_1 \leq U. \tag{5.24}$$

For example, it is practical to restrict the turnover of an asset to be less than 5% of the average daily volume of the asset.

Holding Constraints

It is also common in practice to limit the weights in each asset, that is,

$$L_i \leq w_i \leq U_i, \tag{5.25}$$

where L_i and U_i are lower and upper bounds of the holdings of asset i.

Another issue is that one has to pay a fixed minimum brokerage fee no matter how small the order is. Thus too small holdings are not desired in practice and they can be avoided by adding the following (nonconvex) constraints:

$$|w_i| \geq L_i \mathbb{1}_{\{w_i \neq 0\}}, \tag{5.26}$$

where L_i is the smallest holding size of asset i.

Cardinality Constraints

It is also suggested to restrict the number of assets in some scenarios, e.g., it is practical to use only a few stocks to track the market index.

Mathematically speaking, this constraint reads

$$\|\mathbf{w}\|_0 \leq K. \tag{5.27}$$

5.2 Drawbacks of Markowitz Framework

Even though the Markowitz framework is quantitatively easy to understand, it has two serious drawbacks that have made the framework not used in practice for many years.

5.2.1 Variance Is Not Appropriate

As motivated in Section 4.5 for an order execution problem, variance is not a good risk measurement in practice since it penalizes both the unwanted high transaction costs and the desired low transaction costs (for short-selling it is the opposite).

This argument indeed applies to the portfolio optimization since only the high portfolio losses[3] are unwanted and it is thus more practical to penalize these only but not the low portfolio losses, see [7, 63, 65, 172, 100].

To overcome this drawback, there are many single side risk measurements, e.g., Roy's safety-first, semi-variance, lower partial moment, VaR, CVaR, etc. [65], proposed in the financial literature. Among them, CVaR enjoys the widest popularity due to its mathematical tractability, thus in the next subsection we mainly review the application of CVaR in portfolio optimization.

CVaR Portfolio Optimization

Actually, one of the first popular single side risk measurements was Value-at-Risk (VaR) initially proposed by J.P. Morgan.[4] Denote $\mathbf{r}$ as a multivariate random variable of the asset returns, and the portfolio loss is $-\mathbf{w}^T\mathbf{r}$. Rockafellar and Uryasev [166] first proposed to minimize

[3]The portfolio loss is the negative portfolio return. Thus high portfolio losses mean low portfolio returns.

[4]See http://www.value-at-risk.net/riskmetrics/.

the CVaR of the portfolio loss as follows:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \text{CVaR}_{1-\varepsilon}(-\mathbf{w}^T\mathbf{r}) \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \end{array} \tag{5.28}$$

where the definition of CVaR has been introduced in Section 4.5.1.

Again, the objective of the problem (5.28) contains a conditional expectation exceeding a threshold that is not fixed, which in general is not easy to deal with.

Following the technique in [166] (which has been introduced in Section 4.5.1) and given the past observations $\mathbf{r}_t$, $t = 1, \ldots, T$, of $\mathbf{r}$, one has the sample average approximation (SAA) of (5.28) as follows:

$$\begin{array}{ll} \underset{\mathbf{w},\mathbf{z},\zeta}{\text{minimize}} & \zeta + \dfrac{1}{\varepsilon T}\displaystyle\sum_{t=1}^{T} z_t \\ \text{subject to} & 0 \leq z_t \geq -\mathbf{w}^T\mathbf{r}_t - \zeta, \quad t = 1, \ldots, T \\ & \mathbf{w}^T\mathbf{1} = 1. \end{array} \tag{5.29}$$

Remark 5.2. Similar to the order execution problem in Section 4.4, one can have either an equivalent convex formulation for a Gaussian distribution (e.g., see [166]) or a safe approximation convex approximation for general non-Gaussian distributions satisfying Assumption 4.1. Since it is straightforward, we omit it here. ■

Remark 5.3. Now we have seen that CVaR as a single side risk measurement has been applied in both order execution and portfolio design. Interestingly, researchers in signal processing and wireless communication communities have become aware of this useful mathematical technique. Recently, it has been used to design some chance-constrained wireless communication networks for more reliable communications even under some extreme events, e.g., see [124, 183, 112]. ■

5.2.2 Markowitz Framework Is Too Sensitive (Lack of Robustness)

The second drawback is that the Markowitz framework is very sensitive to the parameters, i.e., the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, but especially the mean vector [63]. For illustrative purposes, here

Table 5.1: Performance of the maximum Sharpe ratio portfolios under different parameter perturbations. The optimal portfolio $\mathbf{w}^\star$ is the portfolio of the case of No. Err.

Param. Err.	$\mathbf{w}$	$\frac{\|\mathbf{w}-\mathbf{w}^\star\|_2}{\|\mathbf{w}^\star\|_2}$	SR
No Err.	$[\ 0.9909,\ 0.4088, -0.3997]^T$	0	0.2551
Mean Err.	$[\ 0.1341,\ 0.5976,\ 0.2683]^T$	0.9639	0.2377
Cov. Err.	$[\ 0.1103,\ 0.6140,\ 0.2757]^T$	0.9865	0.2363
Mean&Cov. Err.	$[-0.2572,\ 0.6576,\ 0.5996]^T$	1.4144	0.2057

we use a simple numerical example to show how a slightly insignificant error can dramatically distort the optimal portfolio.

Example 5.1. Suppose there are three assets with $\mu_1 = \mu_2 = 8\%$ and $\mu_3 = 5\%$ and volatilities of the three assets are $\sigma_1 = 20\%$, $\sigma_2 = 22\%$, $\sigma_3 = 10\%$ and the correlations are $\rho_{ij} = 0.8$.

Let us focus on solving the maximum Sharpe ratio problem (5.12) with $r_f = 3\%$ under four scenarios: i) all the parameters are known exactly (referred to as No Err.), ii) there is a slight error in μ_1 so that the estimated value is $\hat{\mu}_1 = 7\%$ (referred to as Mean Err.), iii) there is an error in σ_1 such that the estimated value is $\hat{\sigma}_1 = 25\%$ (referred to as Cov. Err.), and iv) the combination of ii) and iii) (referred to as Mean&Cov. Err.).

Table 5.1 shows the numerical results of the solved portfolios, the relative differences, and the Sharpe ratios (SR). For example, if we compare Mean Err. with No Err., we can see that changing the mean of the first asset from 8% to 7% dramatically changes the portfolio weights vector: the relative difference is 0.9639. Similar results can be obtained if we compare Cov. Err. with No Err., and the difference becomes even larger if there are both errors in the mean vector and covariance matrix, see Mean&Cov. Err. versus No Err. ■

There are many works, e.g., [55, 200, 86, 63], that focus on overcoming this drawback fully and we will review them separately from this chapter in the upcoming Chapter 6.

5.3 Black-Litterman Model

The Black-Litterman model is an alternative approach dealing with the sensitivity issue in expected excess returns to some degree. It combines market equilibrium and investors' views to result in a more robust expected return estimate, based on which the optimized portfolio is relatively more stable [22, 23, 24, 104].

For simplicity, let us suppose for the Black-Litterman model the true covariance $\boldsymbol{\Sigma}$ is known and the goal is to produce a stable estimate of the expected excess returns $\boldsymbol{\mu}$.

Let us first start with the two information sources based on which the Black-Litterman model can be built, i.e., market equilibrium and investors' views.

Market Equilibrium. The first important assumption is that a market equilibrium can provide an estimate of the expected excess returns, denoted as $\boldsymbol{\pi}$, close to the true unknown expected excess returns $\boldsymbol{\mu}$. Mathematically, it can be expressed as follows:

$$\boldsymbol{\pi} = \boldsymbol{\mu} + \mathbf{w}_{\boldsymbol{\pi}}, \quad \mathbf{w}_{\boldsymbol{\pi}} \sim \mathcal{N}(0, \tau\boldsymbol{\Sigma}) \tag{5.30}$$

where the parameter $\tau > 0$, which measures the uncertainty in the estimate $\boldsymbol{\pi}$, and the smaller τ is, the less uncertain the estimate is. A specific is provided later in Example .

Investors' View. Suppose there are K views summarized from some investors, the Black-Litterman model quantifies them via a linear system:

$$\mathbf{q} = \mathbf{P}\boldsymbol{\mu} + \mathbf{w}_{\mathbf{q}}, \quad \mathbf{w}_{\mathbf{q}} \sim \mathcal{N}(0, \boldsymbol{\Omega}), \tag{5.31}$$

where $\mathbf{P} \in \mathbb{R}^{K\times N}$ and $\mathbf{q} \in \mathbb{R}^K$ characterize the absolute or relative K views and $\boldsymbol{\Omega} \in \mathbb{R}^{K\times K}$ measures the uncertainty in the views. A specific example is provided later in Example .

The expected excess returns based on the market equilibrium (5.30) and the investors' views (5.31) actually can be written together in a

more compact form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{w}_{\mathrm{BL}}, \tag{5.32}$$

where $\mathbf{w}_{\mathrm{BL}} \sim \mathcal{N}(0, \mathbf{V})$ and

$$\mathbf{y} \triangleq \begin{bmatrix} \boldsymbol{\pi} \\ \mathbf{q} \end{bmatrix}, \quad \mathbf{X} \triangleq \begin{bmatrix} \mathbf{I} \\ \mathbf{P} \end{bmatrix}, \quad \mathbf{V} \triangleq \begin{bmatrix} \tau\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix}. \tag{5.33}$$

Obviously, (5.32) is a standard linear model for the true expected excess returns with white Gaussian noise. The Gaussian ML estimator, i.e., the minimizer of the following problem

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}), \tag{5.34}$$

is a better estimate since it combines the market equilibrium and investors views. Easily, setting the derivative of (5.34) to zero yields the closed-form solution:

$$\hat{\boldsymbol{\mu}}_{\mathrm{BL}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \tag{5.35}$$

$$= \left(\begin{bmatrix} \mathbf{I} & \mathbf{P}^T \end{bmatrix} \begin{bmatrix} (\tau\boldsymbol{\Sigma})^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{P} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{I} & \mathbf{P}^T \end{bmatrix} \begin{bmatrix} (\tau\boldsymbol{\Sigma})^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\pi} \\ \mathbf{q} \end{bmatrix} \tag{5.36}$$

$$= \left((\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^T \boldsymbol{\Omega}^{-1} \mathbf{P} \right)^{-1} \left((\tau\boldsymbol{\Sigma})^{-1} \boldsymbol{\pi} + \mathbf{P}^T \boldsymbol{\Omega}^{-1} \mathbf{q} \right). \tag{5.37}$$

We can further understand the above solution (5.37) deeper as follows. Since the objective of (5.34) can be rewritten as

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) \tag{5.38}$$

$$= \left(\begin{bmatrix} \boldsymbol{\pi} \\ \mathbf{q} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{P}\boldsymbol{\mu} \end{bmatrix} \right)^T \begin{bmatrix} (\tau\boldsymbol{\Sigma})^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}^{-1} \end{bmatrix} \left(\begin{bmatrix} \boldsymbol{\pi} \\ \mathbf{q} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{P}\boldsymbol{\mu} \end{bmatrix} \right)$$

$$= \frac{1}{\tau} (\boldsymbol{\pi} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \boldsymbol{\mu}) + (\mathbf{q} - \mathbf{P}\boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1} (\mathbf{q} - \mathbf{P}\boldsymbol{\mu}), \tag{5.39}$$

problem (5.34) actually equals

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad (\boldsymbol{\pi} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\pi} - \boldsymbol{\mu}) + \tau (\mathbf{q} - \mathbf{P}\boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1} (\mathbf{q} - \mathbf{P}\boldsymbol{\mu}). \tag{5.40}$$

The objective combines the market equilibrium towards the investors' views with τ being the trade-off parameter. There are two extreme cases

- when $\tau = 0$, the objective does not consider any view and the optimal solution is only based on the market equilibrium:

$$\hat{\boldsymbol{\mu}}_{\text{me}} = \boldsymbol{\pi}; \tag{5.41}$$

- when $\tau \to +\infty$, the objective emphasizes on the investors' views only and the optimal solution goes to

$$\hat{\boldsymbol{\mu}}_{\text{view}} = (\mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P})^{-1}\mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{q}. \tag{5.42}$$

Interestingly, the general Black-Litterman estimate (5.37) can be rewritten as follows:

$$\hat{\boldsymbol{\mu}}_{\text{BL}} = \left((\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}\right)^{-1}\left((\tau\boldsymbol{\Sigma})^{-1}\boldsymbol{\pi} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{q}\right) \tag{5.43}$$

$$= \left((\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}\right)^{-1}\left((\tau\boldsymbol{\Sigma})^{-1}\hat{\boldsymbol{\mu}}_{\text{me}} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}\hat{\boldsymbol{\mu}}_{\text{view}}\right) \tag{5.44}$$

$$= \underbrace{\left((\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}\right)^{-1}(\tau\boldsymbol{\Sigma})^{-1}}_{\mathbf{W}_{\text{me}} \triangleq}\hat{\boldsymbol{\mu}}_{\text{me}} + \underbrace{\left((\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}^T\boldsymbol{\Omega}^{-1}\mathbf{P}}_{\mathbf{W}_{\text{view}} \triangleq}\hat{\boldsymbol{\mu}}_{\text{view}}, \tag{5.45}$$

which is simply a linear weighted combination of the two extreme solutions $\hat{\boldsymbol{\mu}}_{\text{me}}$ and $\hat{\boldsymbol{\mu}}_{\text{view}}$ and the weight matrices satisfy

$$\mathbf{W}_{\text{me}} + \mathbf{W}_{\text{view}} = \mathbf{I}. \tag{5.46}$$

Clearly, the Black-Litterman expected excess returns (5.45) shrinks the market equilibrium towards the investors' views. This idea of the Black-Litterman model indeed is similar to the previous James-Stein shrinkage estimator (3.30) with three differences:

- the sample mean estimate in (3.30) is replaced by the expected excess returns estimated based on the market equilibrium $\hat{\boldsymbol{\mu}}_{\text{me}}$;

- the specific target in (3.30) is replaced by the estimate of the expected excess returns investors' view $\hat{\boldsymbol{\mu}}_{\text{view}}$; and

- the scalar trade-off (or shrinkage) parameter in (3.30) is changed to a matrix instead.

Thus, we can see that the Black-Litterman model is a more precise model for producing stable and reliable expected excess returns (or equivalently expected returns since the risk-free rate is almost always known).

The above models of market equilibrium (5.30) and investors' views (5.31) are quite general. This generality enables the popularity of the Black-Litterman model. In the following we consider some specific examples for both of them.

Example 5.2. One of the most popular models for market equilibrium is the CAPM[5] (2.20)

$$\mathsf{E}[r_i] - r_f = \beta_i(\mathsf{E}[r_M] - r_f), \tag{5.47}$$

where $\mathsf{E}[r_i]$, $\mathsf{E}[r_M]$, and r_f are the expected returns on the i-stock, the expected return on the market portfolio, and the risk-free rate, respectively. The sensitivity of the expected excess return of the stock to that of the market is captured by the beta (2.21):

$$\beta_i = \frac{\mathsf{Cov}(r_i, r_M)}{\mathsf{Var}(r_M)} \tag{5.48}$$

and $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_N]^T$.

Let $\mathbf{w}_M \triangleq [w_{1M}, \ldots, w_{NM}]^T \in \mathbb{R}^N$ denote the market portfolio of the N stocks, thus the market return is

$$r_M = \mathbf{r}^T \mathbf{w}_M, \tag{5.49}$$

where $\mathbf{r} \triangleq [r_1, \ldots, r_N]^T$ contains the returns of the N stocks.

Substituting (5.48) and (5.49) into (5.47), the estimated expected excess returns of the N stocks are as follows:

$$\boldsymbol{\pi} \triangleq \begin{bmatrix} \mathsf{E}[r_1] - r_f \\ \vdots \\ \mathsf{E}[r_N] - r_f \end{bmatrix} = \boldsymbol{\beta}(\mathsf{E}[r_M] - r_f) \tag{5.50}$$

[5]Actually, the CAPM model was used in the initial derivaiton of the Black-Litterman model [22, 23, 24]. For simplicity, we drop the time index t in this section.

$$= \frac{\mathsf{E}[r_M] - r_f}{\mathsf{Var}(r_M)} \begin{bmatrix} \mathsf{Cov}(r_1, r_M) \\ \vdots \\ \mathsf{Cov}(r_N, r_M) \end{bmatrix} = \frac{\mathsf{E}[r_M] - r_f}{\mathsf{Var}(r_M)} \begin{bmatrix} \mathsf{Cov}(r_1, \mathbf{r}^T \mathbf{w}_M) \\ \vdots \\ \mathsf{Cov}(r_N, \mathbf{r}^T \mathbf{w}_M) \end{bmatrix} \tag{5.51}$$

$$= \underbrace{\frac{\mathsf{E}[r_M] - r_f}{\mathsf{Var}(r_M)}}_{\delta \triangleq} \underbrace{\begin{bmatrix} \mathsf{Cov}(r_1, r_1) & \dots & \mathsf{Cov}(r_1, r_N) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(r_N, r_1) & \dots & \mathsf{Cov}(r_N, r_N) \end{bmatrix}}_{\boldsymbol{\Sigma} \triangleq} \mathbf{w}_M \tag{5.52}$$

$$= \delta \boldsymbol{\Sigma} \mathbf{w}_M. \tag{5.53}$$

That is, $\boldsymbol{\pi}$ in (5.30) is replaced by the quantity $\delta \boldsymbol{\Sigma} \mathbf{w}_M$. ■

Example 5.3. Let us consider an example from [65] to understand how the model (5.31) expresses the views. Suppose there are $N = 5$ stocks and two independent views on them:

- Stock 1 will have excess return of 1.5% with standard deviation 1%;
- Stock 3 will outperform Stock 2 by 4% with a standard deviation 1%.

Mathematically, the above two independent views can be expressed as

$$\begin{bmatrix} 1.5\% \\ 4\% \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \mathbf{w}_{\mathbf{q}}, \tag{5.54}$$

where $\mathbf{w}_{\mathbf{q}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{\Omega} = \begin{bmatrix} 1\%^2 & 0 \\ 0 & 1\%^2 \end{bmatrix}$. ■

Once a Black-Litterman expected excess returns $\hat{\boldsymbol{\mu}}_{\mathrm{BL}}$ has been estimated, we can further plug it and the known true covariance matrix[6] $\boldsymbol{\Sigma}$ into the previously mentioned mean-variance portfolio optimization framework to achieve some desired portfolios.

[6] Keep in mind that the covariance matrix also needs to be estimated in practice.

Remark 5.4. The Black-Litterman expected excess returns (5.37) requires the trade-off parameter τ, investors' view $\mathbf{P}$ and $\mathbf{q}$ and the confidence parameter $\mathbf{\Omega}$. In general, they are difficult to specify. For example, different researchers have different views on selecting the parameter τ: some experience researchers generally set $\tau \in [0.01, 0.05]$ [104], some prefers to use $\tau = 1$ directly [174], while some suggest the value 1 divided by the number of observations [26]. Here we only outline the idea of Black-Litterman model but do not explore these difficulties. The interested readers may please refer to [104] and references therein for more detailed discussions. ■

6

Robust Portfolio Optimization

Markowitz portfolio optimization requires knowledge of the mean return vector and covariance matrix parameters. As it turns out, the resulting optimized portfolio is so highly sensitive to small estimation errors in such parameters that it is unusable in practice (indeed practitioners seldom use such a naive design). One step towards the solution is to make the portfolio design robust to uncertainties in the parameters.

This chapter reviews the robust portfolio optimization that uses some uncertainty sets to capture the estimation errors and then takes such uncertainty sets into problem formulations.

The organization of this chapter is as follows. Section 6.1 reviews the robust mean-variance portfolio optimization and Section 6.2 concentrates on the robust Sharpe ratio maximization. At the end, Section 6.3 makes some specific connections between robust portfolio optimization in financial engineering and robust beamforming in signal processing.

6.1 Robust Mean-Variance Trade-off Portfolio Optimization

Recall that in Section 5.1.1 there are three alternative mean-variance trade-off optimization formulations, i.e., (5.4), (5.8), and (5.9). Since the formulations are equivalent in the sense that they give the same efficient frontier, for simplicity, we focus on (5.9) which is restated as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \mathbf{w}^T\boldsymbol{\mu} - \lambda \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{6.1}$$

where $\lambda \geq 0$ is the trade-off parameter, $\mathcal{W}$ denotes the set of other convex constraints, and we further define $\overline{\mathcal{W}} \triangleq \{\mathbf{w}|\mathbf{w}^T\mathbf{1} = 1\} \cap \mathcal{W}$ and assume $\overline{\mathcal{W}}$ is convex and compact.

To design the robust counterpart of (6.1), here we assume that the uncertainty sets of the mean return $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are separable, convex, and compact, and they are denoted as $\mathcal{U}_{\boldsymbol{\mu}}$ and $\mathcal{U}_{\boldsymbol{\Sigma}}$, respectively. A conservative and practical investment approach is to optimize the worst-case objective over the uncertainty sets, which leads to the following robust counterpart of (6.1):

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}} \mathbf{w}^T\boldsymbol{\mu} - \lambda \max_{\boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{aligned} \tag{6.2}$$

6.1.1 Minimax or Maximin

It is obvious that the objective of (6.2) is concave in $\mathbf{w}$ and is linear (and thus convex) in both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Under the condition that $\overline{\mathcal{W}}$, $\mathcal{U}_{\boldsymbol{\mu}}$, and $\mathcal{U}_{\boldsymbol{\Sigma}}$ are convex and compact sets, one can easily get that

$$\max_{\mathbf{w}\in\overline{\mathcal{W}}} \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}},\, \boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}} \{\mathbf{w}^T\boldsymbol{\mu} - \lambda \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\} = \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}},\, \boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}} \max_{\mathbf{w}\in\overline{\mathcal{W}}} \{\mathbf{w}^T\boldsymbol{\mu} - \lambda \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\} \tag{6.3}$$

based on the minimax theory [165]. Therefore, one can equivalently solve either the minimax or maximin formulations, whichever is computationally cheaper in practice. Some specific examples of numerical iterative algorithms can be found in [127, 200].

However, instead of solving a double-layered minimax or maximin problem numerically, which in general is computationally costly, one

may either find the worst-case mean and variance in closed-form directly or reformulate the worst-case formulation as some simpler maximization problem so that (6.2) reduces into a single-layered convex maximization problem (e.g., QP, QCQP, or SDP). In the following, we will review different types of uncertainty sets such that (6.2) can be reformulated to a simple single-layered convex problem.

6.1.2 Worst-Case Mean

Let us start with the worst-case mean first. We consider two types of the uncertainty set up for the mean vector $\mathcal{U}_{\boldsymbol{\mu}}$, i.e., box and elliptical sets.

Box Uncertainty Set

The box uncertainty set is given by

$$\mathcal{U}_{\boldsymbol{\mu}}^{b} = \{\boldsymbol{\mu} | -\boldsymbol{\delta} \leq \boldsymbol{\mu} - \hat{\boldsymbol{\mu}} \leq \boldsymbol{\delta}\}, \tag{6.4}$$

where the predefined parameters $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\delta}$ denote the location and size of the box uncertainty set, respectively.

We can easily derive the worst-case mean as

$$\min_{\boldsymbol{\mu} \in \mathcal{U}_{\boldsymbol{\mu}}^{b}} \mathbf{w}^T \boldsymbol{\mu} = \mathbf{w}^T \hat{\boldsymbol{\mu}} + \min_{-\boldsymbol{\delta} \leq \boldsymbol{\gamma} \leq \boldsymbol{\delta}} \mathbf{w}^T \boldsymbol{\gamma} = \mathbf{w}^T \hat{\boldsymbol{\mu}} - |\mathbf{w}|^T \boldsymbol{\delta}, \tag{6.5}$$

where $|\mathbf{w}|$ denotes elementwise absolute value of $\mathbf{w}$.

Elliptical Uncertainty Set

The elliptical uncertainty set[1] is

$$\mathcal{U}_{\boldsymbol{\mu}}^{e} = \{\boldsymbol{\mu} | (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^T \mathbf{S}_{\boldsymbol{\mu}}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \leq \delta_{\boldsymbol{\mu}}^2\}, \tag{6.6}$$

where the predefined parameters $\hat{\boldsymbol{\mu}}$, $\delta_{\boldsymbol{\mu}} > 0$, and $\mathbf{S}_{\boldsymbol{\mu}} \succ \mathbf{0}$ denote the location, size, and the shape of the uncertainty set, respectively. The worst-case mean is

$$\min_{\boldsymbol{\mu} \in \mathcal{U}_{\boldsymbol{\mu}}^{e}} \mathbf{w}^T \boldsymbol{\mu} = \min_{\left\| \mathbf{S}_{\boldsymbol{\mu}}^{-1/2} \boldsymbol{\gamma} \right\|_2 \leq \delta_{\boldsymbol{\mu}}} \mathbf{w}^T (\hat{\boldsymbol{\mu}} + \boldsymbol{\gamma}) = \mathbf{w}^T \hat{\boldsymbol{\mu}} + \min_{\left\| \mathbf{S}_{\boldsymbol{\mu}}^{-1/2} \boldsymbol{\gamma} \right\|_2 \leq \delta_{\boldsymbol{\mu}}} \mathbf{w}^T \boldsymbol{\gamma}$$

[1] A special case is $\mathbf{S} = \mathbf{I}$ and the uncertainty set becomes a sphere.

$$= \mathbf{w}^T \hat{\boldsymbol{\mu}} + \min_{\|\tilde{\boldsymbol{\gamma}}\|_2 \leq \delta_{\boldsymbol{\mu}}} \mathbf{w}^T \mathbf{S}_{\boldsymbol{\mu}}^{1/2} \tilde{\boldsymbol{\gamma}} = \mathbf{w}^T \hat{\boldsymbol{\mu}} - \delta_{\boldsymbol{\mu}} \left\| \mathbf{S}_{\boldsymbol{\mu}}^{1/2} \mathbf{w} \right\|_2. \quad (6.7)$$

It is easy to check that both the worst-case values (6.5) and (6.7) are concave in $\mathbf{w}$, which is desired since (6.2) is a maximization problem.

6.1.3 Worst-Case Variance Based on $\boldsymbol{\Sigma}$ Directly

Now let us focus on the worst-case variance and we start by incorporating the uncertainty into the covariance matrix $\boldsymbol{\Sigma}$ directly.

Box Uncertainty Set

Again, let us elementwise first consider the box type uncertainty set as follows:

$$\mathcal{U}_{\boldsymbol{\Sigma}}^b = \{\boldsymbol{\Sigma} | \underline{\boldsymbol{\Sigma}} \leq \boldsymbol{\Sigma} \leq \overline{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma} \succeq \mathbf{0}\}, \quad (6.8)$$

where $\underline{\boldsymbol{\Sigma}}$ and $\overline{\boldsymbol{\Sigma}}$ are as lower and upper bounds.

A special case is that if $\overline{\boldsymbol{\Sigma}} \succeq \mathbf{0}$ and $\mathbf{w} \geq 0$ holds, the worst-case variance can be found directly [200]:

$$\max_{\boldsymbol{\Sigma} \in \mathcal{U}_{\boldsymbol{\Sigma}}^b} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} = \mathbf{w}^T \overline{\boldsymbol{\Sigma}} \mathbf{w}. \quad (6.9)$$

However, when either $\overline{\boldsymbol{\Sigma}} \succeq \mathbf{0}$ or $\mathbf{w} \geq 0$ may not hold, the worst-case variance does not have a closed-form expression anymore. Fortunately, an equivalent formulation can be found as follows. First note that the worst-case value $\max_{\boldsymbol{\Sigma} \in \mathcal{U}_{\boldsymbol{\Sigma}}^b} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}$ is given by the convex problem

$$\begin{array}{ll} \underset{\boldsymbol{\Sigma}}{\text{maximize}} & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} & \underline{\boldsymbol{\Sigma}} \leq \boldsymbol{\Sigma} \leq \overline{\boldsymbol{\Sigma}}, \\ & \boldsymbol{\Sigma} \succeq \mathbf{0}. \end{array} \quad (6.10)$$

Then it is easy to have the equivalent dual problem of (6.10) as [127, 63]

$$\begin{array}{ll} \underset{\overline{\boldsymbol{\Lambda}}, \underline{\boldsymbol{\Lambda}}}{\text{minimize}} & \operatorname{Tr}(\overline{\boldsymbol{\Lambda}} \overline{\boldsymbol{\Sigma}}) - \operatorname{Tr}(\underline{\boldsymbol{\Lambda}} \underline{\boldsymbol{\Sigma}}) \\ \text{subject to} & \begin{bmatrix} \overline{\boldsymbol{\Lambda}} - \underline{\boldsymbol{\Lambda}} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq \mathbf{0}, \\ & \overline{\boldsymbol{\Lambda}} \geq \mathbf{0}, \quad \underline{\boldsymbol{\Lambda}} \geq \mathbf{0}, \end{array} \quad (6.11)$$

which is a convex SDP, and in fact the constraints are jointly convex in the inner dual variable variables $\overline{\mathbf{\Lambda}}$ and $\underline{\mathbf{\Lambda}}$ and the outer variable $\mathbf{w}$.

Now we can easily have a specific equivalent formulation of (6.2) as follows. Given the uncertainty sets $\mathcal{U}_{\boldsymbol{\mu}}^b$ and $\mathcal{U}_{\mathbf{\Sigma}}^b$, (6.2) equals the convex Problem I in Table 6.1. In fact, Table 6.1 summarizes all the convex problems for all the possible combinations of uncertainty sets.

Elliptical Uncertainty Set

The elliptical uncertainty set of the covariance matrix can be defined as [127]

$$\mathcal{U}_{\mathbf{\Sigma}}^e = \left\{ \mathbf{\Sigma} | \left(\text{vec}(\mathbf{\Sigma}) - \text{vec}(\hat{\mathbf{\Sigma}}) \right)^T \mathbf{S}_{\mathbf{\Sigma}}^{-1} \left(\text{vec}(\mathbf{\Sigma}) - \text{vec}(\hat{\mathbf{\Sigma}}) \right) \leq \delta_{\mathbf{\Sigma}}^2, \ \mathbf{\Sigma} \succeq \mathbf{0} \right\}, \tag{6.12}$$

where the predefined parameters $\hat{\mathbf{\Sigma}} \succeq \mathbf{0}$, $\delta_{\mathbf{\Sigma}} > 0$, and $\mathbf{S}_{\mathbf{\Sigma}} \succ \mathbf{0}$ denote the location, size, and the shape of the uncertainty set.

To proceed, we consider a reformulation of (6.2) as follows:

$$\begin{array}{ll} \underset{\mathbf{w}, \mathbf{X}}{\text{maximize}} & \min\limits_{\boldsymbol{\mu} \in \mathcal{U}_{\boldsymbol{\mu}}} \mathbf{w}^T \boldsymbol{\mu} - \lambda \max\limits_{\mathbf{\Sigma} \in \mathcal{U}_{\mathbf{\Sigma}}} \text{Tr}(\mathbf{X}\mathbf{\Sigma}) \\ \text{subject to} & \mathbf{w}^T \mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \\ & \begin{bmatrix} \mathbf{X} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq \mathbf{0}. \end{array} \tag{6.13}$$

Since $\mathbf{\Sigma} \succeq \mathbf{0}$, the last constraint implies $\mathbf{X} \succeq \mathbf{w}\mathbf{w}^T$ which in turn is satisfied with equality at an optimal solution and thus (6.13) is equal to (6.2) and an optimal portfolio $\mathbf{w}$ of (6.13) is also optimal for (6.2). The advantage of (6.13) over (6.2) is that it allows us to derive a final equivalent convex problem.

Similar to (6.10), the inner worst-case variance in (6.13) over the elliptical uncertainty set is given by the following problem:

$$\begin{array}{ll} \underset{\mathbf{\Sigma}}{\text{maximize}} & \text{Tr}(\mathbf{X}\mathbf{\Sigma}) \\ \text{subject to} & \left(\text{vec}(\mathbf{\Sigma}) - \text{vec}(\hat{\mathbf{\Sigma}}) \right)^T \mathbf{S}_{\mathbf{\Sigma}}^{-1} \left(\text{vec}(\mathbf{\Sigma}) - \text{vec}(\hat{\mathbf{\Sigma}}) \right) \leq \delta_{\mathbf{\Sigma}}^2, \\ & \mathbf{\Sigma} \succeq \mathbf{0}. \end{array} \tag{6.14}$$

Problem (6.14) is convex and equals its dual problem:

$$\begin{array}{ll} \underset{\mathbf{Z}}{\text{minimize}} & \operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}+\mathbf{Z}\right)\right)+\delta_{\boldsymbol{\Sigma}}\left\|\mathbf{S}_{\boldsymbol{\Sigma}}^{1/2}\left(\operatorname{vec}(\mathbf{X})+\operatorname{vec}(\mathbf{Z})\right)\right\|_2 \\ \text{subject to} & \mathbf{Z}\succeq\mathbf{0}. \end{array} \tag{6.15}$$

Note that the objective is jointly convex in $\mathbf{X}$ and $\mathbf{Z}$, and the formulations III and IV in Table 6.1 are the resulting convex problems over the elliptical uncertainty set (6.12).

6.1.4 Worst-Case Variance Based on Factor Model

Instead of incorporating the uncertainty into the covariance matrix directly, it is may be more accurate to explore the structure of the covariance matrix and thus the uncertainty can be incorporated in a more proper way. Recall from Chapter 2 that one example of the financial time series modeling is the explicit factor model:

$$\mathbf{r}_t=\boldsymbol{\mu}+\boldsymbol{\Pi}^T\mathbf{f}_t+\mathbf{w}_t. \tag{6.16}$$

Here, for simplicity we assume $\boldsymbol{\mu}\in\mathbb{R}^N$ is the vector of mean returns, $\mathbf{f}\sim\mathcal{N}\left(\mathbf{0},\mathbf{F}\right)\in\mathbb{R}^K$ is the vector of returns of the factors that drive the market, $\boldsymbol{\Pi}\in\mathbb{R}^{K\times N}$ is the matrix of factor loadings, $\mathbf{w}_t\sim\mathcal{N}\left(\mathbf{0},\mathbf{D}\right)$ is the residual noise, and $\mathbf{D}$ is diagonal, i.e., $\mathbf{D}=\operatorname{Diag}(\mathbf{d})$. Then the covariance has the following structure:

$$\boldsymbol{\Sigma}=\boldsymbol{\Pi}^T\mathbf{F}\boldsymbol{\Pi}+\mathbf{D}. \tag{6.17}$$

For this structure (6.17), we assume $\mathbf{F}$ is known exactly and $\boldsymbol{\Pi}$ and $\mathbf{D}$ contain some estimation errors.

Similar to the previous cases, we assume the uncertainty sets of $\boldsymbol{\Pi}$ and $\mathbf{D}$ are separable, convex, and compact, and they are denoted as $\mathcal{U}_{\boldsymbol{\Pi}}$ and $\mathcal{U}_{\mathbf{D}}$, respectively. Now, the worst-case variance turns out to be

$$\begin{aligned} \max_{\boldsymbol{\Pi}\in\mathcal{U}_{\boldsymbol{\Pi}},\,\mathbf{D}\in\mathcal{U}_{\mathbf{D}}}\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} &= \max_{\boldsymbol{\Pi}\in\mathcal{U}_{\boldsymbol{\Pi}},\,\mathbf{D}\in\mathcal{U}_{\mathbf{D}}}\mathbf{w}^T\left(\boldsymbol{\Pi}^T\mathbf{F}\boldsymbol{\Pi}+\mathbf{D}\right)\mathbf{w} \\ &= \max_{\boldsymbol{\Pi}\in\mathcal{U}_{\boldsymbol{\Pi}}}\mathbf{w}^T\boldsymbol{\Pi}^T\mathbf{F}\boldsymbol{\Pi}\mathbf{w}+\max_{\mathbf{D}\in\mathcal{U}_{\mathbf{D}}}\mathbf{w}^T\mathbf{D}\mathbf{w}. \end{aligned} \tag{6.18}$$

Now, the expression in (6.18) is not concave in the uncertainty parameters any more and the results in Section 6.1.1 cannot be used. Here

we can consider the worst-case terms in (6.18) one by one and the goal is to find the worst-case variance either in a closed-form or given by an efficiently solvable convex problem.

Let us start with the second one which is simpler. Since $\mathbf{D}$ is the covariance for the residual noise and is assumed to be diagonal, the following uncertainty set is considered in practice [86]

$$\mathcal{U}_{\mathbf{D}} = \{\mathbf{D} | \mathbf{D} = \text{Diag}(\mathbf{d}),\ \underline{\mathbf{d}} \leq \mathbf{d} \leq \overline{\mathbf{d}}\}. \tag{6.19}$$

Denoting $\overline{\mathbf{D}} = \text{Diag}(\overline{\mathbf{d}})$, we have

$$\max_{\mathbf{D} \in \mathcal{U}_{\mathbf{D}}} \mathbf{w}^T \mathbf{D} \mathbf{w} = \mathbf{w}^T \overline{\mathbf{D}} \mathbf{w}. \tag{6.20}$$

For the first worst-case term in (6.18), i.e., $\max_{\mathbf{\Pi} \in \mathcal{U}_{\mathbf{\Pi}}} \mathbf{w}^T \mathbf{\Pi}^T \mathbf{F} \mathbf{\Pi} \mathbf{w}$, note that the objective is convex in $\mathbf{\Pi}$; however, the goal is to maximize the objective and thus it is nonconvex. In general, it is not easy to compute the worst-case value efficiently.

In the following, we will review some uncertainty sets so that the worst-case value $\max_{\mathbf{\Pi} \in \mathcal{U}_{\mathbf{\Pi}}} \mathbf{w}^T \mathbf{\Pi}^T \mathbf{F} \mathbf{\Pi} \mathbf{w}$ can be either computed in a closed-form or given by solving a convex problem.

Sphere Uncertainty Set

The uncertainty set of $\mathbf{\Pi}$ is assumed to be a sphere[2] and is given by

$$\mathcal{U}^s_{\mathbf{\Pi}} = \{\mathbf{\Pi} | \mathbf{\Pi} = \hat{\mathbf{\Pi}} + \mathbf{\Delta},\ \|\mathbf{\Delta}\|_F \leq \delta_{\mathbf{\Pi}}\}. \tag{6.21}$$

Without loss of generality and for simplicity, we set $\mathbf{F} = \mathbf{I}$ so that

$$\max_{\mathbf{\Pi} \in \mathcal{U}^s_{\mathbf{\Pi}}} \sqrt{\mathbf{w}^T \mathbf{\Pi}^T \mathbf{\Pi} \mathbf{w}} = \max_{\mathbf{\Pi} \in \mathcal{U}^s_{\mathbf{\Pi}}} \|\mathbf{\Pi} \mathbf{w}\|_2, \tag{6.22}$$

which is the square root of $\max_{\mathbf{\Pi} \in \mathcal{U}_{\mathbf{\Pi}}} \mathbf{w}^T \mathbf{\Pi}^T \mathbf{\Pi} \mathbf{w}$. One can upper bound the worst-case value in (6.22) as follows [54]:

$$\begin{aligned} \max_{\mathbf{\Pi} \in \mathcal{U}^s_{\mathbf{\Pi}}} \|\mathbf{\Pi} \mathbf{w}\|_2 &= \max_{\|\mathbf{\Delta}\|_F \leq \delta_{\mathbf{\Pi}}} \left\| \hat{\mathbf{\Pi}} \mathbf{w} + \mathbf{\Delta} \mathbf{w} \right\|_2 \\ &\leq \left\| \hat{\mathbf{\Pi}} \mathbf{w} \right\|_2 + \max_{\|\mathbf{\Delta}\|_F \leq \delta_{\mathbf{\Pi}}} \|\mathbf{\Delta} \mathbf{w}\|_2 \end{aligned}$$

[2]This can be easily extended to an elliptical uncertainty set.

$$\leq \left\|\hat{\mathbf{\Pi}}\mathbf{w}\right\|_2 + \max_{\|\mathbf{\Delta}\|_F \leq \delta_{\mathbf{\Pi}}} \|\mathbf{\Delta}\|_F \|\mathbf{w}\|_2$$
$$= \left\|\hat{\mathbf{\Pi}}\mathbf{w}\right\|_2 + \delta_{\mathbf{\Pi}} \|\mathbf{w}\|_2 . \tag{6.23}$$

In fact, this upper bound is achievable by $\mathbf{\Delta} = \delta_{\mathbf{\Pi}} \mathbf{u} \frac{\mathbf{w}^T}{\|\mathbf{w}\|_2}$ where

$$\mathbf{u} = \begin{cases} \frac{\hat{\mathbf{\Pi}}\mathbf{w}}{\left\|\hat{\mathbf{\Pi}}^T\mathbf{w}\right\|_2}, & \text{if } \hat{\mathbf{\Pi}}\mathbf{w} \neq \mathbf{0}, \\ \text{any unitary vector}, & \text{otherwise.} \end{cases} \tag{6.24}$$

The Problem V in Table 6.1 shows the equivalent convex formulation when only worst-case variance is considered and the uncertainty sets are $\mathcal{U}_{\mathbf{D}}$ in (6.19) and $\mathcal{U}^s_{\mathbf{\Pi}}$ in (6.21). Similar to Problems I-IV in Table 6.1, it is easy to combine the worst-case means over different uncertainty sets to get more equivalent convex formulations. They are quite straightforward and thus are omitted.

Column-Wise Elliptical Uncertainty Set

Another type of uncertainty is the column-wise elliptical uncertainty set [86]

$$\mathcal{U}^{ce}_{\mathbf{\Pi}} = \{\mathbf{\Pi} = \hat{\mathbf{\Pi}} + \mathbf{\Delta},\ \|\mathbf{\Delta}_i\|_g \leq \delta_{\mathbf{\Pi},i},\ i = 1, \ldots, N\}, \tag{6.25}$$

where $\mathbf{\Delta}_i$ is the i-th column of $\mathbf{\Delta}$, $\|\mathbf{x}\|_g = \sqrt{\mathbf{x}^T\mathbf{G}\mathbf{x}}$ and $\mathbf{G}$ is a given positive definite weight matrix, and $\boldsymbol{\delta}_{\mathbf{\Pi}} = [\delta_{\mathbf{\Pi},1}, \ldots, \delta_{\mathbf{\Pi},N}]^T$ represent the sizes of the elliptical uncertainty sets.

Even though the worst-case value $\max_{\mathbf{\Pi}\in\mathcal{U}_{\mathbf{\Pi}}} \mathbf{w}^T\mathbf{\Pi}^T\mathbf{F}\mathbf{\Pi}\mathbf{w}$ indeed is a non-convex problem, it is shown in [86] that the following inequality

$$\max_{\mathbf{\Pi}\in\mathcal{U}_{\mathbf{\Pi}}} \mathbf{w}^T\mathbf{\Pi}^T\mathbf{F}\mathbf{\Pi}\mathbf{w} \leq v \tag{6.26}$$

holds if and only if there exist $\sigma > 0$, $\tau \geq 0$, and $\mathbf{t} \geq \mathbf{0} \in \mathbb{R}^K$ that satisfy the following convex constraints:

$$\tau + \mathbf{1}^T\mathbf{t} \leq v, \tag{6.27}$$
$$|\mathbf{w}|^T\boldsymbol{\delta}_{\mathbf{\Pi}} \leq r, \tag{6.28}$$
$$\sigma \leq \frac{1}{\lambda_{\max}(\mathbf{H})}, \tag{6.29}$$

$$\left\| \begin{bmatrix} 2r \\ \sigma - \tau \end{bmatrix} \right\|_2 \leq \sigma + \tau, \tag{6.30}$$

$$\left\| \begin{bmatrix} 2s_i \\ 1 - \sigma\lambda_i - t_i \end{bmatrix} \right\|_2 \leq 1 - \sigma\lambda_i + t_i, \quad i = 1, \ldots, K, \tag{6.31}$$

where $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the spectral decomposition of $\mathbf{H} = \mathbf{G}^{-1/2}\mathbf{F}\mathbf{G}^{-1/2}$, $\mathbf{\Lambda} = \text{Diag}([\lambda_1, \ldots, \lambda_K])$ and $\mathbf{s} = \mathbf{U}^T\mathbf{H}^{1/2}\mathbf{G}^{1/2}\hat{\mathbf{\Pi}}\mathbf{w}$.

Problem VI in Table 6.1 presents the resulted equivalent convex formulation when the uncertainty sets are $\mathcal{U}_{\mathbf{D}}$ in (6.19) and $\mathcal{U}_{\mathbf{\Pi}}^{ce}$ in (6.25). Again, we omit the cases of considering worst-case mean and worst-case variance together since the derivations of equivalent convex formulations can be obtained straightforwardly based on the previous derivations.

6.1.5 Summary of Different Equivalent Formulations

Table 6.1 summarizes all the previously reviewed cases and, as mentioned before, straightforwardly, we can have many more different equivalent convex formulations for different combinations of the uncertainty sets of the mean vector and variance matrix.

6.2 Robust Sharpe ratio Optimization

Let us first recall the convex reformulation of Sharpe ratio maximization with only the capital budget constraint $\mathbf{w}^T\mathbf{1} = 1$, i.e., (5.14) as follows:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \\ \text{subject to} & \mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1}) = 1, \\ & \mathbf{w}^T\mathbf{1} > 0. \end{array} \tag{6.32}$$

Actually, the equality constraint $\mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1}) = 1$ in (6.32) can be relaxed as the inequality $\mathbf{w}^T(\boldsymbol{\mu} - r_f\mathbf{1}) \geq 1$ since optimality is always achieved at the equality. Then the robust Sharpe ratio problem can be

Table 6.1: Different robustifications of the problem (6.2).

$$\underset{\mathbf{w}}{\text{maximize}} \quad \min_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}} \mathbf{w}^T\boldsymbol{\mu} - \lambda \max_{\boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$$
$$\text{subject to} \quad \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}.$$

	Uncertainty Sets	Equivalent Convex Formulations
I	$\mathcal{U}^b_{\boldsymbol{\mu}} = (6.4)$ $\mathcal{U}^b_{\boldsymbol{\Sigma}} = (6.8)$	$\underset{\mathbf{w},\overline{\boldsymbol{\Lambda}},\underline{\boldsymbol{\Lambda}}}{\text{maximize}} \quad \mathbf{w}^T\hat{\boldsymbol{\mu}} - \lvert\mathbf{w}\rvert^T\boldsymbol{\delta}$ $- \lambda\left(\text{Tr}(\overline{\boldsymbol{\Lambda}}\boldsymbol{\Sigma}) - \text{Tr}(\underline{\boldsymbol{\Lambda}}\boldsymbol{\Sigma})\right)$ $\text{subject to} \quad \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W},$ $\begin{bmatrix} \overline{\boldsymbol{\Lambda}} - \underline{\boldsymbol{\Lambda}} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq \mathbf{0},$ $\overline{\boldsymbol{\Lambda}} \geq \mathbf{0}, \quad \underline{\boldsymbol{\Lambda}} \geq \mathbf{0}.$
II	$\mathcal{U}^e_{\boldsymbol{\mu}} = (6.6)$ $\mathcal{U}^b_{\boldsymbol{\Sigma}} = (6.8)$	$\underset{\mathbf{w},\overline{\boldsymbol{\Lambda}},\underline{\boldsymbol{\Lambda}}}{\text{maximize}} \quad \mathbf{w}^T\hat{\boldsymbol{\mu}} - \delta_{\boldsymbol{\mu}}\left\lVert\mathbf{S}^{1/2}\mathbf{w}\right\rVert_2$ $- \lambda\left(\text{Tr}(\overline{\boldsymbol{\Lambda}}\boldsymbol{\Sigma}) - \text{Tr}(\underline{\boldsymbol{\Lambda}}\boldsymbol{\Sigma})\right)$ $\text{subject to} \quad \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W},$ $\begin{bmatrix} \overline{\boldsymbol{\Lambda}} - \underline{\boldsymbol{\Lambda}} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq \mathbf{0},$ $\overline{\boldsymbol{\Lambda}} \geq \mathbf{0}, \quad \underline{\boldsymbol{\Lambda}} \geq \mathbf{0}.$
III	$\mathcal{U}^b_{\boldsymbol{\mu}} = (6.4)$ $\mathcal{U}^e_{\boldsymbol{\Sigma}} = (6.12)$	$\underset{\mathbf{w},\mathbf{X},\mathbf{Z}}{\text{maximize}} \quad \mathbf{w}^T\hat{\boldsymbol{\mu}} - \lvert\mathbf{w}\rvert^T\boldsymbol{\delta} - \lambda\text{Tr}\left(\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}+\mathbf{Z}\right)\right)$ $- \lambda\delta_{\boldsymbol{\Sigma}}\left\lVert\mathbf{S}_{\boldsymbol{\Sigma}}^{1/2}\left(\text{vec}(\mathbf{X}) + \text{vec}(\mathbf{Z})\right)\right\rVert_2$ $\text{subject to} \quad \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W},$ $\begin{bmatrix} \mathbf{X} & \mathbf{w} \\ \mathbf{w}^T & 1 \end{bmatrix} \succeq \mathbf{0}, \quad \mathbf{Z} \succeq \mathbf{0}.$

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \mathbf{w}^T\boldsymbol{\mu} - \lambda \max\limits_{\boldsymbol{\Pi}\in\mathcal{U}_{\boldsymbol{\Pi}},\, \mathbf{D}\in\mathcal{U}_{\mathbf{D}}} \mathbf{w}^T(\boldsymbol{\Pi}^T\mathbf{F}\boldsymbol{\Pi}+\mathbf{D})\mathbf{w} \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W}. \end{array}$$

	Uncertainty Sets	Equivalent Convex Formulations
IV	$\mathcal{U}^e_{\boldsymbol{\mu}} = (6.6)$ $\mathcal{U}^e_{\boldsymbol{\Sigma}} = (6.12)$	$\begin{array}{ll} \underset{\mathbf{w},\,\mathbf{X},\,\mathbf{Z}}{\text{maximize}} & \mathbf{w}^T\hat{\boldsymbol{\mu}} - \delta_{\boldsymbol{\mu}}\left\|\mathbf{S}^{1/2}\mathbf{w}\right\|_2 \\ & \quad - \lambda\mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}\left(\mathbf{X}+\mathbf{Z}\right)\right) \\ & \quad - \lambda\delta_{\boldsymbol{\Sigma}}\left\|\mathbf{S}_{\boldsymbol{\Sigma}}^{1/2}\left(\mathrm{vec}(\mathbf{X})+\mathrm{vec}(\mathbf{Z})\right)\right\|_2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W}, \\ & \begin{bmatrix}\mathbf{X} & \mathbf{w}\\ \mathbf{w}^T & 1\end{bmatrix} \succeq \mathbf{0}, \quad \mathbf{Z}\succeq\mathbf{0}. \end{array}$
V	$\mathcal{U}_{\mathbf{D}} = (6.19)$ $\mathcal{U}^s_{\boldsymbol{\Pi}} = (6.21)$	$\begin{array}{ll} \underset{\mathbf{w},\,y}{\text{maximize}} & \mathbf{w}^T\boldsymbol{\mu} - \lambda\left(\mathbf{w}^T\overline{\mathbf{D}}\mathbf{w} + y^2\right) \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W}, \\ & \left\|\hat{\boldsymbol{\Pi}}^T\mathbf{w}\right\|_2 + \delta_{\boldsymbol{\Pi}}\left\|\mathbf{w}\right\|_2 \leq y. \end{array}$
VI	$\mathcal{U}_{\mathbf{D}} = (6.19)$ $\mathcal{U}^{ce}_{\boldsymbol{\Pi}} = (6.25)$	$\begin{array}{ll} \underset{\mathbf{w},\,v,\,\sigma,\,\tau,\,\mathbf{t},\,\mathbf{s}}{\text{maximize}} & \mathbf{w}^T\boldsymbol{\mu} - \lambda\left(\mathbf{w}^T\overline{\mathbf{D}}\mathbf{w} + v\right) \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w}\in\mathcal{W}, \\ & \tau + \mathbf{1}^T\mathbf{t} \leq v, \quad \mathbf{t}\geq\mathbf{0}, \\ & \left\|\mathbf{w}\right\|^T\boldsymbol{\delta}_{\boldsymbol{\Pi}} \leq r \\ & \sigma \leq \dfrac{1}{\lambda_{\max}\left(\mathbf{H}\right)}, \\ & \left\|\begin{bmatrix}2r\\ \sigma-\tau\end{bmatrix}\right\|_2 \leq \sigma+\tau, \\ & \mathbf{s} = \mathbf{U}^T\mathbf{H}^{1/2}\mathbf{G}^{1/2}\hat{\boldsymbol{\Pi}}\mathbf{w}, \\ & \left\|\begin{bmatrix}2s_i\\ 1-\sigma\lambda_i-t_i\end{bmatrix}\right\|_2 \leq 1-\sigma\lambda_i+t_i, \\ & \qquad i = 1,\ldots,K. \end{array}$

formulated based on (6.32) as follows:

$$\begin{array}{ll} \underset{\mathbf{w},\,\kappa}{\text{minimize}} & \max\limits_{\boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \\ \text{subject to} & \min\limits_{\boldsymbol{\mu}\in\mathcal{U}_{\boldsymbol{\mu}}} \mathbf{w}^T(\boldsymbol{\mu}-r_f\mathbf{1}) \geq 1, \\ & \mathbf{w}^T\mathbf{1} > 0, \end{array} \tag{6.33}$$

where $\mathcal{U}_{\boldsymbol{\mu}}$ and $\mathcal{U}_{\boldsymbol{\Sigma}}$ denote some general uncertainty sets for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, and the robust techniques stated in the previous Section 6.1 can be directly used to obtain some equivalent convex formulations.

When there exist some other convex constraints apart from the capital budge constraint, the robust formulation is not to simply add them into (6.33) but becomes more complicated. The detailed derivation approach can be found in [200]. Nevertheless, for the derived robust formulation in [200], the robust techniques in Section 6.1 are still applicable.

6.3 Connections with Robust Beamforming

Let us first recall the receive beamforming problem (5.18):

$$\underset{\mathbf{w}}{\text{maximize}} \quad \frac{\sigma_s^2|\mathbf{w}^H\mathbf{a}|^2}{\mathbf{w}^H\mathbf{R}\mathbf{w}} \tag{6.34}$$

where $\mathbf{w} \in \mathbb{C}^N$ is the complex beamforming vector variable denoting the weights of N array observations and $\mathbf{a} \in \mathbb{C}^N$ and $\mathbf{R} \in \mathbb{C}^{N\times N}$ (estimated in advance) are the signal steering vector (also known as the transmission channel) and the positive definite interference-plus-noise covariance matrix, respectively.

Similar to the (real-valued) parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for portfolio design, the (complex-valued) parameters $\mathbf{a}$ and $\mathbf{R}$ need to be estimated first and may contain some estimation errors. Since the objective in (6.34) is invariant to the magnitude of $\mathbf{w}^H\mathbf{a}$, the robust counterpart of (6.34) has the following general form [204, 205]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \max\limits_{\mathbf{R}\in\mathcal{U}_{\mathbf{R}}} \mathbf{w}^H\mathbf{R}\mathbf{w} \\ \text{subject to} & \min\limits_{\mathbf{a}\in\mathcal{U}_{\mathbf{a}}} \left|\mathbf{w}^H\mathbf{a}\right| \geq 1, \end{array} \tag{6.35}$$

where $\mathcal{U}_{\mathbf{a}}$ and $\mathcal{U}_{\mathbf{R}}$ denote the uncertainty sets of $\mathbf{a}$ and $\mathbf{R}$, respectively.

6.3.1 Worst-Case Signal Power Constraint

In this subsection, we deal with the worst-case signal power constraint in (6.35), i.e., $\min_{\mathbf{a}\in\mathcal{U}_\mathbf{a}} \left|\mathbf{w}^H\mathbf{a}\right| \geq 1$.

The authors of [204] considered a sphere uncertainty set

$$\mathcal{U}_\mathbf{a}^s = \{\mathbf{a}|(\mathbf{a}-\hat{\mathbf{a}})^H(\mathbf{a}-\hat{\mathbf{a}}) \leq \delta_\mathbf{a}^2\}, \tag{6.36}$$

where the predefined parameters $\hat{\mathbf{a}}$ and (usually very small) $\delta_\mathbf{a} > 0$ define the location and size of the uncertainty set, respectively.

Denoting $\boldsymbol{\gamma} \triangleq \mathbf{a}-\hat{\mathbf{a}}$, we have $\|\boldsymbol{\gamma}\|_2 \leq \delta_\mathbf{a}$ and

$$\left|\mathbf{w}^H\mathbf{a}\right| = \left|\mathbf{w}^H(\hat{\mathbf{a}}+\boldsymbol{\gamma})\right| \geq \left|\mathbf{w}^H\hat{\mathbf{a}}\right| - \left|\mathbf{w}^H\boldsymbol{\gamma}\right| \geq \left|\mathbf{w}^H\hat{\mathbf{a}}\right| - \delta_\mathbf{a}\left\|\mathbf{w}\right\|_2. \tag{6.37}$$

It can be shown that if $\delta_\mathbf{a}$ is small enough such that $\left|\mathbf{w}^H\hat{\mathbf{a}}\right| > \delta_\mathbf{a}\left\|\mathbf{w}\right\|_2$ always holds, then the inequalities in (6.37) are achieved with equality by [204]

$$\boldsymbol{\gamma} = -\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\delta_\mathbf{a}e^{j\angle(\mathbf{w}^H\hat{\mathbf{a}})}. \tag{6.38}$$

That is to say,

$$\min_{\mathbf{a}\in\mathcal{U}_\mathbf{a}} \left|\mathbf{w}^H\mathbf{a}\right| = \left|\mathbf{w}^H\hat{\mathbf{a}}\right| - \delta_\mathbf{a}\left\|\mathbf{w}\right\|_2. \tag{6.39}$$

However, then the worst-case signal power constraint $\min_{\mathbf{a}\in\mathcal{U}_\mathbf{a}} \left|\mathbf{w}^H\mathbf{a}\right| \geq 1$ turns out to be

$$\left|\mathbf{w}^H\hat{\mathbf{a}}\right| - \delta_\mathbf{a}\left\|\mathbf{w}\right\|_2 \geq 1, \tag{6.40}$$

which is still nonconvex.

Fortunately, the objective of (6.35) is unchanged under any arbitrary phase rotation of $\mathbf{w}$, and one can always rotate $\mathbf{w}$ properly so that $\mathbf{w}^H\hat{\mathbf{a}}$ is real and positive. That is, (6.40) can be further equivalently reformulated as the following convex constraints:

$$\mathbf{w}^H\hat{\mathbf{a}} - \delta_\mathbf{a}\left\|\mathbf{w}\right\|_2 \geq 1, \tag{6.41}$$

$$\mathrm{Im}\{\mathbf{w}^H\hat{\mathbf{a}}\} = 0. \tag{6.42}$$

Interestingly, we can see that the derivations for the (complex-valued) worst-case signal power here are very similar to that for the

(real-valued) worst-case mean under the elliptical uncertainty set in Section 6.1.2. For example, the (complex-valued) worst-case signal power $\mathbf{w}^H\hat{\mathbf{a}} - \delta_{\mathbf{a}} \|\mathbf{w}\|_2$ in (6.41) looks the same as the (real-valued) worst-case mean $\mathbf{w}^T\hat{\boldsymbol{\mu}} - \delta_{\boldsymbol{\mu}} \left\|\mathbf{S}_{\boldsymbol{\mu}}^{1/2}\mathbf{w}\right\|_2$ in (6.7) with $\mathbf{S}_{\boldsymbol{\mu}} = \mathbf{I}$.

6.3.2 Worst-Case Interference-Plus-Noise Power

Now let us consider the worst-case interference-plus-noise power in (6.35), i.e., $\max_{\mathbf{R}\in\mathcal{U}_{\mathbf{R}}} \mathbf{w}^H\mathbf{R}\mathbf{w}$.

The authors of [205] considered replacing the interference-plus-noise covariance matrix with the SCM:

$$\mathbf{R}_{\text{SCM}} = \frac{1}{T}\hat{\mathbf{X}}^H\hat{\mathbf{X}}, \tag{6.43}$$

where $\hat{\mathbf{X}} \in \mathbb{C}^{T\times N}$ is the observation matrix such that the t-th row of $\hat{\mathbf{X}}$ is the transpose of the t-th observation $\mathbf{x}(t)$, and T is the total number of observations. Then they considered a spherical uncertainty set for the underlying true observations $\mathbf{X}$ as follows:

$$\mathcal{U}_{\mathbf{X}}^s = \{\mathbf{X} | \mathbf{X} = \hat{\mathbf{X}} + \boldsymbol{\Delta},\ \|\boldsymbol{\Delta}\|_F \le \delta_{\mathbf{X}}\}. \tag{6.44}$$

Instead of studying the worst-case interference-plus-noise power, one can study its square root value

$$\max_{\mathbf{X}\in\mathcal{U}_{\mathbf{X}}^s} \sqrt{\mathbf{w}^H\mathbf{X}^H\mathbf{X}\mathbf{w}} = \max_{\mathbf{X}\in\mathcal{U}_{\mathbf{X}}^s} \|\mathbf{X}\mathbf{w}\|_2 . \tag{6.45}$$

Given the uncertainty set (6.44), the worst-case value admits a closed-form expression [205]

$$\max_{\mathbf{X}\in\mathcal{U}_{\mathbf{X}}^s} \|\mathbf{X}\mathbf{w}\|_2 = \left\|\hat{\mathbf{X}}\mathbf{w}\right\|_2 + \delta_{\mathbf{X}} \|\mathbf{w}\|_2 . \tag{6.46}$$

Actually, the derivation procedure of (6.46) is exactly the same as that of (6.23) for worst-case portfolio variance and thus it is omitted.

6.3.3 Robust Beamforming Formulation

Finally we can see that with the uncertainties are considered in (6.36) and (6.44), the worst-case robust problem formulation (6.35) can be

reformulated in a convex form as follows:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \left\|\hat{\mathbf{X}}\mathbf{w}\right\|_2 + \delta_{\mathbf{X}} \left\|\mathbf{w}\right\|_2 \\ \text{subject to} & \mathbf{w}^H\hat{\mathbf{a}} - \delta_{\mathbf{a}} \left\|\mathbf{w}\right\|_2 \geq 1, \\ & \mathrm{Im}\{\mathbf{w}^H\hat{\mathbf{a}}\} = 0. \end{array} \tag{6.47}$$

Thus, it is interesting to see that both robust portfolio optimization and robust beamforming can be dealt with using almost the same techniques.

7

Multi-Portfolio Optimization

Portfolio managers usually manage multiple accounts corresponding to different clients, and the portfolios associated with different accounts are pooled together for execution, amplifying the level of the so-called market impact (cf. Chapter 4) on all accounts. In the previous Chapters 5 and 6, each portfolio is considered and optimized individually disregarding the effect or impact on other portfolio, however, if this aggregate market effect is not considered when each account is individually optimized, the actual market impact can be severely underestimated.

Thus, a more realistic way is to analyze and optimize the multiple portfolios jointly while adhering to both the account-specific constraints and also some global constraints present on all accounts. The holistic approach is termed multi-portfolio optimization.

The detailed organization of this chapter is as follows. Section 7.1 reviews some basic concepts and definitions. Section 7.2 states some typical problem formulations and Section 7.3 presents a solving approach based on game theory.

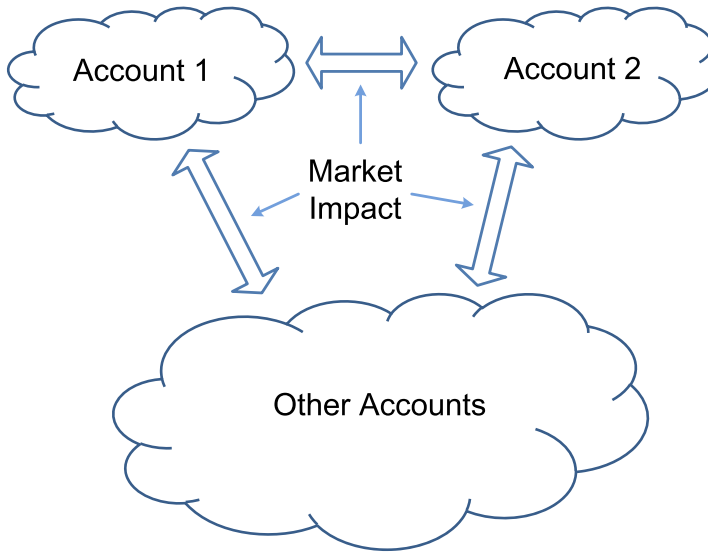

Figure 7.1: Multiple accounts and market impact.

7.1 From Single-Portfolio to Multi-Portfolio

In the real markets, portfolio managers always manage multiple accounts and each account is in fact effected by all the others as shown in Figure 7.1. In practice, such an impact usually is undesired, e.g., the impact on account 1 given by account 2 and the other accounts always tends to weaken the profitability of account 1, and it is referred to as market impact.

Suppose there are N assets with mean vector and covariance matrix given by $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$. Now we consider multiple, say K, accounts, and their corresponding investment portfolios are denoted as $\mathbf{w}_k \in \mathbb{R}^N$, $k = 1, \ldots, K$. So now we have multiple portfolios to optimize at the sample instead of only a single portfolio. In the following we will first quantify the market impact and then consider the utility function and different types of constraints.

7.1.1 Market Impact Cost Function

A key concept of the extension from single-portfolio to multi-portfolio is to understand how one account will be affected by the other accounts, i.e., the market impact among portfolios of different accounts.

Recall there are K portfolios: $\mathbf{w}_k \in \mathbb{R}^N$, $k = 1, \ldots, K$. Let us denote $\mathbf{w}_{-k} \triangleq (\mathbf{w}_l)_{l \neq k}$ and $\mathbf{w} \triangleq (\mathbf{w}_k)_{k=1}^K$ as the other portfolios (i.e., all the portfolios except portfolio k) and all the portfolios, respectively. For simplicity, $\langle \mathbf{x}, \mathbf{y} \rangle$ and $\mathbf{x}^T\mathbf{y}$ are used interchangeably to denote the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$.

A popular market impact on the portfolio $\mathbf{w}_k$ caused by itself and the other ones $\mathbf{w}_{-k}$ is [210]

$$\mathrm{TC}(\mathbf{w}_k, \mathbf{w}) \triangleq \frac{1}{2}\left(\left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^+, \mathbf{c}^+(\mathbf{w}) \right\rangle + \left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^-, \mathbf{c}^-(\mathbf{w}) \right\rangle\right), \tag{7.1}$$

where

$$[\mathbf{w}_k - \mathbf{w}_k^0]^+ \triangleq \max(\mathbf{0}, \mathbf{w}_k - \mathbf{w}_k^0) \tag{7.2}$$

$$[\mathbf{w}_k - \mathbf{w}_k^0]^- \triangleq \max(\mathbf{0}, -(\mathbf{w}_k - \mathbf{w}_k^0)) \tag{7.3}$$

represent the buy and sell trades of the k-th account, respectively, and

$$\mathbf{c}^+(\mathbf{w}) \triangleq \mathbf{\Omega}^+ \sum_{l=1}^K [\mathbf{w}_l - \mathbf{w}_l^0]^+ \tag{7.4}$$

$$\mathbf{c}^-(\mathbf{w}) \triangleq \mathbf{\Omega}^- \sum_{l=1}^K [\mathbf{w}_l - \mathbf{w}_l^0]^- \tag{7.5}$$

are the linear market impact costs of buy and sell trades of all the accounts, respectively, [8, 18]. Here, $\mathbf{\Omega}^+$ and $\mathbf{\Omega}^-$ are positive diagonal matrices representing the market impact of buy and sell trades, respectively.

7.1.2 Mean-Variance Utility Function

Instead of ignoring the market impact in single-portfolio optimization (cf. Chapter 5), the utility function of each portfolio is composed of both the mean-variance trade-off (cf. Section 5.1.1) and the market

impact cost in multi-portfolio optimization, that is, the utility function of account k is

$$u_k(\mathbf{w}_k, \mathbf{w}_{-k}) \triangleq \boldsymbol{\mu}^T \mathbf{w}_k - \frac{1}{2}\rho_k \mathbf{w}_k^T \boldsymbol{\Sigma} \mathbf{w}_k - \mathrm{TC}(\mathbf{w}_k, \mathbf{w}), \tag{7.6}$$

where the first two terms together are the mean-variance trade-off that depends on the portfolio of account k (i.e., $\mathbf{w}_k$) only with $\rho_k > 0$ being the trade-off parameter, and the third term $\mathrm{TC}(\mathbf{w}_k, \mathbf{w})$ as defined in (7.1) is the market impact cost function that measures the impact quantitatively among the portfolios of all accounts.

Substituting (7.1)-(7.5) into (7.6), the utility function $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ can be rewritten more explicitly as follows:

$$\begin{aligned} u_k(\mathbf{w}_k, \mathbf{w}_{-k}) = \boldsymbol{\mu}^T \mathbf{w}_k &- \frac{1}{2}\rho_k \mathbf{w}_k^T \boldsymbol{\Sigma} \mathbf{w}_k \\ &- \frac{1}{2}\left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^+, \boldsymbol{\Omega}^+ \sum_{l=1}^{K} [\mathbf{w}_l - \mathbf{w}_l^0]^+ \right\rangle \\ &- \frac{1}{2}\left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^-, \boldsymbol{\Omega}^- \sum_{l=1}^{K} [\mathbf{w}_l - \mathbf{w}_l^0]^- \right\rangle. \end{aligned} \tag{7.7}$$

7.1.3 Individual and Global Constraints

For multi-portfolio optimization, there are two types of constraints: individual constraints that apply to each specific account and global constraints that apply to all (or a group of) accounts.

Individual Constraints

The individual constraints are similar to the constraints stated in Section 5.1.4, e.g., holding constraint $\mathbf{l}_k \leq \mathbf{w}_k \leq \mathbf{u}_k$, long-only constraint $\mathbf{w}_k \geq \mathbf{0}$, etc., for each account k where $k = 1, \ldots, K$, and they are referred to as individual constraints.

For the multi-portfolio optimization, since each account may have different capital budgets, the capital budget constraints can be mathematically represented as $\mathbf{1}^T \mathbf{w}_k \leq b_k$, where $b_k \geq 0$, $k = 1, \ldots, K$, are the capital budget bounds for the corresponding accounts.

For clarity of presentation, we use $\mathcal{W}_k$ to denote all the individual constraints on account k, and in general we assume it is non-empty,

closed, and convex. We further use $\mathcal{W} \triangleq \mathcal{W}_1 \times \cdots \times \mathcal{W}_K$ to denote their Cartesian product set.

Global Constraints

As to the global constraints, one example is that the total traded volume of each asset over all the accounts must be less than a threshold (e.g., 10% of the average daily trading volume). Mathematically, these global constraints on all the accounts are

$$\sum_{k=1}^{K} |w_{k,i} - w_{k,i}^0| \leq D_i, \quad i = 1, \ldots, N. \tag{7.8}$$

These constraints can be extended so that the traded volume of some groups of assets (e.g., industries, sectors, countries, asset classes, etc.) should be limited, that is,

$$\sum_{k=1}^{K} \sum_{i \in \mathcal{G}_l} |w_{k,i} - w_{k,i}^0| \leq U_i, \quad l = 1, \ldots, L, \tag{7.9}$$

where $\mathcal{G}_l$ denotes the l-th group of assets and there are L groups.

It is easy to see from (7.8) and (7.9) that one account's portfolio design, say $\mathbf{w}_k$, also depends on other accounts' actions $\mathbf{w}_{-k}$. Therefore, the presence of global constraints couple all the portfolios together.

The global constraints (7.8) and (7.9) can be rewritten in a more compact form. We first define a multivariate function

$$\mathbf{g}(\mathbf{w}) = \begin{bmatrix} \left(\sum_{k=1}^{K} |w_{k,i} - w_{k,i}^0| - D_i\right)_{i=1}^{N} \\ \left(\sum_{k=1}^{K} \sum_{i \in \mathcal{G}_l} |w_{k,i} - w_{k,i}^0| - U_l\right)_{l=1}^{L} \end{bmatrix}, \tag{7.10}$$

then (7.8) and (7.9) can be simply rewritten as $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ (or more often $\mathbf{g}(\mathbf{w}_k, \mathbf{w}_{-k}) \leq \mathbf{0}$ for the consistency of notation).

7.2 Multi-Portfolio Problems

For the multi-portfolio case, there exist many different formulations.

7.2.1 Naive Formulation

One of the most direct formulations is to ignore the market impact among different accounts, do not consider the global constraints, and simply optimize each account individually (but include the market impact of the individual account) as follows [175]:

$$\left.\begin{array}{ll} \underset{\mathbf{w}_k}{\text{maximize}} & \boldsymbol{\mu}^T\mathbf{w}_k - \frac{1}{2}\rho_k\mathbf{w}_k^T\boldsymbol{\Sigma}\mathbf{w}_k \\ & \quad - \frac{1}{2}\left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^+, \boldsymbol{\Omega}^+[\mathbf{w}_k - \mathbf{w}_k^0]^+ \right\rangle \\ & \quad - \frac{1}{2}\left\langle [\mathbf{w}_k - \mathbf{w}_k^0]^-, \boldsymbol{\Omega}^-[\mathbf{w}_k - \mathbf{w}_k^0]^- \right\rangle \\ \text{subject to} & \mathbf{w}_k \in \mathcal{W}_k \end{array}\right\} \forall k, \tag{7.11}$$

where the objective contains a mean-variance trade-off with $\rho_k > 0$ being the trade-off parameter and a market impact cost caused only by itself.

We can see both the objectives and the constraints of the problems in (7.11) depend on each individual account portfolio $\mathbf{w}_k$ and the problems can be optimized separately. In other words, (7.11) represents K different single-portfolio optimization problems.

7.2.2 Total Social Welfare Maximization Problem

When the market impact among different accounts is considered, all the accounts are coupled together. Then one direct formulation is simply maximize the summation of all the utilities of all the accounts, i.e., the total social welfare maximization problem [154]

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{maximize}} & \sum_{k=1}^{K} u_k(\mathbf{w}_k, \mathbf{w}_{-k}) \\ \text{subject to} & \mathbf{w} \in \mathcal{W}, \end{array} \tag{7.12}$$

where $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ are defined in (7.7), $\mathbf{w} = (\mathbf{w}_k)_{k=1}^K$, $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_K$, and $\mathbf{w}$ also needs to satisfy the global constraints $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ if they are present.

Even though the central problem (7.12) can achieve the maximum social welfare, it may not result in fair enough portfolios: smaller ac-

counts suffer from a shortage of liquidity and they are forced to sacrifice their own benefits to achieve social optimality [175, 173].

7.2.3 Game Theoretical Formulation Under Individual Constraints

A more fair formulation proposed in [210] is that each account competes against the others and chooses a portfolio that maximizes its own utility under individual constraints. Mathematically, it can be formulated as a Nash Equilibrium Problem (NEP): given the other strategies $\mathbf{w}_{-k}$, account k aims at solving

$$\left.\begin{array}{ll} \underset{\mathbf{w}_k}{\mathsf{maximize}} & u_k(\mathbf{w}_k, \mathbf{w}_{-k}) \\ \mathsf{subject\ to} & \mathbf{w}_k \in \mathcal{W}_k \end{array}\right\} \forall k, \tag{7.13}$$

where $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ is defined in (7.7).

Compared with the naive individual formulation in (7.11), the main difference is that the objectives in (7.13) depend on not only the portfolio of each account $\mathbf{w}_k$ but also the portfolios of the other accounts $\mathbf{w}_{-k}$. Thus all the problems given by (7.13) are coupled via their objectives.

With the NEP formulation, a solution of interest is the well-known notion of the Nash Equilibrium (NE) point from which no account has an incentive to deviate from unilaterally. That is, a solution $\mathbf{w}_{\text{ne}} = (\mathbf{w}_k^\star)_{k=1}^K$ is an NE of the NEP (7.13) if

$$u_k(\mathbf{w}_k^\star, \mathbf{w}_{-k}^\star) \geq u_k(\mathbf{w}_k, \mathbf{w}_{-k}^\star), \quad \forall \mathbf{w}_k \in \mathcal{W}_k, \quad \forall k. \tag{7.14}$$

7.2.4 Game Theoretical Formulation Under Global Constraints

When there are global constraints, incorporating them into (7.13) results in the following Generalized NEP (GNEP) [210]:

$$\left.\begin{array}{ll} \underset{\mathbf{w}_k}{\mathsf{maximize}} & u_k(\mathbf{w}_k, \mathbf{w}_{-k}) \\ \mathsf{subject\ to} & \mathbf{w}_k \in \mathcal{W}_k \\ & \mathbf{g}(\mathbf{w}_k, \mathbf{w}_{-k}) \leq \mathbf{0} \end{array}\right\} \forall k, \tag{7.15}$$

where there is coupling in both utility and constraint sets.

Similar to an NE of (7.13), a solution of interest of (7.15) is referred to as Generalized NE (GNE) such that $\mathbf{w}_{\text{gne}} = (\mathbf{w}_k^\star)_{k=1}^K$ and

$$u_k(\mathbf{w}_k^\star, \mathbf{w}_{-k}^\star) \geq u_k(\mathbf{w}_k, \mathbf{w}_{-k}^\star),\ \forall \mathbf{w}_k \in \mathcal{W}_k,\ \mathbf{g}(\mathbf{w}_k, \mathbf{w}_{-k}^\star) \leq \mathbf{0},\ \forall k. \tag{7.16}$$

The extra coupling in the constrain sets caused by the global constraints makes the GNEP (7.15) much more difficult to analyze than the NEP (7.13).

7.2.5 Difficulties

For the above naive and total social welfare maximization problems, i.e., (7.11) and (7.12), the main difficulty is that the objectives in general are nonconcave and nondifferentiable due to the projections in the utilities $[\cdot]^+$ and $[\cdot]^-$.

For the NEP (7.13) and GNEP (7.15), apart from the above nonconcave and nondifferentiable objectives, the coupling in the objectives and constraints sets of the multi-account problems further complicates the analysis.

7.3 Efficient Solving Methods

In this section, some reformulation techniques are considered to deal with the difficulty caused the projections $[\cdot]^+$ and $[\cdot]^-$ and then the efficient solving methods for all the problems, i.e., (7.11), (7.12), (7.13), and GNEP (7.15) are reviewed.

7.3.1 Reformulations of Objectives and Constraints

To deal with the projections $[\cdot]^+$ and $[\cdot]^-$, one can first introduce some new variables, $\forall k$,

$$\tilde{\mathbf{w}}_k \triangleq \begin{bmatrix} \tilde{\mathbf{w}}_k^+ \\ \tilde{\mathbf{w}}_k^- \end{bmatrix} \geq \mathbf{0} \tag{7.17}$$

such that

$$[\mathbf{w}_k - \mathbf{w}_k^0]^+ = \tilde{\mathbf{w}}_k^+, \tag{7.18}$$

$$[\mathbf{w}_k - \mathbf{w}_k^0]^- = \tilde{\mathbf{w}}_k^-, \tag{7.19}$$

$$\mathbf{w}_k - \mathbf{w}_k^0 = \tilde{\mathbf{w}}_k^+ - \tilde{\mathbf{w}}_k^-, \tag{7.20}$$

$$0 = \left\langle \tilde{\mathbf{w}}_k^+, \tilde{\mathbf{w}}_k^- \right\rangle. \tag{7.21}$$

Then the utility function in (7.7) can be rewritten as (some constants are added)

$$\begin{aligned}\tilde{u}(\tilde{\mathbf{w}}_k, \tilde{\mathbf{w}}_{-k}) = &\underbrace{\begin{bmatrix} \boldsymbol{\mu} - \rho_k \boldsymbol{\Sigma}\mathbf{w}_k^0 \\ -\boldsymbol{\mu} + \rho_k \boldsymbol{\Sigma}\mathbf{w}_k^0 \end{bmatrix}^T}_{\triangleq \tilde{\boldsymbol{\mu}}_k^T} \tilde{\mathbf{w}}_k - \frac{1}{2}\rho_k \tilde{\mathbf{w}}_k^T \underbrace{\begin{bmatrix} \boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \end{bmatrix}}_{\triangleq \tilde{\boldsymbol{\Sigma}}} \tilde{\mathbf{w}}_k \\ &- \frac{1}{2}\tilde{\mathbf{w}}_k^T \underbrace{\begin{bmatrix} \boldsymbol{\Omega}^+ & \\ & \boldsymbol{\Omega}^- \end{bmatrix}}_{\triangleq \tilde{\boldsymbol{\Omega}}} \left(\sum_{l=1}^K \tilde{\mathbf{w}}_l \right) \\ = &\tilde{\boldsymbol{\mu}}_k^T \tilde{\mathbf{w}}_k - \frac{1}{2}\rho_k \tilde{\mathbf{w}}_k^T \tilde{\boldsymbol{\Sigma}} \tilde{\mathbf{w}}_k - \frac{1}{2}\tilde{\mathbf{w}}_k^T \tilde{\boldsymbol{\Omega}} \left(\sum_{l=1}^K \tilde{\mathbf{w}}_l \right), \end{aligned} \tag{7.22}$$

which now is a differentiable function.

For the introduced variable $\tilde{\mathbf{w}}_k$, relaxing the nonconvex constraint (7.21), one can define the following individual set based on (7.17)-(7.20):

$$\widetilde{\mathcal{W}}_k \triangleq \left\{ \tilde{\mathbf{w}}_k \middle| \begin{bmatrix} \mathbf{I} & -\mathbf{I} \end{bmatrix} \tilde{\mathbf{w}}_k + \mathbf{w}_k^0 \in \mathcal{W}_k, \tilde{\mathbf{w}}_k \geq \mathbf{0} \right\}, \tag{7.23}$$

which is convex in $\tilde{\mathbf{w}}_k$.

7.3.2 Naive Solution

For each k, the objective of the naive formulation (7.11) can be obtained by ignoring the market impact terms caused by the other accounts in $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ (cf. (7.7)). Thus, similar to (7.22), a relaxation of the naive formulation (7.11) is

$$\left.\begin{array}{ll} \underset{\tilde{\mathbf{w}}_k}{\mathsf{maximize}} & \tilde{\boldsymbol{\mu}}_k^T \tilde{\mathbf{w}}_k - \frac{1}{2}\rho_k \tilde{\mathbf{w}}_k^T \tilde{\boldsymbol{\Sigma}} \tilde{\mathbf{w}}_k - \frac{1}{2}\tilde{\mathbf{w}}_k^T \tilde{\boldsymbol{\Omega}} \tilde{\mathbf{w}}_k \\ \mathsf{subject\ to} & \tilde{\mathbf{w}}_k \in \widetilde{\mathcal{W}}_k \end{array}\right\} \forall k, \tag{7.24}$$

which is convex since the objective is quadratic concave and the feasible is convex for each given k, and thus it is efficiently solvable.

Similar to (7.11), the relaxation (7.24) actually represents K individual convex problems and, fortunately, it is shown that each optimal $\tilde{\mathbf{w}}_k$ satisfies (7.21) for all k and thus the relaxation (7.24) is tight for and therefore equivalent to (7.11) [210]. An optimal solution of (7.24) is referred to as a naive solution.

7.3.3 Total Social Welfare Maximization

For the total welfare maximization problem (7.12), replacing $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ with $\tilde{u}(\tilde{\mathbf{w}}_k, \tilde{\mathbf{w}}_{-k})$ in (7.22) and rearranging the terms, one can have the following relaxation [210]:

$$\begin{array}{ll} \underset{\tilde{\mathbf{w}}}{\text{maximize}} & P_{\text{so}}(\tilde{\mathbf{w}}) \triangleq \tilde{\boldsymbol{\mu}}^T \tilde{\mathbf{w}} - \dfrac{1}{2} \tilde{\mathbf{w}}^T \mathbf{M}_{\text{so}} \tilde{\mathbf{w}} \\ \text{subject to} & \tilde{\mathbf{w}} \in \widetilde{\mathcal{W}}_1 \times \cdots \times \widetilde{\mathcal{W}}_K \end{array} \tag{7.25}$$

where $\tilde{\boldsymbol{\mu}} \triangleq (\tilde{\boldsymbol{\mu}}_k)_{k=1}^K$, $\tilde{\mathbf{w}} \triangleq (\tilde{\mathbf{w}}_k)_{k=1}^K$,

$$\mathbf{M}_{\text{so}} = \text{Diag}(\boldsymbol{\rho}) \otimes \tilde{\boldsymbol{\Sigma}} + \mathbf{J} \otimes \tilde{\boldsymbol{\Omega}} \tag{7.26}$$

and $\mathbf{J}$ is a $K \times K$ matrix with all entries being 1.

Again, it is shown in [210] that (7.25) is convex and the optimal $\tilde{\mathbf{w}}_k$ satisfies (7.21) for all k and thus the relaxation (7.25) is tight.

7.3.4 Multi-Portfolio Optimization with Individual Constraints

Replacing $u_k(\mathbf{w}_k, \mathbf{w}_{-k})$ in the NEP (7.13) with $\tilde{u}(\tilde{\mathbf{w}}_k, \tilde{\mathbf{w}}_{-k})$ results in the following relaxation NEP:

$$\left.\begin{array}{ll} \underset{\tilde{\mathbf{w}}_k}{\text{maximize}} & \tilde{u}(\tilde{\mathbf{w}}_k, \tilde{\mathbf{w}}_{-k}) \\ \text{subject to} & \tilde{\mathbf{w}}_k \in \widetilde{\mathcal{W}}_k \end{array}\right\} \forall k. \tag{7.27}$$

And it is shown that (7.21) is satisfied by an NE of (7.27), thus the NEP (7.27) indeed equals the NEP (7.13).

Since the constraint sets of $\tilde{\mathbf{w}}_k$ are decoupled, based on potential game theory [148], it is further shown in [210] that the NEP is equal to the following optimization problem:

$$\begin{array}{ll} \underset{\tilde{\mathbf{w}}}{\text{maximize}} & P_{\text{ne}}(\tilde{\mathbf{w}}) \triangleq \tilde{\boldsymbol{\mu}}^T \tilde{\mathbf{w}} - \dfrac{1}{2} \tilde{\mathbf{w}}^T \mathbf{M}_{\text{ne}} \tilde{\mathbf{w}} \\ \text{subject to} & \tilde{\mathbf{w}} \in \widetilde{\mathcal{W}}_1 \times \cdots \times \widetilde{\mathcal{W}}_K, \end{array} \tag{7.28}$$

where

$$\mathbf{M}_{\text{ne}} = \text{Diag}(\boldsymbol{\rho}) \otimes \tilde{\boldsymbol{\Sigma}} + \frac{1}{2}(\mathbf{I} + \mathbf{J}) \otimes \tilde{\boldsymbol{\Omega}}, \tag{7.29}$$

in the sense that $\tilde{\mathbf{w}}$ is an NE of (7.27) if and only if it is optimal to (7.28). Later, the authors of [210] showed that (7.28) is strongly convex and thus its optimal solution, or equivalently, the NE of (7.27) or (7.13), is unique.

7.3.5 Multi-Portfolio Optimization with Global Constraints

Even when there exist global constraints, one can still show that the GNEP (7.15) is equal to the following GNEP [210]:

$$\left.\begin{array}{ll} \underset{\tilde{\mathbf{w}}_k}{\text{maximize}} & \tilde{u}(\tilde{\mathbf{w}}_k, \tilde{\mathbf{w}}_{-k}) \\ \text{subject to} & \tilde{\mathbf{w}}_k \in \widetilde{\mathcal{W}}_k \\ & \tilde{\mathbf{g}}(\tilde{\mathbf{w}}) \leq \mathbf{0} \end{array}\right\} \forall k, \tag{7.30}$$

where

$$\tilde{\mathbf{g}}(\tilde{\mathbf{w}}) \triangleq \sum_{k=1}^{K} \begin{bmatrix} \left(\tilde{w}_{k,i}^{+} + \tilde{w}_{k,i}^{-}\right)_{i=1}^{N} \\ \left(\sum_{i \in \mathcal{G}_l} \left(\tilde{w}_{k,i}^{+} + \tilde{w}_{k,i}^{-}\right) - U_l\right)_{l=1}^{L} \end{bmatrix} - \begin{bmatrix} (D_i)_{i=1}^{N} \\ (U_l)_{l=1}^{L} \end{bmatrix}. \tag{7.31}$$

Similar to (7.28), one can construct the following convex problem with global constraints:

$$\begin{array}{ll} \underset{\tilde{\mathbf{w}}}{\text{maximize}} & P_{\text{ne}}(\tilde{\mathbf{w}}) = \tilde{\boldsymbol{\mu}}^T \tilde{\mathbf{w}} - \frac{1}{2} \tilde{\mathbf{w}}^T \mathbf{M}_{\text{ne}} \tilde{\mathbf{w}} \\ \text{subject to} & \tilde{\mathbf{w}} \in \widetilde{\mathcal{W}}_1 \times \cdots \times \widetilde{\mathcal{W}}_K, \\ & \tilde{\mathbf{g}}(\tilde{\mathbf{w}}) \leq \mathbf{0}. \end{array} \tag{7.32}$$

However, now the constraint sets of all the $\tilde{\mathbf{w}}_k$ are coupled and one can only conclude that an optimal solution to (7.32) is a GNE of the GNEP (7.30), but not vice versa [210]. An optimal solution to (7.32) is referred to as a Variational Equilibrium (VE), and it is actually unique since (7.32) is strongly convex.

Example 7.1. Let us now consider some numerical experiments. The mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and market impact coefficient

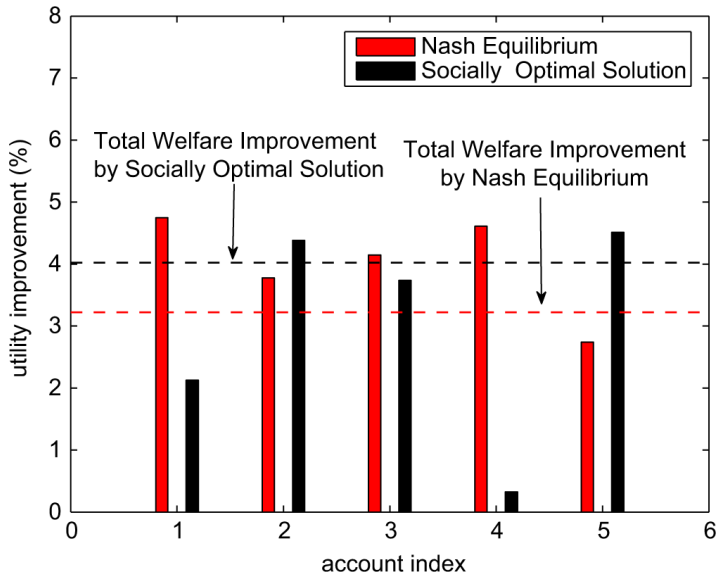

Figure 7.2: Utility improvement of the NE, and socially optimal solution against the naive solution.

matrices $\mathbf{\Omega}^+$ and $\mathbf{\Omega}^-$ are randomly generated. Suppose there are $N = 5$ assets.

For the moment, the number of accounts is fixed to $K = 5$ with individual constraints. We compare three methods, i.e., i) the naive problem (7.11), ii) the total social welfare maximization problem (7.12), and iii) the NEP (7.13), in terms of two criteria:

- the relative utility improvement of each account:

$$\frac{u_k(\mathbf{w}) - u_k(\mathbf{w}_{\text{naive}})}{u_k(\mathbf{w}_{\text{naive}})} \tag{7.33}$$

- the relative utility improvement of all the accounts:

$$\frac{\sum_{k=1}^{K} u_k(\mathbf{w}) - \sum_{k=1}^{K} u_k(\mathbf{w}_{\text{naive}})}{\sum_{k=1}^{K} u_k(\mathbf{w}_{\text{naive}})} \tag{7.34}$$

where $\mathbf{w}$ is either $\mathbf{w}_{\text{ne}}$ and $\mathbf{w}_{\text{so}}$ and $\mathbf{w} = \mathbf{0}$.

Figure 7.2 shows the numerical results measured by (7.33) and 7.34. We can see that the social welfare maximization problem (7.12)

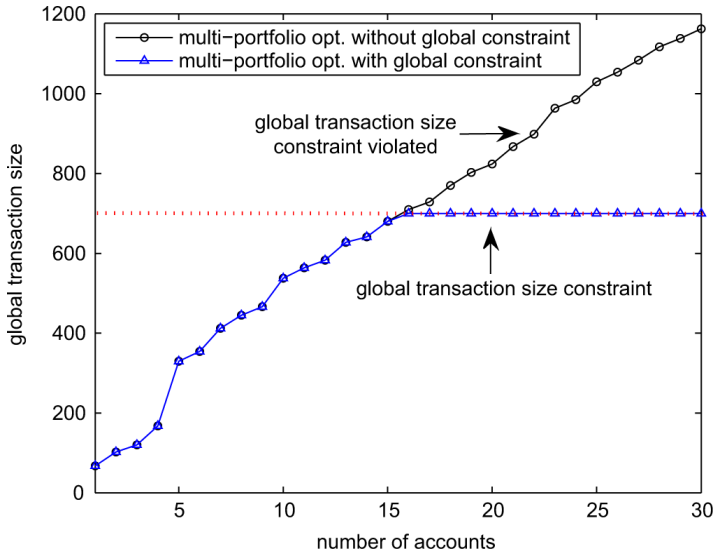

Figure 7.3: Global transaction size versus number of accounts.

achieves the best total social welfare (see the horizontal dashed black line), but at the price of sacrificing accounts 1, 3, and especially, account 4 (see the vertical black bars). The NEP improves the total welfare significantly (see the horizontal dashed red line) albeit below the social solution; however, opposed to the social solution, it does not sacrifice individual accounts as much as the social formulation (see the red bars vs the black bars).

Later we also considered to include some global constraints, e.g., a global transaction size constraint. Figure 7.3 shows the total transaction size versus of the number of accounts. Clearly, we see that the global transaction size constraint may be violated if it is not properly considered. ■

Remark 7.1. In this chapter, we have focused on the reformulation of different nonconvex problems in convex form, but without going into the details of the specific algorithms to solve such problems. It is possible to derive highly efficient parallel and distributed algorithms for the above convex problems [210]. ■

8

Index Tracking

Active investment strategies assume that the markets are not perfectly efficient and fund managers can identify mispriced stocks and/or make superior predictions and then collect (hopefully positive) profits based on them (cf. Chapters 5-7).

Passive investment strategies, on the other hand, assume the markets are efficient enough and cannot be beaten in the long run, therefore, the investment philosophy is to directly follow the markets.

This chapter reviews one of the most popular and important passive investment strategies: index tracking. The goal of index tracking is to construct a tracking portfolio whose value follows a market index (or some preferred benchmark index).

The detailed organization of this chapter is as follows. Section 8.1 reviews different methods of index tracking, i.e., full index tracking, synthetic index tracking, and sparse index tracking. Sections 8.2 and 8.3 focus on two approaches of sparse index tracking, i.e., the two-step approach and joint optimization approach, separately.

8.1 Different Index Tracking Methods

Suppose that a benchmark index is composed of N stocks, let $\mathbf{r}^b = [r_1^b, \ldots, r_T^b]^T \in \mathbb{R}^T$ and $\mathbf{X} = [\mathbf{r}_1, \ldots, \mathbf{r}_T]^T \in \mathbb{R}^{T \times N}$ denote the returns of the benchmark index and the N stocks in the past T days, respectively. Let $\mathbf{b} \in \mathbb{R}^N$ denote the (normalized) benchmark index weights such that $\mathbf{b} > \mathbf{0}$, $\mathbf{b}^T\mathbf{1} = 1$, and $\mathbf{X}\mathbf{b} = \mathbf{r}^b$. Further, let $\mathbf{w}$ denote the tracking portfolio to be designed, which must satisfy $\mathbf{w} \geq \mathbf{0}$ and $\mathbf{w}^T\mathbf{1} = 1$.

8.1.1 Tracking Performance

Tracking Error

Given the covariance matrix of the benchmark stocks $\mathbf{\Sigma}$ and the benchmark index weight vector $\mathbf{b}$, the theoretical tracking error is defined as

$$\mathrm{TTE}(\mathbf{w}) = (\mathbf{w} - \mathbf{b})^T \mathbf{\Sigma} (\mathbf{w} - \mathbf{b}). \tag{8.1}$$

Since $\mathbf{\Sigma}$ needs to be estimated first and $\mathbf{b}$ may not be available, the empirical tracking error, defined as

$$\mathrm{TE}(\mathbf{w}) = \frac{1}{T}\|\mathbf{X}\mathbf{w} - \mathbf{r}^b\|_2^2, \tag{8.2}$$

is more popular in practice [134, 177]. It measures how closely the tracking portfolio mimics the benchmark index empirically. In principal, the smaller, the better. Note that the daily stock returns are in general very small and if we suppose $\mathsf{E}[\mathbf{r}_t] = \mathbf{0}$, the expected value of TE equals TEE:

$$\mathsf{E}[\mathrm{TE}(\mathbf{w})] = \frac{1}{T}\mathsf{E}\left[\|\mathbf{X}\mathbf{w} - \mathbf{r}^b\|_2^2\right], \tag{8.3}$$

$$= (\mathbf{w} - \mathbf{b})^T \mathsf{E}\left[\frac{1}{T}\mathbf{X}^T\mathbf{X}\right](\mathbf{w} - \mathbf{b}), \tag{8.4}$$

$$= (\mathbf{w} - \mathbf{b})^T \mathbf{\Sigma} (\mathbf{w} - \mathbf{b}). \tag{8.5}$$

Excess Return

Apart from tracking error, another important criterion is excess return (ER):

$$\mathrm{ER}(\mathbf{w}) = \frac{1}{T}\mathbf{1}^T(\mathbf{X}\mathbf{w} - \mathbf{r}^b), \tag{8.6}$$

It represents how much the tracking portfolio outperforms the benchmark index, and the larger, the better.

Combined Criterion

To achieve a trade-off between the tracking error (8.2) and the excess return (8.6), a combined objective is thus considered [17, 19]:

$$U(\mathbf{w}) = \alpha\mathrm{TE}(\mathbf{w}) - (1-\alpha)\mathrm{ER}(\mathbf{w}), \tag{8.7}$$

where $\alpha \in [0, 1]$ is a predefined trade-off parameter.

Goal

Since the excess return (8.6) is linear in the tracking portfolio $\mathbf{w}$, without loss of generality and for clarity of presentation, we focus on the tracking error (8.2) only. The goal of index tracking is to construct a portfolio $\mathbf{w}^\star$ (or a derivative like future contract) to track the performance of the benchmark index with the tracking error (8.2) being small or, even better, minimized.

8.1.2 Full Index Tracking

The most straightforward tracking method, referred to as full index tracking, is to purchase all the index constituents in appropriate quantities to perfectly track the index, i.e., $\mathbf{w}^\star = \mathbf{b}$ and $\|\mathbf{X}\mathbf{w}^\star - \mathbf{r}^b\|_2^2 = 0$. However, it has several significant disadvantages [17, 134], for example:

- including all the stocks may not be practical especially when the index contains some illiquid stocks and it is hard to purchase such stocks; and
- allocating capital in all assets would also incur significant trading cost.

8.1.3 Synthetic Index Tracking

The second method is to use derivatives, like future contracts, to track the index (e.g., E-mini S&P future contract is used to track the S&P500 Index). The advantage of future contracts is that the trading cost is relatively lower than stocks, however, dynamically tracking the index by rolling the contracts can be both expensive and risky because of the counterparty risk and illiquidity of contracts. These drawbacks make future contracts less attractive in tracking the index [19, 106].

8.1.4 Sparse Index Tracking

To make the index tracking more practical (i.e., relatively lower trading cost and less risky), a third method was proposed: to use a subset of stocks to track the index (i.e., $\|\mathbf{w}^\star\|_0 \ll N$) with only a small sacrifice in tracking error (i.e., $\|\mathbf{X}\mathbf{w}^\star - \mathbf{r}^b\|_2$ is still close to 0) [17]. This method is referred to as sparse index tracking and in fact it is the core business of ETFs, which now have been very popular in the markets[1].

In the following content of this chapter, we will focus on two main approaches of sparse index tracking, namely, the two-step approach and joint optimization approach.

8.2 Sparse Index Tracking: Two-Step Approach

The first approach of sparse index tracking is to decompose the task into two steps [19, 52, 155]:

- stock selection: selecting a subset of K ($K \ll N$) stocks; and
- capital allocation: distributing the capital among the selected stocks.

8.2.1 Stock Selection

Let us first introduce different stock selection methods.

[1] Many funds provide some ETF products and some of them have very large assets under management (AUM) even at the magnitude of $10 billion USD, e.g., see http://etfdb.com/type/size/large-cap/.

Random Selection

One simple and naive idea is to randomly select K stocks from the N index stocks [52]. This method in general is used as a benchmark.

Selection Based Market Capitals

A widely used stock selection method, especially for a market capital weighted index, is to select the largest K stocks according to their market capitals (e.g., the product of outstanding shares[2] and prices) [155]. For the market capital weighted index, if the index weight vector $\mathbf{b}$ is available, one can select the stocks with K largest weights b_i.

Selection Based on Correlation

Another idea is to select the stocks that have similar return performances as the index [19, 52]. For example, given the correlation between the i-th stock and the benchmark index

$$\rho_{ib} = \mathsf{Cov}(\mathbf{X}_{\cdot i}, \mathbf{r}^b), \tag{8.8}$$

this method selects the stocks with K largest correlations ρ_{ib}.

Selection Based on Cointegration

The idea is to select K stocks so that there exists a linear combination of their log-prices cointegrated well with the value of the benchmark index [5, 19]. Mathematically, based on the following model:

$$I_t = \sum_{i=1}^{N} s_i \beta_i p_{i,t} + w_t, \tag{8.9}$$

where $s_i \in \{0, 1\}$, one needs to find the optimal $\mathbf{s}$ with $\mathbf{s}^T\mathbf{1} = K$ (i.e., selection of K stocks) and the weights β_i such that w_t is most likely stationary (e.g., resulting the smallest p-value of the stationary test). This problem itself is NP-hard. Exhaustive search can be employed

[2]Outstanding shares refer to a company's stock currently held by all its shareholders, including share blocks held by institutional investors and restricted shares owned by the company's officers and insiders.

when N is small; some heuristic method, e.g., genetic algorithm, is needed otherwise [19].

8.2.2 Capital Allocation

Once a subset of K stocks has been selected, the second step is to design the capital allocation among them. Before we proceed, let us use the binary vector $\mathbf{s}^\star \in \mathbb{R}^N$:

$$s_i^\star = \begin{cases} 1, & \text{if stock } i \text{ is selected} \\ 0, & \text{otherwise} \end{cases} \tag{8.10}$$

with $\mathbf{1}^T\mathbf{s}^\star = K$ to represent the selected K stocks.

Naive Allocation

When the benchmark portfolio weight vector $\mathbf{b}$ is known, a naive allocation is to distribute the capital among the selected stocks proportional to the original weights with their summation equal to 1. That is, the naive allocation weight vector is

$$\mathbf{w}^\star = \frac{\mathbf{b} \odot \mathbf{s}^\star}{\mathbf{1}^T (\mathbf{b} \odot \mathbf{s}^\star)}, \tag{8.11}$$

where $\odot$ means Hadamard product.

Optimization Allocation

The naive allocation weight (8.11) is simple enough, however, the tracking error is not optimized and sometimes the benchmark weight vector $\mathbf{b}$ may not be available. The optimization allocation overcomes this drawback by minimizing the tracking error based on the selected stocks directly as follows [155][3]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \frac{1}{T}\|\mathbf{X}(\mathbf{w} \odot \mathbf{s}^\star) - \mathbf{r}^b\|_2^2 \\ \text{subject to} & \mathbf{1}^T(\mathbf{w} \odot \mathbf{s}^\star) = 1, \\ & \mathbf{w} \geq \mathbf{0}. \end{array} \tag{8.12}$$

[3]The authors of [155] considered a more complicated nonconvex objective and they employed a genetic algorithm to solve their nonconvex problem.

Problem (8.12) is convex and can be solved efficiently. The optimal allocation simply is the optimal solution of (8.12).

8.3 Sparse Index Tracking: Joint Optimization Approach

The previous approach executes the two steps of stock selection and capital allocation sequentially; however, it is not clear how optimal the resulting tracking portfolio is. A better approach may be to conduct these two steps jointly and systematically.

8.3.1 Problem Formulation

A direct way is to regularize the cardinality of the tracking portfolio weights [106]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \frac{1}{T}\|\mathbf{X}\mathbf{w}-\mathbf{r}^b\|_2^2+\lambda\|\mathbf{w}\|_0 \\ \text{subject to} & \mathbf{1}^T\mathbf{w}=1, \\ & \mathbf{w}\geq\mathbf{0}, \end{array} \tag{8.13}$$

where $\lambda \geq 0$ is a predefined parameter.

8.3.2 ℓ_1-norm Approximation

Generally, problem (8.13) is hard to solve due to the nonconvex and discontinuous cardinality term $\|\mathbf{w}\|_0$ (note that $\|\mathbf{w}\|_0 = \sum_{i=1}^N \mathbb{1}_{\{w_i\neq 0\}}$). Figure 8.1 shows the indicator function $\mathbb{1}_{\{x\neq 0\}}$ (see the solid black line).

A popular approximation of $\|\mathbf{w}\|_0$ that is convex and promotes sparsity is the ℓ_1-norm function $\|\mathbf{w}\|_1$ as indicated by the dashed red line in Figure 8.1, i.e., the LASSO (least absolute shrinkage and selection operator) technique [96]. LASSO has indeed been used in portfolio optimization [3, 33, 48, 73, 74].

Unfortunately, this technique does not work for index tracking with long only constraints (i.e., $\mathbf{1}^T\mathbf{w}=1$ and $\mathbf{w}\geq\mathbf{0}$) since

$$\|\mathbf{w}\|_1 = \sum_{i=1}^N |w_i| = \sum_{i=1}^N w_i = \mathbf{1}^T\mathbf{w} = 1 \tag{8.14}$$

is constant.

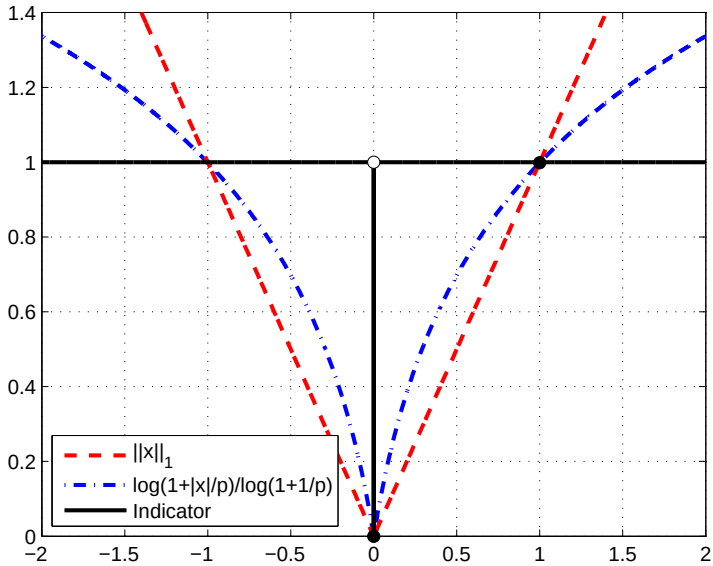

Figure 8.1: Indicator function and approximations.

8.3.3 Reweighted ℓ_1-norm Approximation

Since the convex ℓ_1-norm approximation does not work for an index tracking problem, a better (possibly nonconvex) approximation is needed. An example is [37]

$$\rho_p(x) = \frac{\log(1 + |x|/p)}{\log(1 + 1/p)}, \tag{8.15}$$

where $p > 0$ is a parameter and $\rho_p(x) \to \mathbb{1}_{\{x \neq 0\}}$ as $p \to 0$. Figure 8.1 shows an illustrative example of $p = 0.2$, i.e., the dashed-dotted blue line.

Replacing the indicator function $\mathbb{1}_{\{x \neq 0\}}$ by the approximation function $\rho_p(x)$ results in the following problem:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \frac{1}{T}\|\mathbf{X}\mathbf{w} - \mathbf{r}^b\|_2^2 + \lambda \sum_{i=1}^N \rho_p(w_i) \\ \text{subject to} & \mathbf{1}^T\mathbf{w} = 1, \\ & \mathbf{w} \geq \mathbf{0}. \end{array} \tag{8.16}$$

In fact, there are also some other approximations for the indicator function, e.g., see [185] and references therein. For example, the $|x|^p$ with $0 < p < 1$ is used in [75, 106] and a smoothed version of $|x|^p$ is used in [41]. However, there does not exist either efficient algorithms [41, 106] or heuristic algorithms that can guarantee the quality of the solution [75].

Following [37], we will present an iterative algorithm that interestingly turns out to replace $\|\mathbf{w}\|_0$ with a sequence of reweighted ℓ_1-norm approximations. The idea also applies to the problems in [41, 106].

The idea is, at each iteration point, say x^0, to approximate $\rho_p(x)$ with its first-order Taylor approximation, as follows:

$$\begin{aligned}
\rho_p(x) &= \frac{\log(1+|x|/p)}{\log(1+1/p)} && (8.17)\\
&\approx \frac{1}{\log(1+1/p)}\left[\frac{|x|}{p+|x^0|}+\log\left(1+\left|x^0\right|/p\right)-\frac{|x^0|}{p+|x^0|}\right] && (8.18)\\
&= \underbrace{\frac{1}{(p+|x^0|)\log(1+1/p)}}_{\triangleq d(x^0)}|x|+\text{const} && (8.19)\\
&= d(x^0)\,|x|+\text{const} && (8.20)\\
&\triangleq u(x,x^0). && (8.21)
\end{aligned}$$

Figure 8.2 shows an illustrative example of $u(x, x^0)$ at point $x^0 = 1$ (see the dashed magenta line).

Then at the k-th iteration point $\mathbf{w}^k$, one can solve the following reweighted approximation problem to get the next iteration point $\mathbf{w}^{k+1}$:

$$\begin{aligned}
\underset{\mathbf{w}}{\text{minimize}} \quad & \frac{1}{T}\|\mathbf{X}\mathbf{w}-\mathbf{r}^b\|_2^2+\lambda\left\|\mathbf{D}\left(\mathbf{w}^k\right)\mathbf{w}\right\|_1\\
\text{subject to} \quad & \mathbf{1}^T\mathbf{w}=1,\\
& \mathbf{w}\geq\mathbf{0},
\end{aligned} \tag{8.22}$$

where

$$\mathbf{D}\left(\mathbf{w}^k\right)=\text{Diag}\left(d\left(w_1^k\right),\ddots,d\left(w_N^k\right)\right). \tag{8.23}$$

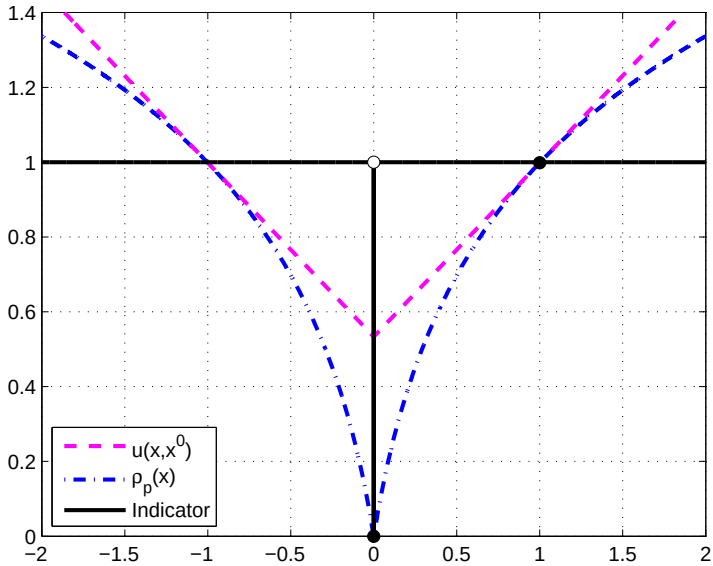

Figure 8.2: Reweighted ℓ_1-norm approximation.

Algorithm 6 summarizes the iterative procedure. It can be easily shown that Algorithm 6 converges to a stationary point of problem (8.16) following [164].

Algorithm 6 Reweighted ℓ_1-norm approximation for index tracking.

Input: $\mathbf{w}^0$

Output: a stationary point of problem (8.16)

1: **repeat**
2: Compute $d\left(w_i^k\right)$ according to (8.19)
3: Compute $\mathbf{D}\left(\mathbf{w}^k\right)$ according to (8.23)
4: Solve (8.22) and set the optimal solution as $\mathbf{w}^{k+1}$
5: $k \leftarrow k+1$
6: **until** convergence

Example 8.1. For illustration purposes, here we conduct some synthetic experiments in MATLAB.

The data is synthetically generated as follows. We consider

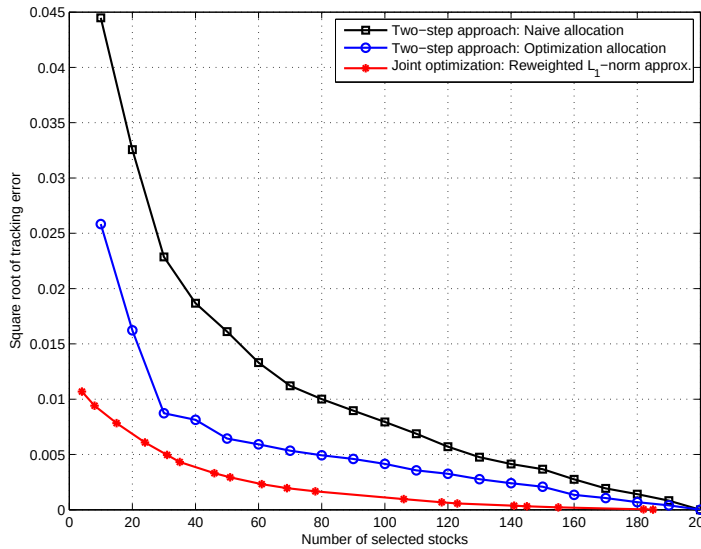

Figure 8.3: Comparisons of different sparse index tracking methods.

$N = 200$ stocks and draw $T = 1000$ i.i.d. samples, denoted as $\mathbf{r}_1, \ldots, \mathbf{r}_{1000}$, from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \texttt{randn}(\texttt{N}, 1)/252$ and $\boldsymbol{\Sigma} = \mathbf{DCD}$ with $\mathbf{D} = 2 * \texttt{diag}(\texttt{rand}(\texttt{N}, 1))/\texttt{sqrt}(252)$ and $\mathbf{C}_{ij} = 0.7^{|i-j|}$. The data matrix is $\mathbf{X} = \left[\mathbf{r}_1^T, \ldots, \mathbf{r}_{1000}^T\right]^T \in \mathbb{R}^{1000 \times 200}$.

Next, we construct an artificial index. We first randomly generate a temporary vector $\mathbf{t} = \texttt{rand}(\texttt{N}, 1)$ and then set the artificial index weights by normalizing $\mathbf{t}$ so that the summation of the weights equals one, i.e., $\mathbf{b} = \frac{\mathbf{t}}{\mathbf{1}^T\mathbf{t}}$. The historical returns of the constructed benchmark index are $\mathbf{r}^b = \mathbf{Xb} \in \mathbb{R}^T$.

We compare the following sparse index tracking methods:

- two-step approach: we select the stocks with K largest correlations (8.8) and then consider both the naive allocation (i.e., (8.11)) and the optimization allocation (i.e., (8.12)); and
- joint optimization approach: Algorithm 6.

Figure 8.3 shows the square root of tracking error versus the number of selected stocks. We can clearly see that: i) for the two-step method,

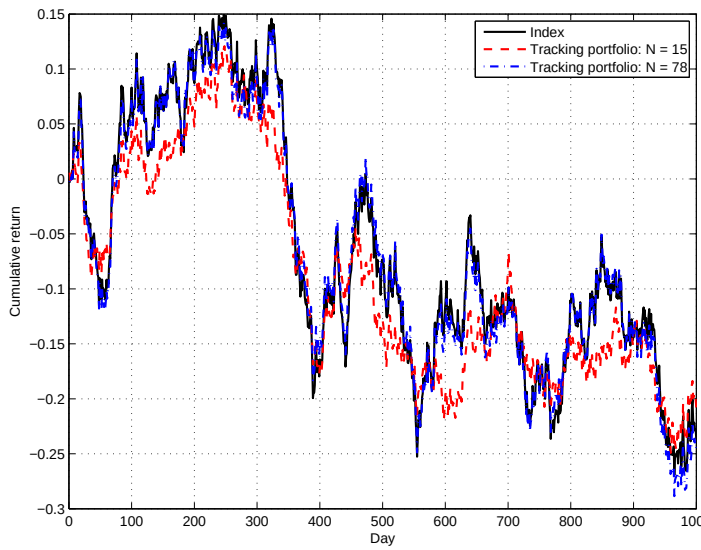

Figure 8.4: Tracking performances of some sparse index tracking portfolios.

the optimization allocation method outperforms the naive allocation method, e.g., the square root of the tracking error is reduced from 4.45% to 2.58% when $K = 10$; and ii) the joint optimization approach outperforms the methods of the two-step approach, e.g., the joint optimization approach even achieves a much lower square root of tracking error at 0.94% with fewer stocks $K = 8$ compared with the results 4.45% and 2.58% of the two-step approach with $K = 10$.

Figure 8.4 shows the tracking performances of the joint optimization approach: the tracking path deviates from the index path significantly when $K = 15$ (see the dashed red line) and the tracking path mimics the index path very closely when $K = 78$ (see the dashed-dotted blue line). ■

8.3.4 Nonconvex Constraints

For simplicity, we imposed only the long only constraints (i.e., $\mathbf{1}^T\mathbf{w} = 1$ and $\mathbf{w} \geq \mathbf{0}$) in the previous parts of this chapter. In practice, some fund managers may also impose some holding constraints (see Section 5.1.4)

and the joint optimization problem becomes

$$\begin{aligned} \underset{\mathbf{w},\mathbf{s}}{\text{minimize}} \quad & \frac{1}{T}\|\mathbf{X}\mathbf{w}-\mathbf{r}^b\|_2^2+\lambda\mathbf{1}^T\mathbf{s} \\ \text{subject to} \quad & \mathbf{1}^T\mathbf{w}=1, \\ & \mathbf{w}\geq\mathbf{0}, \\ & s_iL_i\leq w_i\leq s_iU_i, \quad \forall i \\ & s_i\in\{0,1\}, \quad \forall i \end{aligned} \tag{8.24}$$

where L_i and U_i are the holding lower and upper bound for the i-th stock, respectively, only if it is selected, and $0 \leq L_i \leq U_i$.

The binary variable $\mathbf{s}$ complicates problem (8.24). There are several different methods to deal with it. In the following, we briefly explain each method and list the corresponding references:

- thresholding method: a practical heuristic is to solve the problem (8.24) without the binary variable $\mathbf{s}$ and then select the stocks with weights larger than a certain threshold (i.e., decide $\mathbf{s}$ based on the optimized $\mathbf{w}$) and then optimize (8.24) with $\mathbf{s}$ fixed. To make the solution more robust, one can remove a few stocks each time and apply the idea several times to achieve enough sparsity in the portfolio [106];

- mixed-integer programming (MIP): problem (8.24) indeed is an MIP and there are some commercial solvers like GUROBI[4] and CPLEX[5] that can solve MIPs with small and medium sizes efficiently. Thus, one can directly apply such standard solvers to solve small and medium size MIP type index tracking problems [36, 177];

- heuristic algorithms: for MIP with a large size, standard solvers may fail, and some heuristic algorithms, e.g., genetic algorithms [10, 17, 177], and differential evolution [10, 134], are used in practice. However, the solution in general may be far from optimal.

[4]http://www.gurobi.com/

[5]http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/index.html

9

Risk Parity Portfolio Optimization

The Markowitz portfolio (cf. Chapters 5-7) has never been embraced by practitioners, among other reasons because it only considers the risk of the portfolio as a whole and ignores the risk diversification.

Recently, an alternative risk parity portfolio design has been receiving significant attention from both the theoretical and practical sides due to its advantage in diversification of (ex-ante) risk contributions among assets. Such risk contributions can be deemed good predictors for the (ex-post) loss contributions, especially when there exist huge losses. The main goal of this chapter is to introduce the concepts of risk parity portfolio, review different existing formulations, and study different efficient solving algorithms.

The detailed organization is as follows. Section 9.1 introduces the concepts of risk contribution and risk parity portfolio. Section 9.2 lists several existing specific risk parity formulations and presents a general risk parity portfolio problem formulation that can fit most of the listed specific risk parity formulations. To solve the risk parity problems, Section 9.3 details an efficient numerical solving approach for the general risk parity portfolio problem formulation based on successive convex optimization methods.

9.1 What is a Risk Parity Portfolio?

Let us first start with introducing the concept of risk contribution based on which we can define the risk parity portfolio.

9.1.1 Risk Contribution

Suppose there are N assets and the mean vector and (positive definite) covariance matrix of the returns are denoted as $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$, respectively. For a portfolio $\mathbf{w} \in \mathbb{R}^N$, to study the risk parity portfolio, we need some well defined risk measurements $f(\mathbf{w})$ so that the "risk contribution" of each asset to the risk of the whole portfolio can be quantified. We start with the following desired property as it will be the key to quantify the risk parity.

Theorem 9.1 (Euler's Theorem)**.** Let a continuous and differentiable function $f\colon \mathbb{R}^N \mapsto \mathbb{R}$ be a positively homogeneous function of degree one [1]. Then

$$f(\mathbf{w}) = \sum_{i=1}^{N} w_i \frac{\partial f}{\partial w_i}. \tag{9.1}$$

One observation from property (9.1) is that the component $w_i \frac{\partial f}{\partial w_i}$ can be regarded as the risk contribution from asset i to the total risk $f(\mathbf{w})$.

Interestingly and fortunately, most of the existing risk measurements do satisfy the Euler property (9.1) either directly (VaR and CVaR) or indirectly (variance) as we show next.

Volatility

Note that variance $\sigma^2(\mathbf{w}) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ does not satisfy (9.1) directly. Fortunately, it is easy to check that volatility $\sigma(\mathbf{w}) = \sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$ does satisfy (9.1):

$$\sum_{i=1}^{N} w_i \frac{\partial \sigma}{\partial w_i} = \sum_{i=1}^{N} w_i \left(\frac{\boldsymbol{\Sigma}\mathbf{w}}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}}\right)_i = \frac{1}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} \sum_{i=1}^{N} w_i (\boldsymbol{\Sigma}\mathbf{w})_i$$

[1] A function $f(\mathbf{w})$ is a positively homogeneous function of degree one if $f(c\mathbf{w}) = cf(\mathbf{w})$ holds for any constant $c > 0$.

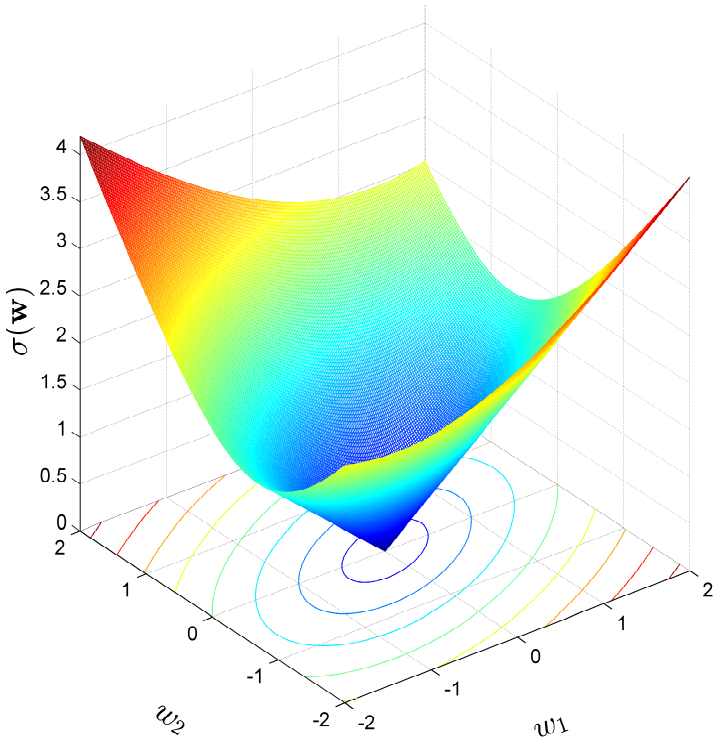

Figure 9.1: One example that satisfies the Euler property (9.1).

$$= \frac{1}{\sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} = \sigma(\mathbf{w}). \tag{9.2}$$

Thus variance fits (9.1) indirectly via volatility. Figure 9.1 shows an example of $\sigma(\mathbf{w})$ and we can see that the function is linear along any direction starting from the origin.

VaR and CVaR

For simplicity, we consider the Gaussian case VaR and CVaR in this chapter. For the Gaussian distribution, VaR and CVaR can be expressed explicitly as [141]

$$\mathrm{VaR}_{1-\varepsilon}(\mathbf{w}) = -\boldsymbol{\mu}^T \mathbf{w} + \kappa_1(\varepsilon) \sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}, \tag{9.3}$$

$$\mathrm{CVaR}_{1-\varepsilon}(\mathbf{w}) = -\boldsymbol{\mu}^T \mathbf{w} + \kappa_2(\varepsilon) \sqrt{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}, \tag{9.4}$$

where $\kappa_1(\varepsilon) \triangleq Q^{-1}(\varepsilon)$ and $\kappa_2(\varepsilon) \triangleq \frac{1}{\sqrt{2\pi}\varepsilon} e^{-\frac{\left(Q^{-1}(\varepsilon)\right)^2}{2}}$, and $Q^{-1}(\cdot)$ is the inverse of the Q-function (see (4.18)). Here, we implicitly assume that ε is small (e.g., $\varepsilon \leq 20\%$) and $\kappa_1(\varepsilon)$ and $\kappa_2(\varepsilon)$ are both positive.

From (9.3) and (9.4) we can see that if $\boldsymbol{\mu} \propto \mathbf{1}$, ignoring the constant terms, the volatility, VaR, and CVaR are equal up to a positive scalar.

More generally, the Gaussian distribution can be extended to elliptical distributions [119] for which VaR and CVaR both are mean and standard deviation trade-off expressions.

Remark 9.1. For the more general non-Gaussian VaR and CVaR, it can be shown that they both satisfy (9.1), however, they do not have closed-form expressions and some approximations are needed. For more discussions, please refer to [76] and references therein. ■

9.1.2 Risk Parity Portfolio

The risk parity portfolio is a portfolio such that each asset has the same risk contribution. That is, given the risk measurement $f(\mathbf{w})$, the risk parity portfolio should satisfy [162, 163, 131]

$$w_i \frac{\partial f(\mathbf{w})}{\partial w_i} = w_j \frac{\partial f(\mathbf{w})}{\partial w_j}, \quad \forall i, j. \tag{9.5}$$

Risk budgeting portfolio is a more general concept. Given a budget vector $\mathbf{b} = [b_1, \ldots, b_N]^T > \mathbf{0}$, and $\mathbf{b}^T\mathbf{1} = 1$, where budget $\mathbf{b}$ is interpreted as a perdetermined percentage risk contribution target for all the assets, the risk budgeting portfolio should satisfy

$$w_i \frac{\partial f(\mathbf{w})}{\partial w_i} = b_i f(\mathbf{w}), \quad \forall i. \tag{9.6}$$

Obviously, the risk parity portfolio is a special case of the risk budgeting portfolio with $\mathbf{b} = \mathbf{1}/N$.

Due to the popularity of the terminology "risk parity", it is always used to refer to a broad portfolio allocation method of risk diversification (e.g., including both risk parity and risk budgeting portfolios) [167]. We take the broad concept of "risk parity" unless specified otherwise in this chapter.

9.2 Risk Parity Portfolio Formulations

There are many different existing specific formulations on risk parity portfolios due to different risk measurements used or different profiles of investors. In this section, we first review some specific formulations and then consider a general risk parity portfolio problem formulation.

9.2.1 Some Specific Formulations

Recall that the risk contribution of asset i is $\frac{w_i(\mathbf{\Sigma w})_i}{\sqrt{\mathbf{w\Sigma w}}}$, then the risk parity (9.5) and risk budgeting (9.6) relationships turn out to be

$$\text{risk parity}: \qquad w_i\left(\mathbf{\Sigma w}\right)_i = w_j\left(\mathbf{\Sigma w}\right)_j, \tag{9.7}$$

$$\text{risk budgeting}: \qquad w_i\left(\mathbf{\Sigma w}\right)_i = b_i \mathbf{w}^T\mathbf{\Sigma w}, \tag{9.8}$$

respectively, where $\mathbf{b} = [b_1, \ldots, b_N]^T > \mathbf{0}$ is the given risk budgeting for n assets and $\mathbf{b}^T\mathbf{1} = 1$. Actually, relationship (9.7) is a special case of relationship (9.8) with $b_i = 1/N$ for all i.

Again, we denote the feasible set as $\overline{\mathcal{W}} \triangleq \{\mathbf{w}|\mathbf{w}^T\mathbf{1} = 1\} \cap \mathcal{W}$ where $\mathbf{w}^T\mathbf{1} = 1$ denotes the capital budget constraint and $\mathcal{W}$ is a convex set that denotes the other constraints.

Only when $\mathbf{\Sigma}$ is diagonal and there exists a long-only constraint, i.e., $\mathcal{W} = \{\mathbf{w}|\mathbf{w} \geq \mathbf{0}\}$, the nonlinear equation systems (9.8) admit a unique solution [167]:

$$w_i = \frac{\sqrt{b_i}/\sqrt{\Sigma_{ii}}}{\sum_{k=1}^n \sqrt{b_k}/\sqrt{\Sigma_{kk}}}, \quad i = 1, \ldots, N. \tag{9.9}$$

However, for non-diagonal $\mathbf{\Sigma}$ or when there are some additional constraints, the closed-form solution does not exist anymore and some optimization problems are constructed instead.

Paper [131] is one of the first few papers that focuses on finding the risk parity portfolio via optimization. The proposed problem formulation is to penalize the summation of squared differences among risk contributions:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \sum_{i,j=1}^N \left(w_i\left(\mathbf{\Sigma w}\right)_i - w_j\left(\mathbf{\Sigma w}\right)_j\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{array} \tag{9.10}$$

Motivated by problem (9.10), Bai et al. [14] simplified the objective of (9.10) to solve:

$$\begin{array}{ll} \underset{\mathbf{w},\theta}{\text{minimize}} & \sum_{i=1}^{N} \left(w_i \left(\boldsymbol{\Sigma}\mathbf{w}\right)_i - \theta\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{array} \tag{9.11}$$

To find a portfolio that meets the risk budgeting targets $\mathbf{b}$ as much as possible, Bruder and Roncalli proposed to solve [34]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \sum_{i=1}^{N} \left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}} - b_i\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{array} \tag{9.12}$$

Similarly, it is easy to have some other alternative (but different) problem formulations, e.g.,

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \sum_{i,j=1}^{N} \left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{b_i} - \frac{\mathbf{w}_j(\boldsymbol{\Sigma}\mathbf{w})_j}{b_j}\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{array} \tag{9.13}$$

and

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \sum_{i=1}^{N} \left(w_i \left(\boldsymbol{\Sigma}\mathbf{w}\right)_i - b_i\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{array} \tag{9.14}$$

and

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \sum_{i=1}^{N} \left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}} - b_i\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{array} \tag{9.15}$$

and

$$\begin{array}{ll} \underset{\mathbf{w},\theta}{\text{minimize}} & \sum_{i=1}^{N} \left(\frac{w_i(\boldsymbol{\Sigma}\mathbf{w})_i}{b_i} - \theta\right)^2 \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}. \end{array} \tag{9.16}$$

Note that all the above formulations are nonconvex and they are only some examples. Actually, there are many more specific formulations; for more a comprehensive summary, please see [76, Table I].

Unfortunately, all the above problem formulations are generally nonconvex in general. In the following we review a numerical approach that attacks all of them in a unified way.

9.2.2 General Risk Parity Portfolio Problem

Let us start with a general risk parity formulation proposed in [76] that can fit all the previously stated specific formulations. The general risk parity formulation can be expressed as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & U(\mathbf{w}) \triangleq R(\mathbf{w}) + \lambda F(\mathbf{w}) \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{array} \tag{9.17}$$

where

- $R(\mathbf{w})$ measures the risk concentration and has the form

$$R(\mathbf{w}) \triangleq \sum_{i=1}^{N} (g_i(\mathbf{w}))^2 \tag{9.18}$$

  in which each $g_i(\mathbf{w})$ is a smooth differentiable nonconvex function that measures the risk concentration of the i-th asset. The smaller the quantity $R(\mathbf{w})$ is, the more uniform the risk is distributed among n assets;[2]

- $F(\mathbf{w})$ is a convex function that represents some traditional preferences on the portfolio. For example, it can be the expected portfolio loss (e.g., $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w}$), the mean-variance trade-off of the portfolio loss (e.g., $F(\mathbf{w}) = -\boldsymbol{\mu}^T\mathbf{w} + \nu\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ where $\nu > 0$ is the trade-off parameter), or $F(\mathbf{w}) = 0$ when the goal is to distribute the risk only;

- $\lambda \geq 0$ is some trade-off parameter between the portfolio preference and risk concentration; and

- $\mathbf{w}^T\mathbf{1} = 1$ denotes the capital budget constraint and $\mathcal{W}$ is a convex set that denotes the investor's profiles, capital limitations, market regulations, etc.

This problem formulation is quite general to fit the previously stated specific formulations, for example, setting $g_i(\mathbf{w}) = w_i(\boldsymbol{\Sigma}\mathbf{w})_i - b_i\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$ and $\lambda = 0$ recovers the problem (9.14).

[2] In some problem formulations, the definition $\sum_{i,j=1}^{N} (g_{ij}(\mathbf{w}))^2$ is used where $g_{ij}(\mathbf{w})$ measures the difference between the risk contributions of assets i and j, for which the analytical approach derived in this paper still applies.

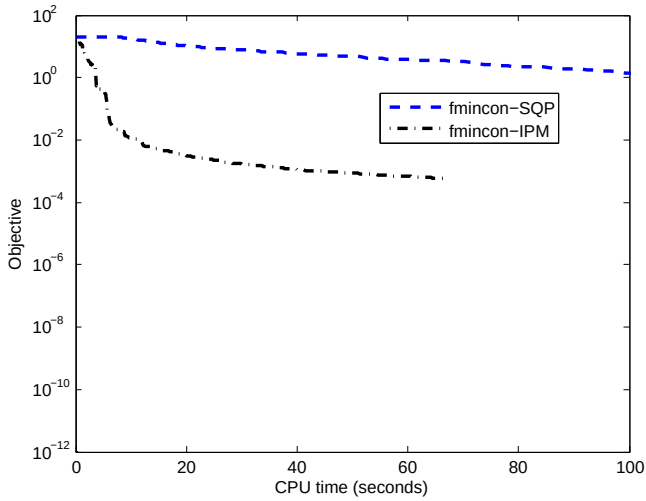

Figure 9.2: Performances of SQP and IPM.

Since each function $g_i(\mathbf{w})$ is highly nonconvex, the problem (9.17) is also nonconvex and hard to solve. In the literature, usually traditional off-the-shelf nonlinear optimization methods, like sequential quadratic programming (SQP) [153] and interior point methods (IPM) [35] built in the MATLAB function `fmincon`, are used in practice [14, 34, 131, 168, 90, 184]. However, for the nonconvex risk parity problem, in general they are time consuming and sometimes may not even converge globally [14, 90, 184]. This can be shown by a simple numerical example as follows.

Example 9.1. We set $N = 500$ and simulate the problem (9.14). The covariance matrix is randomly generated as $\mathbf{\Sigma} = \mathbf{V}\mathbf{V}^T$ where $\mathbf{V} = \texttt{rand}(\texttt{N}, \texttt{N})$. For simplicity and for illustrative purposes, we consider the long-only constraints, e.g., $\mathbf{w}^T\mathbf{1} = 1$ and $\mathbf{w} \geq \mathbf{0}$, and for this special case it is known that the optimal objective is zero [167].[3]

Figure 9.2 shows one typical realization of the performance of objective vs CPU time of the SQP and IPM methods built in the MATLAB

[3]More comprehensive numerical experiments based on both synthetic and real data can be found in [76].

function `fmincon`, and we have similar results for all the realizations. Basically, we observe that the SQP and IPM methods may either not even converge or converge to a unsatisfactory point very slowly. ■

9.3 SCRIP: An Efficient Numerical Solving Approach

Just as shown before, the general standard off-the-shelf numerical nonconvex nonlinear optimization methods, like SQP and IPM, are not efficient for nonconvex problems like (9.17).

To overcome this drawback, the authors of [76] explored the structure of the nonconvex part of $U(\mathbf{w})$, e.g., $R(\mathbf{w}) = \sum_{i=1}^{N} (g_i(\mathbf{w}))^2$, as follows. At the k-th iteration, the proposed method aims to solve

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \overbrace{\sum_{i=1}^{N} \left(g_i\left(\mathbf{w}^k\right) + \left(\nabla g_i\left(\mathbf{w}^k\right)\right)^T \left(\mathbf{w} - \mathbf{w}^k\right)\right)^2}^{P\left(\mathbf{w};\mathbf{w}^k\right) \triangleq} \\ & + \tfrac{\tau}{2} \left\|\mathbf{w} - \mathbf{w}^k\right\|_2^2 + \lambda F(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T \mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{9.19}$$

where $\tau > 0$ is the parameter for the regularization term. Here, the nonconvex term $R(\mathbf{w})$ is convexified by linearizing each $g_i(\mathbf{w})$ inside the square operation. The added proximal term $\left\|\mathbf{w} - \mathbf{w}^k\right\|_2^2$ is for convergence reasons [178].

The beauty of the approximation $P\left(\mathbf{w};\mathbf{w}^k\right)$ is that it is an easily computable quadratic convex function and has the same gradient as $R(\mathbf{w})$ at each iteration point $\mathbf{w}^k$:

$$\nabla P\left(\mathbf{w};\mathbf{w}^k\right)|_{\mathbf{w}=\mathbf{w}^k} = \nabla R(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^k}, \tag{9.20}$$

where $\nabla P\left(\mathbf{w};\mathbf{w}^k\right)$ denotes the partial gradient of $P\left(\mathbf{w};\mathbf{w}^k\right)$ with respect to the first argument $\mathbf{w}$.

Note that $P\left(\mathbf{w};\mathbf{w}^k\right)$ can be rewritten more compactly as

$$P\left(\mathbf{w};\mathbf{w}^k\right) = \left\|\mathbf{A}^k\left(\mathbf{w} - \mathbf{w}^k\right) + \mathbf{g}\left(\mathbf{w}^k\right)\right\|_2^2 \tag{9.21}$$

where

$$\mathbf{A}^k \triangleq \left[\nabla g_1\left(\mathbf{w}^k\right), \ldots, \nabla g_N\left(\mathbf{w}^k\right)\right]^T, \tag{9.22}$$

$$\mathbf{g}\left(\mathbf{w}^k\right) \triangleq \left[g_1\left(\mathbf{w}^k\right), \ldots, g_N\left(\mathbf{w}^k\right)\right]^T. \tag{9.23}$$

Then the problem (9.19) can be further rewritten as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \frac{1}{2}\mathbf{w}^T\mathbf{Q}^k\mathbf{w} + \mathbf{w}^T\mathbf{q}^k + \lambda F\left(\mathbf{w}\right) \\ \text{subject to} & \mathbf{w}^T\mathbf{1} = 1, \quad \mathbf{w} \in \mathcal{W}, \end{array} \tag{9.24}$$

where

$$\mathbf{Q}^k \triangleq 2\left(\mathbf{A}^k\right)^T\mathbf{A}^k + \tau\mathbf{I}, \tag{9.25}$$

$$\mathbf{q}^k \triangleq 2\left(\mathbf{A}^k\right)^T\mathbf{g}\left(\mathbf{w}^k\right) - \mathbf{Q}^k\mathbf{w}^k. \tag{9.26}$$

In general, under the assumption that $F\left(\mathbf{w}\right)$ is convex, for nonempty convex set $\overline{\mathcal{W}}$ (recall that $\overline{\mathcal{W}} = \left\{\mathbf{w}|\mathbf{w}^T\mathbf{1} = 1\right\} \cap \mathcal{W}$) and $\tau > 0$, problem (9.24) is strongly convex and can be solved by the existing efficient solvers (e.g., MOSEK [150], SeDuMi [189], SDPT3 [194], etc.). Moreover, if $F\left(\mathbf{w}\right)$ is linear or convex quadratic, and $\overline{\mathcal{W}}$ only contains linear constraints, problem (9.24) reduces to a QP.

Algorithm 7 summarizes the sequential solving approach and it is referred to as SCRIP (Successive Convex optimization for RIsk Parity portfolio) since it is based on a successive convex optimization method.

Algorithm 7 Successive Convex optimization for RIsk Parity portfolio (SCRIP).

Input: $k = 0$, $\mathbf{w}^0 \in \overline{\mathcal{W}}_1$, $\tau > 0$, $\{\gamma^k\} > 0$
Output: a stationary point of problem (9.17)

1: **repeat**
2: Solve (9.24) to get the optimal solution $\hat{\mathbf{w}}^k$
3: $\mathbf{w}^{k+1} = \mathbf{w}^k + \gamma^k\left(\hat{\mathbf{w}}^k - \mathbf{w}^k\right)$
4: $k \leftarrow k + 1$
5: **until** convergence

Based on the result of [178, Theorem 3], it can be shown [76] that under some technical assumptions and $\tau > 0$, $\gamma^k \in (0, 1]$, $\gamma^k \to 0$, $\sum_k \gamma^k = +\infty$, and $\sum_k \left(\gamma^k\right)^2 < +\infty$, then either Algorithm 7 converges in a finite number of iterations to a stationary point of (9.17) or every

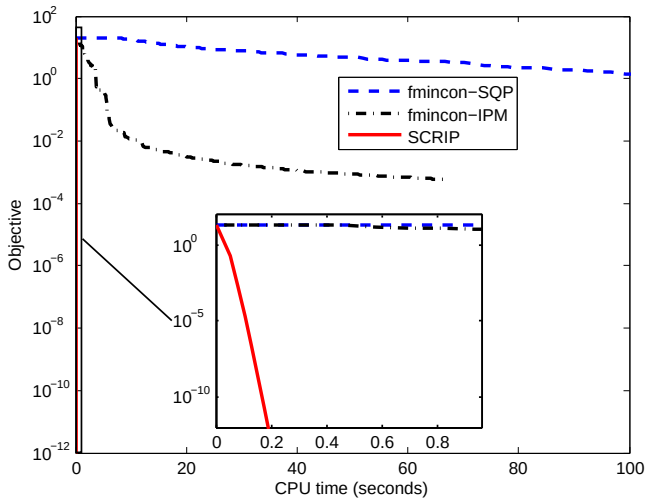

Figure 9.3: Performances of SQP, IPM, and SCRIP.

limit of $\mathbf{w}^k$ (at least one such point exists) is a stationary point of (9.17). There are also some interesting remarks on Algorithm 7.

Remark 9.2. Actually, one can easily have more algorithms by exploring two ideas: i) constructing a simpler QP approximation at each iteration, e.g., the quadratic coefficient matrix can be even diagonal, and ii) deriving some fast numerical updates when solving the inner QP approximation for some specific constraints. Here we do not explore them; however the interested reader is referred to [76, Algorithms 2-5] for detailed information. ■

Let us now revisit the previous Example 9.1 to conclude this chapter. Figure 9.3 shows the performance of objective vs CPU time of the existing SQP and IPM methods and the iterative SCRIP method. Clearly, we can see that SCRIP converges much more quickly and achieves a better objective value. This is also observed in more comprehensive numerical experiments, cf. [76].

Part III

Statistical Arbitrage (Mean-Reversion)

10

Statistical Arbitrage

Markowitz mean-variance portfolio optimization mainly follows the trends of the prices, i.e., the mean vector of the returns, to maximize the portfolio return with the portfolio risk under control, i.e., the portfolio variance is under a given threshold (cf. Part II, i.e., Chapters 5-9).

Conversely, the mean-reversion type of quantitative investment strategies aims at making profit based on the noisy fluctuations in the market prices regardless of the trends. This will be covered in this part, Part III, which contains only this chapter, Chapter 10. The underlying rough idea is to short-sell the (relatively) overvalued stocks and buy the (relatively) undervalued stocks, and hopefully a positive profit will be generated if their values converge. Such a type of quantitative investment strategy is referred to as "statistical arbitrage".

The detailed organization of this chapter is as follows. Section 10.1 explains the concept of cointegration and compares it with correlation. Sections 10.2-10.4 focus on introducing the "ancestor" of statistical arbitrage, that is, pairs trading. Section 10.2 studies different methods of discovering the potential pairs, once the potential pairs have been detected, Section 10.3 then tests whether they are indeed cointegrated or not, and Section 10.4 proceeds to checking the tractability and de-

signing trading rules. At the end, Section 10.5 generalizes pairs trading to statistical arbitrage.

10.1 Cointegration versus Correlation

To begin with, let us first recall the definition of cointegration introduced in Section 2.6: a time series is called integrated of order p, denoted as $I(p)$, if the time series obtained by differencing the time series p times is weakly stationary, while by differencing the time series $p-1$ times is not weakly stationary, and a multivariate time series is said to be cointegrated if it has at least one linear combination being integrated of a lower order.

Unlike correlation, which generally characterizes (relatively short-term) co-movements in log-returns, cointegration refers to (relatively long-term) co-movements in log-prices [61]. Correlation and cointegration are two related but different concepts. High correlation of log-returns does not necessarily imply high cointegration in log-prices, and neither does high cointegration in log-prices imply high correlation of log-returns [6].

Since cointegration is the key concept for statistical arbitrage and it is often confused with correlation, let us use some simple numerical examples to introduce cointegration and illustrate its relationship with correlation.

10.1.1 Log-Price Series with High Cointegration

Recall that in Example 2.1 in Section 2.6, we introduced cointegration based on a VECM model. Here, to understand the relationship between cointegration and correlation, we consider a more direct stochastic common trend model of two stocks, as follows [188]:

$$y_{1t} = \gamma x_t + w_{1t} \tag{10.1}$$

$$y_{2t} = x_t + w_{2t} \tag{10.2}$$

$$x_t = x_{t-1} + w_t, \tag{10.3}$$

where y_{1t} and y_{2t} are the log-prices, x_t is the stochastic common trend (which is a random walk), γ is a (positive) loading coefficient, and w_{1t},

w_{2t}, and w_t are i.i.d. errors that are independent of each other. For simplicity, suppose w_{1t}, w_{2t}, and w_t follow Gaussian distributions and their means are zero and variances are σ_1^2, σ_2^2, and σ^2, respectively. W.l.o.g., we also assume $\gamma = 1$ in this section.

Based on the model (10.1)-(10.3) and according to the definition of cointegration, we can first conclude that y_{1t} and y_{2t} are always cointegrated because the following linear combination

$$z_t \triangleq y_{1t} - y_{2t} = w_{1t} - w_{2t} \tag{10.4}$$

is stationary regardless of the values of σ_1^2, σ_2^2, and σ^2. In the literature, the above obtained stationary process z_t is referred to as "spread".[1] However, the correlation between the first order differences of y_{1t} and y_{2t}, i.e., the log-returns of the two stocks, is

$$\rho = \frac{\sigma^2}{\sqrt{\sigma^2 + 2\sigma_1^2}\sqrt{\sigma^2 + 2\sigma_2^2}} = \frac{1}{\sqrt{1 + \frac{2\sigma_1^2}{\sigma^2}}\sqrt{1 + \frac{2\sigma_2^2}{\sigma^2}}}, \tag{10.5}$$

which depends on the value of σ_1^2, σ_2^2, and σ^2. If $\sigma_1^2 \gg \sigma^2$ and/or $\sigma_2^2 \gg \sigma^2$, ρ is very close to zero and the correlation is very low. Therefore, high cointegration in log-prices does not necessarily imply high correlation in log-returns.

Example 10.1. Let us now study an illustrative numerical example. We set $\sigma_1 = \sigma_2 = 0.2$ and $\sigma = 0.1$ and randomly generate a sample path with 200 observations for each random process in the model (10.1)-(10.3).

Figure 10.1 shows the realization paths of y_{1t} (the blue curve), y_{2t} (the red curve), and their difference $y_{1t} - y_{2t}$ (the black curve). Clearly, we can see that y_{1t} and y_{2t} have a co-movement and indeed they are cointegrated since $y_{1t} - y_{2t}$ is stationary, as shown by the black curve.

However, the empirical correlation coefficient is 0.1124 (the theoretical value based on (10.5) is 0.1111) which means the correlation in the log-return series is quite low. Figure 10.2 shows the log-returns of stock 2 versus that of stock 1 and it verifies the low correlation since the points spread out in all directions.

[1]The spread z_t in (10.4) happens to have zero mean because the means of w_{1t} and w_{2t} are assumed to be zero. Generally, the spread mean is different from zero in practice.

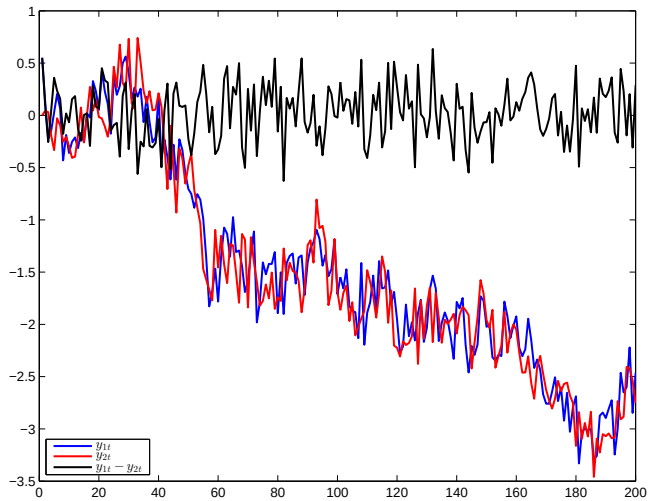

Figure 10.1: Some sample paths of the log-price series of the two stocks: the cointegration is high.

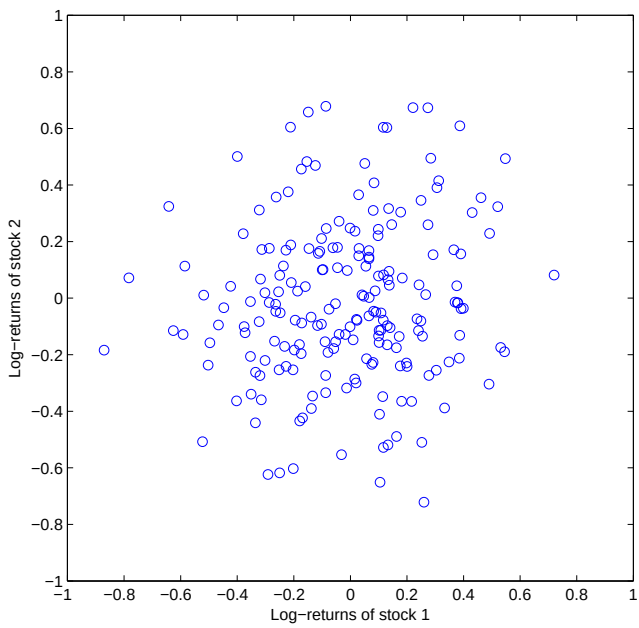

Figure 10.2: Log-returns of stock 2 versus that of stock 1: the correlation is low.

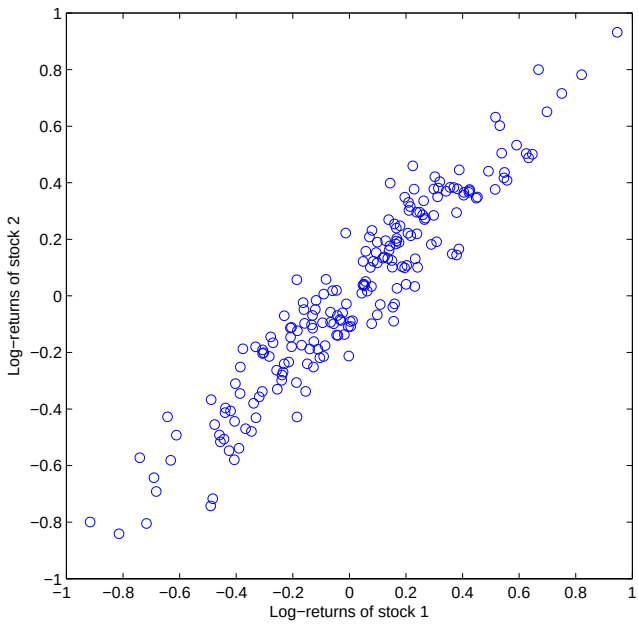

Figure 10.3: The correlation between the log-return series is high.

Therefore, we can conclude that high cointegration in log-prices series does not necessarily imply high correlation in log-return series.■

10.1.2 Log-Return Series with High Correlation

Let us still focus on the previous stochastic common trend model (10.1)-(10.3), but further consider a log-price series $\tilde{y}_{1t}$ as follows:

$$\tilde{y}_{1t} = y_{1t} + c_0 t, \tag{10.6}$$

that is, we add a temporal trend in the log-price series y_{1t}.

The correlation between the first order differences of $\tilde{y}_{1t}$ and y_{2t} is still given by (10.5); however, $\tilde{y}_{1t}$ and y_{2t} are no longer cointegrated since they will diverge increasingly as time goes by. In fact, this relationship is called cointegration with deterministic trend.

Example 10.2. Let us now consider a modification of Example 10.1. We set $\sigma_1 = \sigma_2 = 0.05$, and $\sigma = 0.3$ and generate 200 samples of y_{1t}

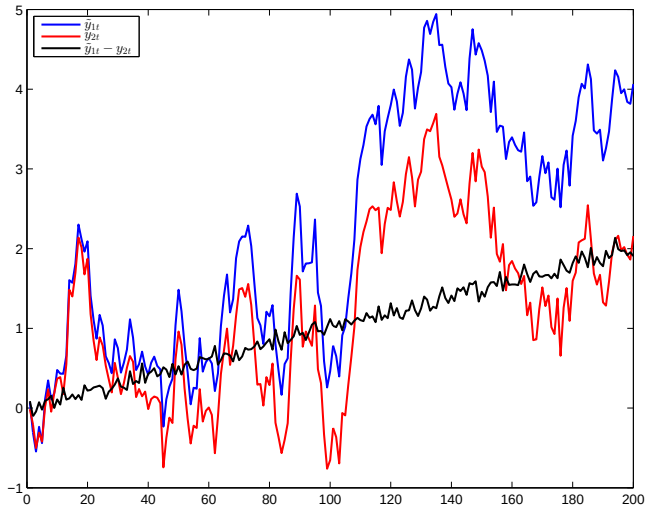

Figure 10.4: The log-price series are not cointegrated.

and y_{2t} according to the common stochastic model (10.1)-(10.3) and $\tilde{y}_{1t}$ according to (10.6) with $c_0 = 0.01$.

The empirical correlation between the first order differences of $\tilde{y}_{1t}$ and y_{2t} is 0.9504 (the theoretical value based on (10.5) is 0.9474), which means the corresponding log-return series are highly correlated. Figure 10.3 shows the log-returns of stock 2 versus that of stock 1 and it verifies the high correlation since the points fall closely along a straight line.

As to the cointegration, Figure 10.4 shows the realization paths of $\tilde{y}_{1t}$ (see the blue curve) and y_{2t} (see the red curve). Clearly, we can see that $\tilde{y}_{1t}$ and y_{2t} are not tied together by a stationary spread and indeed they are diverging increasingly since the spread $\tilde{y}_{1t} - y_{2t}$ keeps growing.

Thus, high correlation in log-return series does not necessarily imply high cointegration in log-prices series. ■

10.1.3 The Idea of Statistical Arbitrage Based on Cointegration

The idea behind statistical arbitrage is to short-sell the overvalued spread, and buy the undervalued stocks and unwind the position when the spread converges to its normal stage.

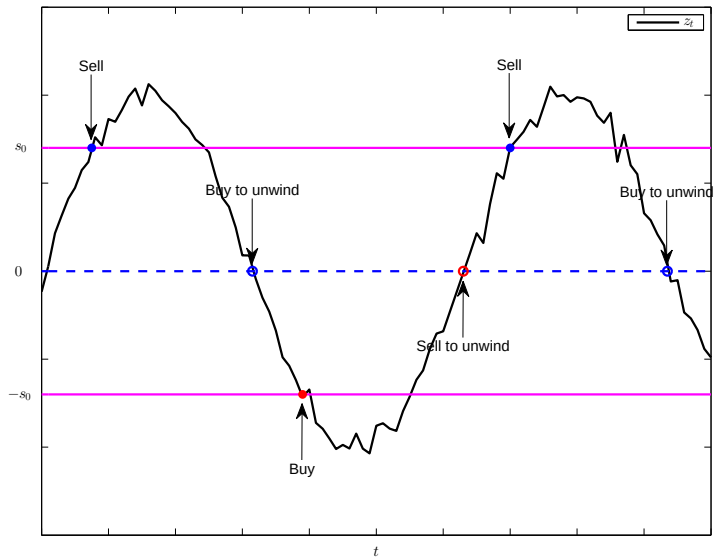

Figure 10.5: Investing on the mean-reversion spread.

Figure 10.5 pictorially shows the idea of investing on the mean-reversion spread. For example, suppose the stationary spread $z_t = y_{1t} - \gamma y_{2t}$ has mean zero, then one can buy the spread when it low at $z_t = -s_0$, i.e., buy one dollar stock 1 and short-sell γ dollar stock 2 as indicated in Figure 10.5 by the red point, and unwind the positions when the spread reverts to zero after i time steps, i.e., $z_{t+i} = 0$ as indicated in Figure 10.5 by the red circle. The resulting log-return of the strategy is $z_{t+i} - z_t = s_0$. Similarly, one can sell the spread when the spread is high at $z_t = s_0$, i.e., short-sell one dollar stock 1 and buy γ dollar stock 2 as indicated in Figure 10.5 by the red point, and unwind the positions when the spread reverts to zero again. The resulting log-return is also $z_t - z_{t+i} = s_0$.[2]

For illustrative purposes, let us revisit Example 10.1 and set $s_0 = 0.25$. Figure 10.6(a) shows the resulting mean-reversion stationary spread z_t and the buy and sell thresholds $\pm s_0$, Figure 10.6(b) reports the raw signaling for buying or shorting the spread, and Figure 10.6(c)

[2]For simplicity, we ignore the trading costs, e.g., brokerage fee, stamp fee, slippage, etc. in this chapter.

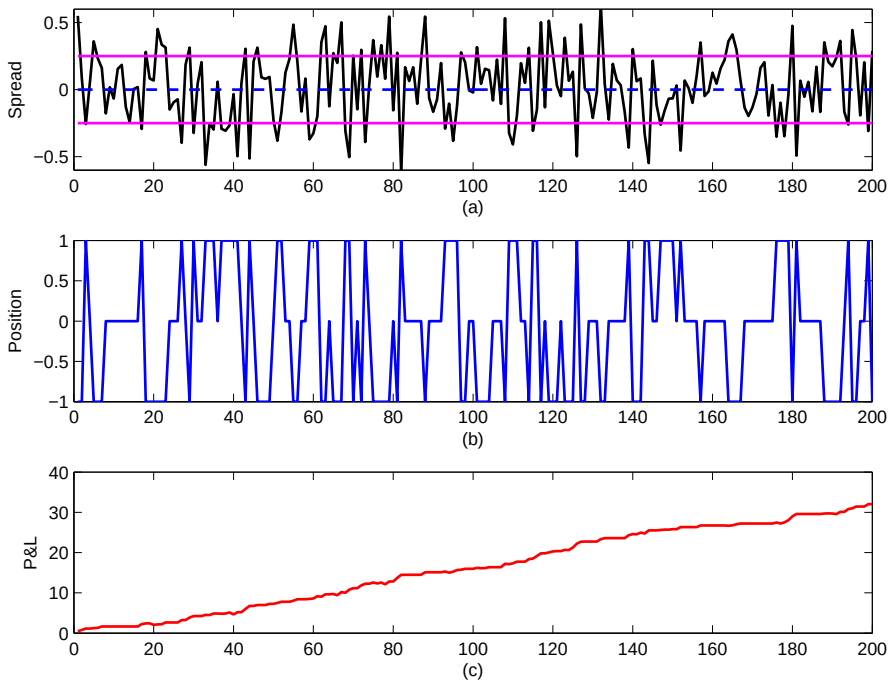

Figure 10.6: A simple example of statistical arbitrage: (a) the mean-reversion spread and the thresholds ± 0.25; (b) the positions, $+1$ and -1 means buy and sell the spread, respectively; (c) the cumulative profit and loss (P&L).

states the cumulative profit and loss (P&L)[3]. We can see that the statistical arbitrage does generate consistent positive profit from Figure 10.6(c). However, note that this is the in-sample result of a synthetic experiment without accounting for any trading costs. In practice, one needs to focus on the out-of-sample results and take the trading costs into consideration as well. Still, statistical arbitrage has generated significant positive profits in the real markets.

Fact 10.1. Pairs trading probably is the first practically implemented statistical arbitrage trading strategy. It was first invented in industry by a quantitative trading team led by the quant Nunzio Tartaglia in Morgan Stanley around the mid 1980s. Tartaglia's team enjoyed signif-

[3]We invest one dollar in each asset whenever we buy or sell the spread, and the P&L is computed as the cumulative summation of the profits and losses.

icant success in pairs trading in 1987. The team was disbanded in 1989 and the members joined various other trading firms. However, pairs trading became widely known. Until now, pairs trading has generated hundreds of millions of dollars in profits for large institutions or hedge funds, e.g., Morgan Stanley, Renaissance Technologies, D. E. Shaw & Co., etc. ■

In the following Sections 10.2-10.4, we first focus on pairs trading as the example to introduce the main steps of statistical arbitrage. In practice, pairs trading can be mainly decomposed into three steps [203]:

- Pairs selection: identify stock pairs that could potentially be cointegrated.
- Cointegration test: test whether the identified stock pairs are indeed cointegrated or not.
- Trading strategy design: study the spread dynamics and design proper trading rules.

In the literature, the papers focusing on pairs trading are usually categorized into different approaches [92, 161], namely minimum distance approach [85, 151, 9], stochastic approach [57, 50, 195], and cointegration approach [203, 125, 6]. However, most of them mainly focus on only one (or two) of the above three steps and do not conduct the other steps properly. Here, we prefer to review different papers following the above three steps structure.

Later in Section 10.5, we will consider more general statistical arbitrage among multiple stocks.

10.2 Pairs Selection

The markets usually contain thousands of stocks which can form millions of pairs. It is too computationally costly to check whether each pair is cointegrated or not. A more practical way is to define an easy and straightforward measure to preliminarily identify the most potentially cointegrated pairs and then focus on testing the cointegration of such identified pairs only.

10.2.1 Normalized Price Distance

Probably the most simple and straightforward measurement is the normalized price distance (NPD) [85, 9]:

$$\text{NPD} \triangleq \sum_{t=1}^{T} (\tilde{p}_{1t} - \tilde{p}_{2t})^2 \tag{10.7}$$

where the normalized price $\tilde{p}_{1t}$ of stock 1 is given by

$$\tilde{p}_{1t} = \prod_{i=1}^{t} (1 + R_{1i}) \tag{10.8}$$

with R_{1i} being the i-th simple return of stock 1. Actually, this criterion implicitly assume the cointegration coefficient between the log-price of two stocks equals 1, i.e., $\gamma = 1$. The normalized prices of the other stocks are defined similarly. Then one can easily compute the NPDs for all the possible pairs and select some pairs with the smallest NPDs as the potentially cointegrated pairs.

The authors of [85] conduct pairs trading as follows. First, they use the past 12 months daily data to construct pairs with the smallest NPDs. Once the pairs are formed, they simply buy one dollar in the undervalued stock and short-sell one dollar in the overvalued stock when the normalized prices diverge more than two standard deviations, and unwind the positions when the normalized prices cross later. After 6 months, the positions are forced to unwind regardless of whether the prices have crossed or not.

Later, a following paper [9] provides more out-of-sample numerical results and another one [151] incorporates a stop-loss trigger if the distance diverges too much to limit the potential huge losses. Also, since the ℓ_2-norm distance in (10.7) is too sensitive to outliers, it is also suggested to consider some robust distance measurements, e.g., ℓ_1-norm distance [92].

Actually, the methods in [85, 9, 151] are just some specific practical implementations and the cointegration test and trading strategy design steps are either ignored or not properly conducted.

10.2.2 Measurements Based on Stochastic Common Trend Model and Factor Model

Now let us revisit the stochastic common trend model (10.1)-(10.3) and do not assume $\gamma = 1$. Then, the log-returns can be decomposed into two components as follows:

$$r_{1t} = y_{1t} - y_{1,t-1} = \underbrace{\gamma w_t}_{\triangleq r^c_{1t}} + \underbrace{(w_{1t} - w_{1,t-1})}_{\triangleq r^s_{1t}} \tag{10.9}$$

$$r_{2t} = y_{2t} - y_{2,t-1} = \underbrace{w_t}_{\triangleq r^c_{2t}} + \underbrace{(w_{2t} - w_{2,t-1})}_{\triangleq r^s_{2t}} \tag{10.10}$$

where r^c_{1t} and r^c_{2t} are the log-returns due to the nonstationary stochastic common trend with $r^c_{1t} = \gamma r^c_{2t}$, and r^s_{1t} and r^s_{2t} are the log-returns due to the stationary components (and thus the cumulative summations of r^s_{1t} and r^s_{2t} are stationary).

Note that the factor model for stock i at time t has the form:

$$r_{it} = \boldsymbol{\pi}_i^T \mathbf{f}_t + \varepsilon_{it}, \tag{10.11}$$

where $\mathbf{f}_t$ is the factor which is the same for all the stocks, $\boldsymbol{\pi}_i$ is the vector of loading coefficients, and ε_{it} is the idiosyncratic noise.

The factor model (10.11) is a more general approach than the trend model (10.9)-(10.10) in the sense that if we further assume $\boldsymbol{\pi}_1 = \gamma \boldsymbol{\pi}_2$ and the cumulative summations of the specific noise component ε_{it}, i.e., $\sum_{k=1}^{t} \varepsilon_{ik}$, are stationary for all the stocks, then r^c_{1t}, r^c_{2t}, r^s_{1t}, and r^s_{2t} are modeled by $\boldsymbol{\pi}_1^T \mathbf{f}_t$ and $\boldsymbol{\pi}_2^T \mathbf{f}_t$, ε_{1t}, and ε_{1t}, respectively.

Based on the above connection, one can always first estimate the factor model parameters, e.g., the factor loading coefficient estimates $\hat{\boldsymbol{\pi}}_i$, the factor covariance matrix estimates $\hat{\boldsymbol{\Sigma}}_f$, and then define different measurements to efficiently select the potentially cointegrated pairs. For simplicity, we arbitrarily study the pair of stocks 1 and 2.

Normalized Factor Loadings Difference

The first idea is that, for a cointegrated pair, the log-returns due to the common trend should be proportional to each other, which means the

factor loading coefficients should be proportional to each other. Therefore, one can define the normalized factor loadings difference (NFLD) as follows [203]:

$$\text{NFLD} \triangleq \left\| \frac{\hat{\boldsymbol{\pi}}_1}{\|\hat{\boldsymbol{\pi}}_1\|_2} - \frac{\hat{\boldsymbol{\pi}}_2}{\|\hat{\boldsymbol{\pi}}_2\|_2} \right\|_2 \tag{10.12}$$

and then identify the pairs with the smallest NFLDs as the potentially cointegrated ones.

Correlation Between Log-Returns due to Common Trend

Since the log-returns due to the common trend should be proportional to each other, i.e., they should share the same direction, an alternative idea is to compute the correlation coefficient between them [203]:

$$\begin{aligned} |\rho| &= \left| \frac{\mathsf{Cov}(r_{1t}^c, r_{2t}^c)}{\sqrt{\mathsf{Var}(r_{1t}^c)\mathsf{Var}(r_{2t}^c)}} \right| = \left| \frac{\mathsf{Cov}(\hat{\boldsymbol{\pi}}_1^T \mathbf{f}_t, \hat{\boldsymbol{\pi}}_2^T \mathbf{f}_t)}{\sqrt{\mathsf{Var}(\hat{\boldsymbol{\pi}}_1^T \mathbf{f}_t)\mathsf{Var}(\hat{\boldsymbol{\pi}}_2^T \mathbf{f}_t)}} \right| \\ &= \left| \frac{\hat{\boldsymbol{\pi}}_1^T \hat{\boldsymbol{\Sigma}}_f \hat{\boldsymbol{\pi}}_2}{\sqrt{(\hat{\boldsymbol{\pi}}_1^T \hat{\boldsymbol{\Sigma}}_f \hat{\boldsymbol{\pi}}_1)(\hat{\boldsymbol{\pi}}_2^T \hat{\boldsymbol{\Sigma}}_f \hat{\boldsymbol{\pi}}_2)}} \right| = |\cos\theta|, \end{aligned} \tag{10.13}$$

where θ is the angle between the log-return series, and the potentially cointegrated pairs are the ones with θ being close to zero or, equivalently, $|\rho|$ being close to one.

Remark 10.1. Note that it is the log-returns due to the common trend only that are used to compute the absolute value of the correlation in (10.13). We should not use the overall log-returns, i.e., difference of log-prices, here because high correlation in the log-returns does not necessarily imply high cointegration in the log-prices, as we have shown in Section 10.1. ■

10.3 Cointegration Test

Once a potentially cointegrated pair, e.g., stocks 1 and 2, has been selected, the next step is to check whether they are cointegrated or

not. That is, we need to find out whether or not there exists a value of γ so that the spread

$$z_t = y_{1t} - \gamma y_{2t} \tag{10.14}$$

is stationary. Note that in practice the mean of spread z_t is not necessarily zero and in general γ may not be one.

To test for cointegration of two stocks, one of the most simple and direct methods is the Engle and Granger test [61] which usually contains two steps:

1. linearly regress the log-prices of one stock against that of the other stock and use the LS to compute the linear regression parameter; and
2. test whether the estimated residuals of the linear regression are stationary or not.

10.3.1 Linear Relationship

If z_t in (10.14) is stationary, it can be rewritten into the following form:

$$z_t = y_{1t} - \gamma y_{2t} = \mu + \varepsilon_t, \tag{10.15}$$

where μ represents the equilibrium value and ε_t is a zero mean stationary process that can be interpreted as the disturbance in the equilibrium [203]. The relationship (10.15) can be further rearranged as

$$y_{1t} = \mu + \gamma y_{2t} + \varepsilon_t, \tag{10.16}$$

which has the same expression as a linear regression. Then naturally the LS is employed to estimate the cointegration coefficient γ and the equilibrium value μ, and in fact, if y_{1t} and y_{2t} are $I(1)$ and are cointegrated, the estimates converge to the true values at the rate of number of observations [61].

Remark 10.2. In the literature, once the pairs are selected, many papers, e.g., [85, 151, 9, 57, 195], always long one and short the other with equal dollars so that the strategy is dollar neutral. Actually, this is equivalent to artificially fixing γ to be one and hoping that the spread

$$z_t = y_{1t} - y_{2t} \tag{10.17}$$

is stationary. Based on (10.17), we can have the following relationship

$$z_t - z_{t-1} = r_{1t} - r_{2t}, \tag{10.18}$$

which implies that the two stocks should have the same average return to ensure z_t is stationary. The number of pairs sharing the same average return in the real markets may be too few. Therefore, fixing $\gamma = 1$ may reduce the chance of identifying truly cointegrated pairs. ■

Remark 10.3. Please note that the log-prices are used for constructing cointegrated pairs here and the cointegration coefficients (i.e., the 1 and γ in (10.15) in front of log-prices) mean the invested dollars in each stock. To keep the invested dollar in each stock constant requires daily rebalancing since the price change may deviate the invested value from the constant level. One drawback of this method is that daily rebalancing may incur significant transaction costs and thus reduce total profit. One way to avoid this daily rebalancing via constructing cointegration pairs based on price series directly (as opposed to log-prices), in which case the cointegration coefficients (i.e., the estimated linar coefficients in front of prices) mean the numbers of shares invested in each stock. However, using prices directly may reduce the chance of cointegration since the noise in price is less symmetric than that in log-prices and the resulting cointegration spread may be less stationary compared that one obtained based on log-prices. Since the two approaches can be analyzed almost in the same way, for clarity of presentation and without loss of generality, we focus on pairs trading using log-prices only. ■

10.3.2 Cointegration and Strength of Mean-Reversion

The spread z_t is stationary if and only if the true residual series is stationary. In practice, we do not know the true values of the cointegration coefficient γ and the equilibrium value μ, and we cannot know the true residuals. However, the parameters γ and μ can be estimated by LS as shown before and we denote their estimates as $\hat{\gamma}$ and $\hat{\mu}$, respectively. Then we can use the estimated residuals

$$\hat{\varepsilon}_t = y_{1t} - \hat{\gamma} y_{2t} - \hat{\mu} \tag{10.19}$$

as the approximations of the true ones and test the stationarity of the estimated residuals instead.

Intuitively, without stepping into any statistical hypothesis test, an ad hoc method may be to use a high mean crossing rate as an indicator of mean-reversion: the higher the mean crossing rate is, the stronger the strength of mean-reversion is [203]. Even though it is simple and straightforward, it is not clear how to set the corresponding testing critical value for a given statistical significance value. This actually can be overcome by some statistical hypothesis tests as follows.

Dickey-Fuller (SF) Test

The DF test [49] is a hypothesis test for unit root nonstationarity. For any given time series x_t, the DF test first fits it to the following model:

$$\Delta x_t = \phi_0 + c_0 t + \phi_1 x_{t-1} + e_t, \tag{10.20}$$

where e_t denotes white noise, and then consider the null hypothesis $H_0 : \phi_1 = 0$ versus the the alternative hypothesis $H_a : \phi_1 < 0$. Here the null hypothesis means the time series x_t is a random walk, thus unit root nonstationary. The intuition here is that, if x_t is stationary, that is $\phi_1 < 0$, then it tends to revert to its long term mean; for example, supposing $\phi_0 = c_0 = 0$, a large value (or a small value) tends to be followed by a smaller value, that is, a negative change (or a large value, that is, a positive change, respectively).

The DF statistic is defined as the t-statistic of the LS estimate of ϕ_1 under the null hypothesis

$$\mathrm{DF} = \frac{\hat{\phi}_1}{\mathrm{std}(\hat{\phi}_1)}, \tag{10.21}$$

where $\hat{\phi}_1$ is the (expected) LS estimate and $\mathrm{std}(\hat{\phi}_1)$ is the standard deviation of the estimate [196]. Then given the statistical significance value, the null hypothesis is rejected if the DF statistic is less than a critical value.

Ideally, it is the true residuals ε_t that should be used in the above DF test to test it is stationary and thus the cointegration between the log-prices series y_{1t} and y_{2t}. However, as we mentioned, we can only

use the estimated residuals instead in practice and the critical value in the above DF test should be adjusted accordingly [130].

Augmented DF (ADF) Test

The ADF test is an extension by removing all the structural effects (autocorrelation) in the time series as follows:

$$\Delta x_t = \phi_0 + c_0 t + \phi_1 x_{t-1} + \sum_{i=1}^{p} \phi_{i+1} \Delta x_{t-i} + e_t, \tag{10.22}$$

where the null and alternative hypotheses are the same as that of the DF test. The remaining procedure of the ADF test for the cointegration test is the same as that of the DF test.

Remark 10.4. Earlier we investigated the Engle and Granger cointegration test based on two stocks. However, such a cointegration test has several drawbacks: the two-step cointegration test is sensitive to the ordering of variables in the regression; the first step "cointegration regression" may lead to spurious estimators if the bivariate series are not cointegrated, and it is not suitable for more than two stocks. An alternative method is the Johansen test, which tests the rank of the matrix $\mathbf{\Pi}$ (recall (2.49)) and obtains the corresponding MLE estimate in the VECM [107, 108] so one can get all the possible cointegration vectors. For more detailed discussions on different tests for cointegration, please refer to [93]. The good thing is that there already exist highly developed functions for the different tests, e.g., `egcitest` for the Engle and Granger test and `jcitest` for the Johansen test in MATLAB or packages egcm for the Engle and Granger test and urca for the Johansen test in R programming language. ■

For illustrative purposes, let us revisit previous Examples 10.1 and 10.2 to see how the simple Engle and Granger cointegration test works.

Example 10.3. Consider the generated sample paths of y_{1t} and y_{2t} in Example 10.1. We simply use the MATLAB function `egcitest` with default settings to test the cointegration. The estimated values are $\hat{\mu} = -0.0521$ and $\hat{\gamma} = 0.9492$, which are close to their true values $\mu = 0$

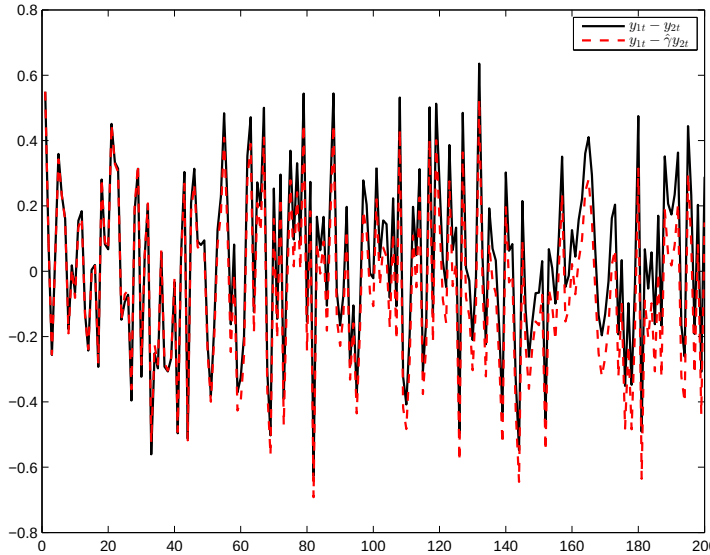

Figure 10.7: Engle and Granger cointegration test of Example 10.1.

and $\gamma = 1$. Then `egcitest` uses the ADF test to test the stationarity of the estimated residuals. Here, the ADF statistic computed by (10.21) is -14.009, less than the 5% significance level critical value -3.3669, thus the ADF test is to reject the null hypothesis and y_{1t} and y_{2t} are cointegrated. Figure 10.7 shows that the true and estimated residuals look close to each other and they look stationary. As for Example 10.2, the default settings of `egcitest` do not contain the time trend and it fails to reject the null hypothesis (note that the null hypothesis is that $\tilde{y}_{1t}$ and y_{2t} are not cointegrated). However, if we allow the time trend component estimation in `egcitest`, it produces the estimates of the parameters $\hat{\mu} = 0.0037$, $\hat{\gamma} = 0.9945$, and $\hat{c}_0 = 1$, which again are close to their true values $\mu = 0$, $\gamma = 1$, and $c_0 = 1$, respectively. Now the ADF statistic computed by (10.21) is -15.1876, less than the 5% significance level critical value -3.8283. Thus the ADF test rejects the null hypothesis and $\tilde{y}_{1t}$ and y_{2t} are not cointegrated but cointegrated with a deterministic trend. Similarly, Figure 10.8 shows that the true and estimated residuals look close to each other and they look stationary with a deterministic trend. ■

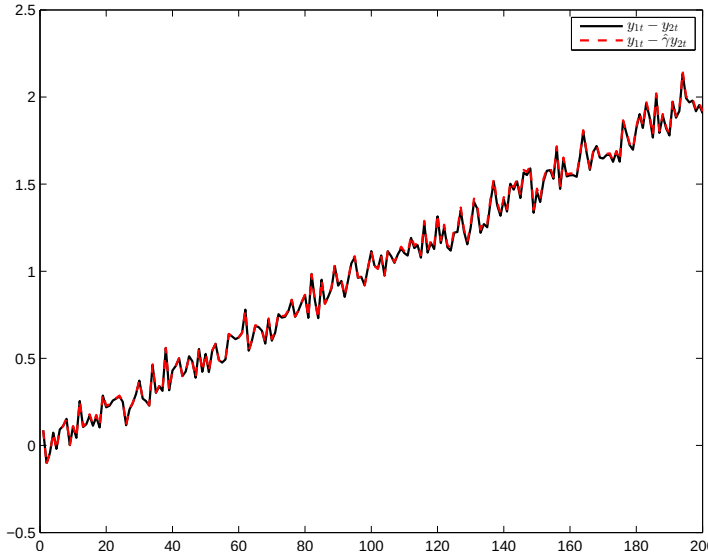

Figure 10.8: Engle and Granger cointegration test of the modification Example 10.2.

Furthermore, we consider one more simple example based on real data to show how to retrieve real data and how the cointegration test performs in practice.

Example 10.4. We focus on two main Chinese banks listed in the Hong Kong Stock Exchange, i.e., Industrial and Commercial Bank of China (ICBC, Code: 1398.HK) and China Construction Bank (CCB, Code: 0939.HK).

Figure 10.9 shows their adjusted log-prices from 01-Jan-2013 to 31-Dec-2015. The data is retrieved from Yahoo! Finance using the MATLAB function hist_stock_data[4]. We can see the two paths look really close to each other.

Indeed, the cointegration test shows that they are cointegrated and Figure 10.10 shows the (in-sample) spread (as indicated by the solid black line), its mean level (as indicated by the dashed blue line), and the thresholds deviating from the mean by one standard deviation (as

[4]http://www.mathworks.com/matlabcentral/fileexchange/18458-historical-stock-data-downloader/content/hist_stock_data.m

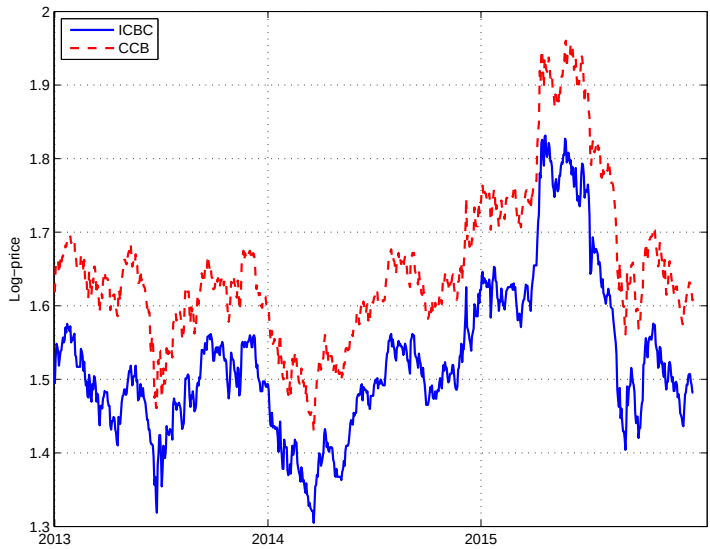

Figure 10.9: Log-prices of ICBC and CCB.

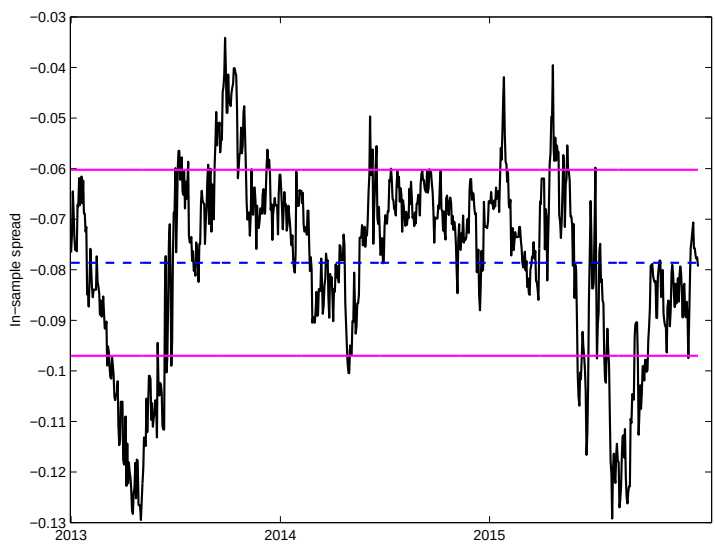

Figure 10.10: Log-prices of ICBC and CCB.

indicated by the two solid magenta lines). The MATLAB code is included in Appendix C. ■

10.4 Investing in Cointegrated Pairs

Once cointegrated pairs have been identified, there are different trading strategies that can be employed, for example, one can short the spread z_t when it is larger than its long term mean by a significant value (i.e., entry threshold) and unwind the position when the spread converges to a smaller value (i.e., exit threshold). The analysis of the optimal entry and exit thresholds for different rules is similar. For simplicity of presentation and w.l.o.g., we take the following trading rule: buy or sell the spread when it diverges from its long-term mean by s_0 and unwind the position when it passes through its mean. Thus, the key problem now is how to design the value of s_0 such that the total profit is maximized.

10.4.1 Optimal Threshold Value

Intuitively, a large threshold provides a large profit for each trade, albeit at a lower frequency, and a small threshold results in more frequent trades but a smaller profit for each trade. Both of these two extremes may not give the best total profit and an optimal threshold must be found between them. To compute the total profit, one needs to know two things: the profit of each single trade and the trading frequency. The former is simply the threshold value s_0 (based on the previous trading rule and note that log-prices are used here) and the latter is a monotonically decreasing function of the threshold value s_0 which is the key issue. Both the parametric and nonparametric approaches for computing the trading frequency function are introduced next.

Parametric Approach

The idea of the parametric approach is to fit the spread dynamic with a specific model based on which the trading frequency can be either theoretically or numerically efficiently computed. There are several different parametric models that satisfy the above requirement, e.g., the

white Gaussian noise model, mixture Gaussian model, ARMA model, and hidden Markov ARMA model [203].

For illustrative purposes, let us focus on the white Gaussian noise model and we further arbitrarily assume the noise is i.i.d. following a standard Normal distribution since otherwise we can always standardize the noise first.

The probability that a white Gaussian noise process at any time deviates above from the mean by s_0 or more is $1 - \Phi(s_0)$, where $\Phi(\cdot)$ is the c.d.f. of the standard Normal distribution. Therefore, in T steps we expect to have $T(1-\Phi(s_0))$ events greater than s_0 and the number of shorts is one half of that, i.e., $T(1 - \Phi(s_0))/2$, since the spread may need to cross the threshold s_0 again before it reverts to the mean level (cf. Figure 10.5). Similarly, we can get the number of buys is also $T(1 - \Phi(s_0))/2$, and the total number of trades is $T(1 - \Phi(s_0))$. For each trade, the profit is s_0 and then the total profit is $s_0 T(1 - \Phi(s_0))$.

Example 10.5. Let us use a simple numerical example to illustrate the idea. We first randomly generate $T = 70$ samples from the standard Normal distribution. The sample mean and variance are 0.1752 and 0.9928, respectively.

Figure 10.11(a) shows the true theoretical function $(1-\Phi(s_0))$ and the estimated one which is computed based on the sample mean and sample variance. Figure 10.11(b) shows the profit of each single trade, and Figure 10.11(c) shows the total profit. The maximum of the estimated total profit is achieved at the threshold $s_0 = 0.8$ which is close to the optimal threshold, i.e., the maximizer of the theoretical total profit, at $s_0 = 0.75$. ■

Nonparametric Approach

For the previous parametric approach, one always needs to calibrate a predefined model from the spread samples and then compute the trading frequency for any given trading threshold either theoretically or numerically. Is there an alternative way to find the trading frequency directly from an observed spread path? The answer is affirmative and it is the nonparametric approach [203].

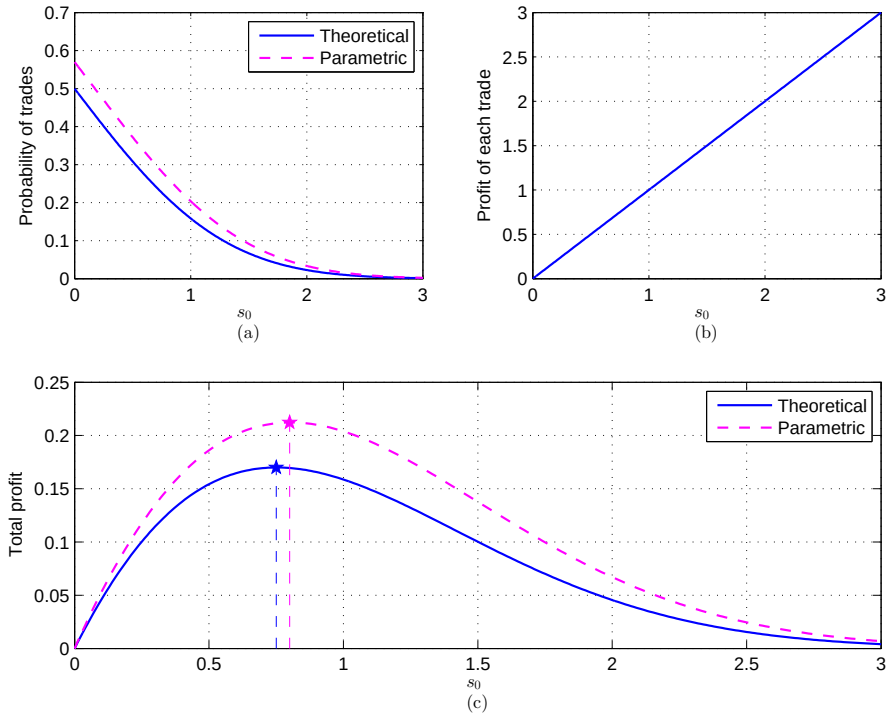

Figure 10.11: The computation of the total profit: parametric approach.

The idea is as follows: given a sample path of the spread realization, one can always compute the empirical trading frequency for any given threshold. That is, suppose the observed sample path has length T, and it is denoted as $z_1, z_2, \ldots, z_T$. We consider J discretized threshold values as $s_0 \in \{s_{01}, s_{02}, \ldots, s_{0J}\}$ and the empirical trading frequency for the threshold s_{0j} is

$$\bar{f}_j = \frac{\sum_{t=1}^{T} \mathbb{1}_{\{z_t > s_{0j}\}}}{T}. \tag{10.23}$$

However, in practice, the empirical values may not be a smooth function in the discretized thresholds and the resulted total profit function may be not accurate enough. To overcome this issue and obtain a smoother trading frequency function, one can employ the regularization idea,

which has been heavily used in Chapter 3, as follows:

$$\underset{\mathbf{f}}{\text{minimize}} \quad \sum_{j=1}^{J}(\bar{f}_j - f_j) + \lambda \sum_{j=1}^{J-1}(f_j - f_{j+1})^2, \tag{10.24}$$

where the second term is the regularization to induce smoothness and $\lambda > 0$ is the regularization parameter which can be chosen according to the rule in [203]. We can see that $\mathbf{f}$ is a smoothed version of the empirical trading frequency $\bar{\mathbf{f}}$. The problem (10.24) can be rewritten as a unconstrained convex QP:

$$\underset{\mathbf{f}}{\text{minimize}} \quad \|\bar{\mathbf{f}} - \mathbf{f}\|_2^2 + \lambda \|\mathbf{D}\mathbf{f}\|_2^2, \tag{10.25}$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(J-1)\times J} \tag{10.26}$$

is the first order difference matrix. Setting the derivative of the objective of (10.25) w.r.t. $\mathbf{f}$ to zero yields the optimal solution

$$\mathbf{f} = (\mathbf{I} + \lambda \mathbf{D}^T\mathbf{D})^{-1}\bar{\mathbf{f}}. \tag{10.27}$$

Similar to the parametric approach, let us use a simple example to illustrate the idea of the nonparametric approach.

Example 10.6. We use the same observations as Example 10.5. Figure 10.12(a) shows the empirical and regularized trading frequencies and Figure 10.12(c) shows the resulting total profit functions. We can see the total profit function given by the nonparametric approach is not smooth and is sensitive to the errors, for example, it gives the maximizer at the threshold $s_0 = 0.6$. This issue indeed is overcome by the regularization approach with $\lambda = 2^{4.5}$, as shown by the red star curve in Figure 10.12(c). The new maximizer now is $s_0 = 0.75$, which is exactly the same as the optimal theoretical threshold. ■

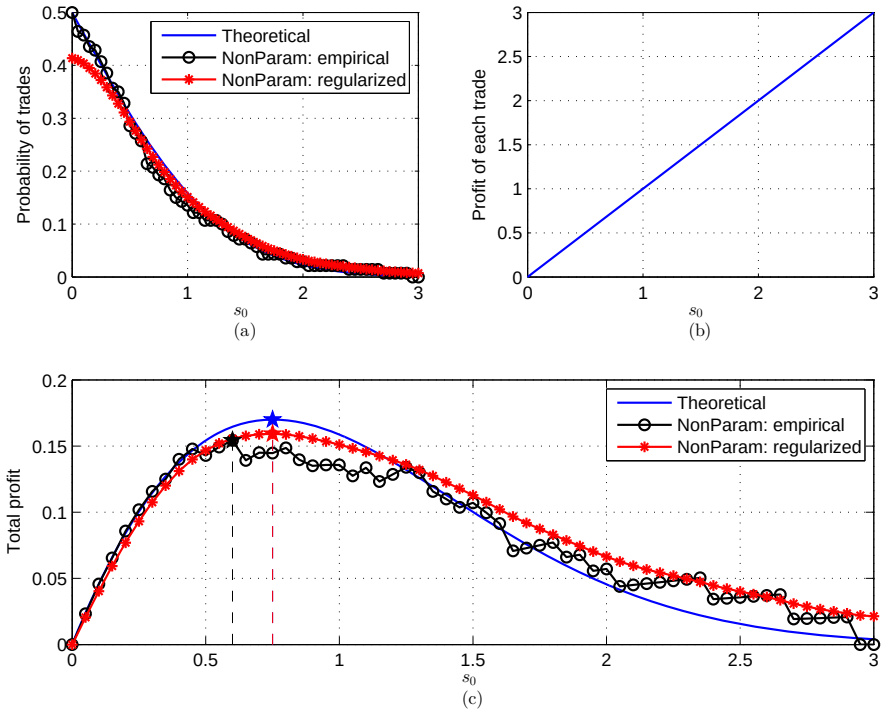

Figure 10.12: The computation of the total profit: nonparametric approach.

10.4.2 Holding Time

The above contents focused on designing the optimal threshold. Once the threshold has been designed, investors may also be interested in the corresponding holding time. For this purpose, we need to resort to some continuous-time mean-reversion models first.

One of the most widely used mean-reversion models is the Ornstein-Uhlenbeck process [114]:

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t, \tag{10.28}$$

where μ denotes long term mean, $\kappa > 0$ represents the strength of reversion, $\sigma > 0$ is the conditional volatility, and $\{W_t | t \geq 0\}$ is a standard Brownian motion. It can be shown that the long term variance is $\frac{\sigma^2}{2\kappa}$ which depends on both the conditional volatility σ and also the strength of reversion κ.

Intuitively, if the current value X_t is larger (or smaller) than the long term mean, i.e., $\mu - X_t < 0$ (or $\mu - X_t > 0$, respectively) since $\kappa > 0$, the change has a higher probability to be negative (or positive, respectively) and thus the process tends to revert to its long term mean. For example, if $X_0 = \mu + c\frac{\sigma}{\sqrt{2\kappa}}$, then the most likely time T it reverts to the long term mean μ is [57]

$$T = \frac{1}{\kappa} \log \left[1 + \frac{1}{2} \left(\sqrt{(c^2 - 3)^2 + 4c^2} + c^2 - 3 \right) \right]. \tag{10.29}$$

Thus, we can see that the larger κ is, the faster the process reverts from the deviation $c\frac{\sigma}{\sqrt{2\kappa}}$ (note that this deviation is measured as c multiples of the long term standard deviation $\frac{\sigma}{\sqrt{2\kappa}}$) to its long term mean.

In practice, the discretized model of (10.28) may be more useful and it turns out to be

$$x_{t+1} - x_t = \kappa(\mu - x_t)\tau + \sigma\sqrt{\tau}\varepsilon_{t+1}, \tag{10.30}$$

where $\tau > 0$ is the discretization period and ε_t is i.i.d. and follows the standard Normal distribution. It can be easily shown [57] that given $x_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$, the distribution of x_t is $x_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ where

$$\mu_t = \mu + (\mu_0 - \mu)(1 - b\tau)^t, \tag{10.31}$$

$$\sigma_t^2 = \frac{\sigma^2\tau}{1 - (1 - \kappa\tau)^2}[1 - (1 - \kappa\tau)^{2t}] + \sigma_0^2(1 - \kappa\tau)^{2t}, \tag{10.32}$$

and $\mu_t \to \mu$ and $\sigma_t^2 \to \frac{\sigma^2\tau}{1-(1-\kappa\tau)^2}$ provided that the discretization period $\tau > 0$ is small enough so that $|1 - \kappa\tau| < 1$. Note that when the discretization period τ goes to 0, the long term variance of the discretized model $\frac{\sigma^2\tau}{1-(1-\kappa\tau)^2} = \frac{\sigma^2}{2\kappa-\kappa^2\tau}$ goes to $\frac{\sigma^2}{2\kappa}$, which is the long term variance for the continuous model.

The relationship (10.30) can be rewritten as:

$$x_{t+1} = A + Bx_t + C\varepsilon_{t+1}, \tag{10.33}$$

where $A = \mu$, $0 < B = 1 - \kappa\tau < 1$, and $C = \sigma\sqrt{\tau}$. Actually, model (10.33) is also a univariate AR(1) model introduced in Section 2.5.1.

To infer a relatively smoother spread dynamic procedure, instead of using (10.33) to model a spread process directly, the authors of [57]

modeled the practical observed spread z_t as the underlying true spread x_t plus an observation noise, as follows:

$$z_t = x_t + Dw_t, \tag{10.34}$$

where $D > 0$ is a model parameter and the i.i.d. noise w_t follows the standard Normal distribution and are independent of the noise in (10.33).

In fact, the model (10.33)-(10.34) is a homogeneous Kalman filter, which has been widely used in various fields, including system control [111], signal processing [176], financial engineering [206], etc., its parameters can be easily estimated via the Expectation-Maximization algorithm, and the filtering procedure admits closed-form update steps under the Gaussian assumption.

Later, the paper [195] extends the work of [57] by considering a time varying Kalman filter since the market regime may change with time.

Again, let us consider a simple example to see how the above Kalman filter can help to improve the modeling of the spread dynamics.

Example 10.7. Here, for the state transition process (10.33) we artificially set $\tau = 1/252$, $\kappa = 150$, $\mu = 0$, and $\sigma = 0.02$, which means $A = \mu = 0$, $B = 1 - \kappa\tau = 1 - 150/252$, and $C = \sigma\sqrt{\tau} = 0.02/\sqrt{252}$. For the observation process (10.34) we set $D = 2C$.

Figure 10.13 shows the randomly generated realization paths of the underlying true spread x_t, the noisy observed spread z_t, and the Kalman filtering spread $\hat{x}_t$. Compared with z_t, we can see that the filtering spread $\hat{x}_t$ is relatively not as noisy and is closer to x_t. This is because, in principle, the Kalman filter can filter out the noise in the observed spread to some degree and the trading threshold designed based on the filtered spread process is relatively more reliable. ■

10.5 From Pairs Trading to Statistical Arbitrage

Now, let us move one step further from pairs trading based on only two stocks to statistical arbitrage for multiple stocks. The idea is still based on cointegration: try to construct some linear combinations of

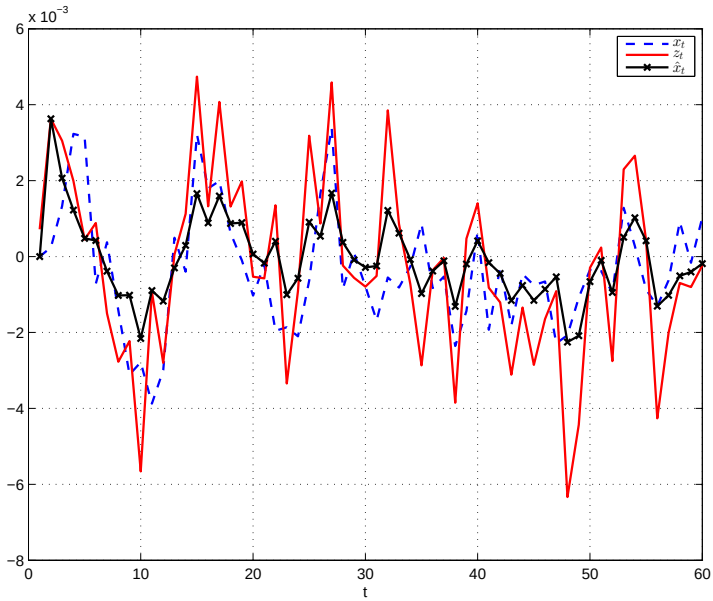

Figure 10.13: A realization of a spread based on the Kalman filter model (10.33)-(10.33): x_t is the underlying hidden spread, z_t is the observed noisy spread, and $\hat{x}_t$ is the Kalman filtering spread.

the log-prices of multiple (more than two) stocks such that the resulting spread series are mean-reversion processes.

10.5.1 Statistical Arbitrage Based on VECM

Until now we have explained the cointegration of only two stocks. As we have introduced the VECM before in Section 2.6, it is possible to find some cointegration components among multiple stocks. Actually, if we look at the VECM model (2.49) which is stated as follows:

$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \tilde{\boldsymbol{\Phi}}_1\mathbf{r}_{t-1} + \cdots + \tilde{\boldsymbol{\Phi}}_{p-1}\mathbf{r}_{t-p+1} + \mathbf{w}_t. \tag{10.35}$$

If $0 < \text{rank}(\boldsymbol{\Pi}) = r < N$, $\boldsymbol{\Pi}$ can be decomposed as

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T \tag{10.36}$$

and then each of the r components of $\boldsymbol{\beta}^T\mathbf{y}_t$ is stationary and thus a mean-reversion process.

Thus, following the procedure in Section 10.4, one can study the spread and find the optimal trading threshold for each component. Among all the cointegrated components, usually the one with strongest strength of mean-reversion is preferred in practice [46].

10.5.2 Statistical Arbitrage Based on Factor Models

Let us now introduce the second method based on factor models. First recall the factor model (10.11) for stock i at time t used in Section 10.2.2 as follows:

$$r_{it} = \boldsymbol{\pi}_i^T \mathbf{f}_t + \varepsilon_{it}, \tag{10.37}$$

where $\mathbf{f}_t$ is the factor which is the same for all the stocks, $\boldsymbol{\pi}_i$ is the vector of loading coefficients, and ε_{it} is the specific noise.

Then the idea of trading the mean-reversion pattern based on (10.37) is to first properly select some tradeable factors and then test whether the cumulative summations of the resulted specific noise ε_{it} are stationary. If positive, then one can define $z_{it} = \sum_{j=0}^{t}(r_{ij} - \boldsymbol{\pi}_i^T \mathbf{f}_j)$ as a spread. Some tradeable examples of $\mathbf{f}_t$ are the log-returns of the sector ETFs and/or that of several largest eigen-portfolios[5] [13].

Again, for each constructed cointegration component, one can study the spread and find the optimal trading threshold following the procedure in Section 10.4.

[5] An eigen-portfolio is a portfolio whose weight is a eigenvector of the covariance (or correlation) matrix of the stock returns.

11

Conclusions

This monograph has discussed the underlying connections between financial engineering and signal processing.

Part I has focused on financial modeling and order execution. The idea of decomposing a financial time series into a trend and noise components is the same as that of decomposing discrete-time signal series into useful signal and noise components; financial time series modeling is similar to filter modeling in signal processing, e.g., the ARMA model in financial engineering is the same as the pole-zero model in signal processing; the order execution problem of minimizing the execution cost is also similar to sensor scheduling in dynamic wireless sensor networks and power allocation problems in broadcast channels.

Part II has mainly explored the (robust) portfolio optimization. In fact, portfolio optimization is mathematically identical to beamforming/filter design and the robust techniques to handling those two problems are also the same, e.g., the shrinkage technique in financial engineering is exactly diagonal loading in beamforming design.

Part III has reviewed statistical arbitrage with three steps: pairs selection, cointegration test, and trading strategy design. It is interesting to see that some quantitative tools familiar to researchers in signal pro-

cessing and control theory, e.g., the Kalman filter, have been applied in financial engineering to improve the statistical arbitrage trading strategy.

Based on the detailed explorations in this entire monograph and in the above brief summary, we believe this monograph may serve as a comprehensive tutorial on financial engineering from a signal processing perspective. We hope it can help researchers in signal processing and communication societies as a starting point to access financial engineering problems more straightforwardly and systematically, and apply signal processing techniques to deal with appropriate financial problems.

Appendices

A

MATLAB Code of Example 3.1

```
clear all; clc; close all;
%% settings
% N: dim; T: #samples; OutT: #outliers
N = 2;
T = 40;
OutT = 4;
CovMatrix = zeros(N,N);
MeanVec = zeros(1, N);
OutMeanVec = [-2, 2];

for i = 1:N
    for j = 1:N
        CovMatrix(i,j) = (0.8)^abs(i-j);
    end
end

%% generate samples and outliers
SamPoints = mvnrnd(MeanVec, CovMatrix, T);
OutPoints = mvnrnd(OutMeanVec, CovMatrix, OutT);

%% data: samples + outliers
X = [SamPoints; OutPoints];

%% sample covariance matrix, or equivalently, the Gaussian
    MLE
CovNormal = X'*X./(T+OutT);

%% Cauchy MLE
CovCauchy = eye(N);
CovCauchyInv = inv(CovMatrix);

cvg = 0;
while (~cvg)
    w = (N+1) ./ (1 + diag(X*CovCauchyInv*X'));
    tmpCov = (X'*diag(w)*X) ./ (T + OutT);
```

```
    if (norm(CovCauchy - tmpCov, 'fro') ./ norm(tmpCov, 'fro
        ') < 1e-8)
        cvg = 1;
    else
        CovCauchy = tmpCov;
        CovCauchyInv = inv(CovCauchy);
    end
end

% get size c: solving Eq. (3.56) yields c =  0.4944.
fun = @(x,c,N) ((N+1)./(1+x./c) .* (x./c) .* chi2pdf(x,N));
cmin = 0.01; cmax = 20;
Tol_c = 1e-6;
while 1
    cc = (cmin + cmax) ./ 2;
    q = integral(@(x)fun(x,cc,N), 0, Inf);

    if q > N + Tol_c
        cmin = cc;
    elseif q < N - Tol_c
        cmax = cc;
    else
        break;
    end
end
CovCauchy = CovCauchy ./ cc;

%% plot results
RG = 4;
x1 = -RG:.2:RG;
x2 = -RG:.2:RG;
[X1,X2] = meshgrid(x1,x2);

figure()
% plot the sample points
hsam = plot(SamPoints(:,1),SamPoints(:,2), 'k+');
hold on;

% plot the outliers
hout = plot(OutPoints(:,1),OutPoints(:,2), 'rs', '
    MarkerFaceColor', 'r');

% plot the true shape
ALL_Points = mvnpdf([X1(:) X2(:)], MeanVec, CovMatrix);
ALL_points = reshape(ALL_Points,length(x2),length(x1));
```

```
[c, hTrue] = contour(x1, x2, ALL_points,[0.01], 'LineWidth'
    , 2, 'Color', 'k', 'LineStyle', ':');

% plot the shape based on Gaussian MLE
ALL_Points = mvnpdf([X1(:) X2(:)], MeanVec, CovNormal);
ALL_points = reshape(ALL_Points,length(x2),length(x1));
[c, hNormal] = contour(x1,x2,ALL_points,[0.01], 'LineWidth'
    , 2, 'Color', 'r', 'LineStyle', '-');

% plot the shape based on Cauchy MLE
ALL_Points = mvnpdf([X1(:) X2(:)], MeanVec, CovCauchy);
ALL_points = reshape(ALL_Points,length(x2),length(x1));
[c, hCauchy] = contour(x1,x2,ALL_points,[0.01],'LineWidth',
     2, 'Color', 'b', 'LineStyle','-.');

axis square
xlim([-RG, RG])
ylim([-RG, RG])
legend([hsam, hout, hTrue, hCauchy, hNormal], 'Samples', '
    Outliers', 'Oracle', 'MLE: Cauchy', 'MLE: Gaussian', '
    Location', 'SouthEast')
print('-depsc', ['Normal_vs_Cauchy_Cov'])
```

B

MATLAB Code of Figure 5.1

```
clear all; clc; close all;

%% initial settings
NumStocks = 3;
Sigma = eye(NumStocks);
mu = 0.5*[1 2 3]';
SigmaInv = inv(Sigma);

%% portfolio optimization: solution (5.8)
Lams = 2.^[-2:0.1:10];
NumLams = length(Lams);
meanVec = NaN(NumLams, 1);
stdVec = NaN(NumLams, 1);
onesNumStocks = ones(NumStocks,1);

for whichLam = 1:NumLams
   lam = Lams(whichLam);

   nu = (2*lam - onesNumStocks' * SigmaInv * mu) / (
       onesNumStocks' * SigmaInv * onesNumStocks);
   w = SigmaInv * (mu + nu * onesNumStocks) / (2*lam);

   meanVec(whichLam) = w'*mu;
   stdVec(whichLam) = sqrt(w'*Sigma*w);
end

%% Sharpe ratio portofilo, (5.13)
rf = 0.4;
wm = SigmaInv*(mu - rf*onesNumStocks);
wm = wm ./ sum(wm);
meanm = wm'*mu;
stdm = sqrt(wm'*Sigma*wm);
xx = linspace(0, 2, 400);
slope = (meanm - rf) ./ stdm;
yy = slope .* xx + rf;
```

```
%% plot the results
figure()
plot(stdVec, meanVec, 'k-','LineWidth',1.5);
hold on;
plot(xx, yy, 'b-','LineWidth',1.5);
xlim([0 1.7028]);
ylim([0 2.2172]);
text(-0.05,rf,'r_f')

% GMVP
scatter(min(stdVec), min(meanVec),25 ,'k','filled')
annotation('textarrow',[0.72,0.62]/(1.5480)
    ,[1-0.2,1-0.05]/(2.0156),...
          'String','Global minimum variance')

% Sharpe ratio
scatter(stdm, meanm,25 ,'b','filled')
annotation('textarrow',[stdm+0.1, stdm+0.005]/(1.5480),[
    meanm-0.28, meanm-0.125]/(2.0156),...
           'String','Maximum Sharpe ratio')

% Efficient frontier
annotation('textarrow',[1.2,1.05]/(1.7028)
    ,[1.35,1.55]/(2.2172),...
           'String','Efficient frontier')

% Capital market line
annotation('textarrow',[0.5, 0.57]/(1.7028),[1.5,
    1.05]/(2.2172),...
'String','Capital market line')

xlabel('Standard deviation')
ylabel('Expected return')

% remove ticks
set(gca,'xtick',[])
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])

print('-depsc','Efficient_Frontier')
```

C

MATLAB Code of Example 10.4

```
clear all; clc; close all;

%% retrive data
% the hist_stock_data function is available at
% http://www.mathworks.com/matlabcentral/fileexchange
    /18458-historical-stock-data-downloader/content//
    hist_stock_data.m
RealData = hist_stock_data('01012013', '31122015', '1398.HK
    ', '0939.HK', 'frequency', 'd');

%% process data
AllDates = sort(intersect(RealData(1).Date, RealData(2).
    Date));

[~, DateIdx] = intersect(RealData(1).Date, AllDates);
LogPrice1 = log(RealData(1).AdjClose(DateIdx));

[~, DateIdx] = intersect(RealData(2).Date, AllDates);
LogPrice2 = log(RealData(2).AdjClose(DateIdx));

% plot the log prices
figure()
NumDays = length(LogPrice1);
h1 = plot(1:NumDays, LogPrice1, 'b', 'LineWidth', 1.5);
hold on;
h2 = plot(1:NumDays, LogPrice2, 'r--', 'LineWidth', 1.5);
grid on;
legend([h1, h2], {'ICBC', 'CCB'}, 'location', 'NorthWest')
ylabel('Log-price')

DateIdxShow = find([1; diff(year(AllDates))]);
set(gca,'XTick',DateIdxShow)
set(gca, 'XTickLabel', datestr({AllDates{DateIdxShow}}, '
    yyyy'))
print('-depsc','Real_log_prices')
```

```
%% cointegration test
Y = [LogPrice1 LogPrice2];
[h, pValue, stat, cValue, reg] = egcitest(Y);

%% plot the in-sample spread
figure()
spread = Y * [1; -reg.coeff(2)];
hspread = plot(1:NumDays, spread, 'k-','LineWidth', 1.5);
hold on;
plot(1:NumDays, mean(spread) + std(spread) .* ones(NumDays
    ,1), 'm', 'LineWidth', 1.5)
plot(1:NumDays, mean(spread) - std(spread) .* ones(NumDays
    ,1), 'm', 'LineWidth', 1.5)
plot(1:NumDays, mean(spread) .* ones(NumDays,1), 'b--', '
    LineWidth', 1.5)
ylabel('In-sample spread')
DateIdxShow = find([1; diff(year(AllDates))]);
set(gca,'XTick',DateIdxShow)
set(gca, 'XTickLabel', datestr({AllDates{DateIdxShow}}, '
    yyyy'))
print('-depsc','Real_in_sample_spread')
```

Abbreviations

AR	Autoregressive.
ARCH	Autoregressive Conditional Heteroskedasticity.
ARMA	Autoregressive Moving Average.
CVaR	Conditional Value-at-Risk.
GARCH	Generalized Autoregressive Conditional Heteroskedasticity.
GMVP	Global Minimum Variance Portfolio.
GNE	Generalized Nash Equilibrium.
GNEP	Generalized Nash Equilibrium Problem.
i.i.d./I.I.D.	Independent and Identically Distributed.
IPM	Interior Point Methods.
LS	Least-Square.
MA	Moving Average.
MAP	Maximum A Posterior.
ML	Maximum Likelihood.
MLE	Maximum Likelihood Estimator.
MSE	Mean Squared Error.
MV	Minimum Variance.

NE	Nash Equilibrium.
NEP	Nash EquilibriumE Problem.
PCA	Principal Component Analysis.
PSD	Positive Semidefinite.
QCQP	Quadratically Constrained Quadratic Programming.
QP	Quadratic Programming.
RMT	Random Matrix Theory.
SAA	Sample Average Approximation.
SCA	Successive Convex Approximation.
SCM	Sample Covariance Matrix.
SCRIP	Successive Convex optimization for RIsk Parity portfolio.
SDP	Semidefinite Programming.
SDR	Semidefinite Programming Relaxation.
SINR	Signal-to-Interference-plus-Noise Ratio.
SNR	Signal-to-Noise Ratio.
SQP	Sequential Quadratic Programming.
SR	Sharpe Ratio.
VaR	Value-at-Risk.
VAR	Vector Autoregressive.
VARMA	Vector Autoregressive Moving Average.
VECM	Vector Error Correction Model.
VMA	Vector Moving Average.
w.r.t.	With Respect To.

Notation

Boldface lower-case letters denote column vectors, boldface upper-case letters denote matrices, lower-case italics denote scalars, and upper-case italics denote random scalar variables. For the financial time series, at time t, we use p_t to denote the price, $R_t \triangleq \frac{p_t - p_{t-1}}{p_{t-1}}$ to denote net return, $y_t \triangleq \log p_t$ to denote the log-price, $r_t \triangleq y_t - y_{t-1} = \log p_t - \log p_{t-1} = \log(1 + R_t)$ to denote the compound return or log-return, and w_t to denote the white noise.

$\propto$	Proportional to.
$\triangleq$	Defined as.
$\mathbf{A}^T, \mathbf{A}^H$	Transpose, conjugate transpose (Hermitian) of the matrix $\mathbf{A}$, respectively.
$\mathbf{A}^{-1}$	Inverse of the matrix $\mathbf{A}$.
$\mathbf{A}^\dagger$	Matrix Moore-Penrose pseudoinverse of the matrix $\mathbf{A}$.
a_i	The i-th entry of the vector $\mathbf{a}$.
$\mathbf{A}_{ij}$	The element of matrix $\mathbf{A}$ at the i-th row and j-th column.
$\mathbf{A}^{1/2}$	The principal square root of the matrix $\mathbf{A}$, i.e., $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.
$\text{Diag}\,(\mathbf{A})$	A diagonal matrix with diagonal elements equal to that of $\mathbf{A}$.
$\lvert\mathbf{A}\rvert$ or $\det(\mathbf{A})$	Determinant of the matrix $\mathbf{A}$.

$\mathrm{Tr}\,(\mathbf{A})$	Trace of the matrix $\mathbf{A}$.
$\lvert a\rvert$	Absolute value of the scalar a.
$\lVert\mathbf{a}\rVert_1$	ℓ_1-norm of the vector $\mathbf{a}$, i.e., $\lVert\mathbf{a}\rVert_1 \triangleq \sum_i \lvert a_i\rvert$.
$\lVert\mathbf{a}\rVert_2$	Euclidean norm (i.e., ℓ_2-norm) of the vector $\mathbf{a}$, i.e., $\lVert\mathbf{a}\rVert_1 \triangleq \sqrt{\mathbf{a}^T\mathbf{a}}$.
$\lVert\mathbf{A}\rVert_F$	Frobenius norm of matrix $\mathbf{A}$, i.e., $\lVert\mathbf{A}\rVert_F \triangleq \sqrt{\mathrm{Tr}\,(\mathbf{A}^T\mathbf{A})}$.
$\mathbf{I}$	Identity matrix with proper size. A subscript can be used to indicate the dimension as well.
$\mathbf{a} \geq \mathbf{b}$	Elementwise relation $a_i \geq b_i$.
$\mathbf{A} \succeq \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix.
$\mathbf{A} \succ \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is a positive definite matrix.
$\mathbb{R}, \mathbb{C}$	The set of real and complex numbers, respectively.
$\mathbb{R}^{m\times n}, \mathbb{C}^{m\times n}$	The set m-by-n matrices with real- and complex-valued entries, respectively.
$\mathbb{S}^n$	The set of symmetric n-by-n matrices $\mathbb{S}^n \triangleq \left\{\mathbf{X} \in \mathbb{R}^{n\times n} \middle\vert \mathbf{X} = \mathbf{X}^T\right\}$.
$\mathbb{S}^n_+$	The set of positive semidefinite n-by-n matrices $\mathbb{S}^n_+ \triangleq \left\{\mathbf{X} \in \mathbb{R}^{n\times n} \middle\vert \mathbf{X} = \mathbf{X}^T \succeq \mathbf{0}\right\}$.
$\mathbf{x}^\star$	The optimal solution $\mathbf{x}$ to a problem. The notation $\mathbf{x}$ denotes the vector form of all the variables of the problem.
$v^\star\,((\cdot))$	The optimal value of problem $(\cdot)$.
$\sim$	Distributed according to.
$\mathcal{N}\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\log\,(\cdot)$	Natural logarithm.
$\mathsf{E}\,[\cdot]$	Statistical expectation.
$\mathsf{Var}\,[\cdot]$	Statistical variance.
$\mathsf{Cov}\,[\cdot]$	Statistical covariance.
$[a]^+$	Positive part of a, i.e., $[a]^+ \triangleq \max\,(0, a)$.
sup, inf	Supremum and infimum.
$\cup, \cap$	Union and intersection.
$\nabla_{\mathbf{x}} f\,(\mathbf{x})$	Gradient of function $f\,(\mathbf{x})$ with respect to $\mathbf{x}$.

Acknowledgments

The work of Yiyong Feng and Daniel P. Palomar was supported by the Hong Kong Research Grants Council under research grants 16207814 and 16206315. Both the authors would like to thank the anonymous reviewer, whose comments have significantly contributed to improve the quality of this monograph.

References

[1] Y. Abramovich. Controlled method for adaptive optimization of filters using the criterion of maximum SNR. *Radio Engineering and Electronic Physics*, 26(3):87–95, 1981.

[2] Y. Abramovich and N. K. Spencer. Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages III–1105. IEEE, 2007.

[3] A. N. Akansu, S. R. Kulkarni, and D. M. Malioutov, editors. *Financial Signal Processing and Machine Learning*. Wiley-IEEE Press, 2016.

[4] I. Aldridge. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. John Wiley & Sons, 2013.

[5] C. Alexander. Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1758):2039–2058, 1999.

[6] C. Alexander, I. Giblin, and W. Weddington. Cointegration and asset allocation: A new active hedge fund strategy. *ISMA Centre Discussion Papers in Finance Series*, 2002.

[7] S. Alexander, T. F. Coleman, and Y. Li. Minimizing CVaR and VaR for a portfolio of derivatives. *Journal of Banking & Finance*, 30(2):583–605, 2006.

[8] R. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[9] S. Andrade, V. Di Pietro, and M. Seasholes. Understanding the profitability of pairs trading. *Unpublished working paper, UC Berkeley, Northwestern University*, 2005.

[10] K. Andriosopoulos, M. Doumpos, N. C. Papapostolou, and P. K. Pouliasis. Portfolio optimization and index tracking for the shipping stock and freight markets using evolutionary algorithms. *Transportation Research Part E: Logistics and Transportation Review*, 52:16–34, 2013.

[11] A. Ang and A. Timmermann. Regime changes and financial markets. Technical report, National Bureau of Economic Research, 2011.

[12] O. Arslan. Convergence behavior of an iterative reweighting algorithm to compute multivariate m-estimates for location and scatter. *Journal of Statistical Planning and Inference*, 118(1):115–128, 2004.

[13] M. Avellaneda and J.-H. Lee. Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782, 2010.

[14] X. Bai, K. Scheinberg, and R. Tutuncu. Least-squares approach to risk parity in portfolio selection. *Available at SSRN 2343406*, 2013.

[15] M. Bańbura, D. Giannone, and L. Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

[16] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006.

[17] J. E. Beasley, N. Meade, and T.-J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621–643, 2003.

[18] D. Bertsimas and A. W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1:1–50, 1998.

[19] D. Bianchi and A. Gargano. High-dimensional index tracking with cointegrated assets using an hybrid genetic algorithm. *Available at SSRN*, 1785908, 2011.

[20] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

[21] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[22] F. Black and R. Litterman. Asset allocation: combining investor views with market equilibrium. *The Journal of Fixed Income*, 1(2):7–18, 1991.

[23] F. Black and R. Litterman. Global asset allocation with equities, bonds, and currencies. *Fixed Income Research*, 2:15–28, 1991.

[24] F. Black and R. Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

[25] F. Black and M. Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, pages 637–654, 1973.

[26] D. Blamont and N. Firoozy. Asset allocation model. *Global Markets Research: Fixed Income Research*, 2003.

[27] Z. Bodie, A. Kane, and A. J. Marcus. *Investments*. Tata McGraw-Hill Education, 10th edition, 2013.

[28] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[29] T. Bollerslev. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The Review of Economics and Statistics*, pages 498–505, 1990.

[30] T. Bollerslev, R. F. Engle, and J. M. Wooldridge. A capital asset pricing model with time-varying covariances. *The Journal of Political Economy*, pages 116–131, 1988.

[31] J.-P. Bouchaud. Economics needs a scientific revolution. *Nature*, 455(7217):1181–1181, 2008.

[32] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[33] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[34] B. Bruder and T. Roncalli. Managing risk exposures using the risk budgeting approach. Technical report, University Library of Munich, Germany, 2012.

[35] R. H. Byrd, M. E. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

[36] N. A. Canakgoz and J. E. Beasley. Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*, 196(1):384–399, 2009.

[37] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[38] B. D. Carlson. Covariance matrix estimation errors and diagonal loading in adaptive arrays. *IEEE Transactions on Aerospace and Electronic Systems*, 24(4):397–401, 1988.

[39] Y. Chen, A. Wiesel, and A. O. Hero III. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.

[40] X. Cheng, Z. Liao, and F. Schorfheide. Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, 2016.

[41] T. F. Coleman, Y. Li, and J. Henniger. Minimizing tracking error while restricting the number of assets. *Journal of Risk*, 8(4):33, 2006.

[42] G. Connor. The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal*, 51(3):42–46, 1995.

[43] R. Couillet and M. McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120, 2014.

[44] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[45] H. Cox, R. M. Zeskind, and M. M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, 1987.

[46] A. d'Aspremont. Identifying small mean-reverting portfolios. *Quantitative Finance*, 11(3):351–364, 2011.

[47] R. A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 0:1–53, 2015.

[48] V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

[49] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

[50] B. Do, R. Faff, and K. Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 Financial Management Association European Conference*, 2006.

[51] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[52] C. Dose and S. Cincotti. Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1):145–151, 2005.

[53] B. Efron and C. Morris. Stein's estimation rule and its competitors-an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[54] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

[55] L. El Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, pages 543–556, 2003.

[56] Y. C. Eldar. Rethinking biased estimation: Improving maximum likelihood and the Cramér–Rao bound. *Foundations and Trends® in Signal Processing*, 1(4):305–449, 2008.

[57] R. J. Elliott, J. Van Der Hoek, and W. P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3):271–276, 2005.

[58] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons, 2009.

[59] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[60] R. F. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

[61] R. F. Engle and C. W. J. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276, 1987.

[62] R. F. Engle and K. F. Kroner. Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11(01):122–150, 1995.

[63] F. J. Fabozzi. *Robust Portfolio Optimization and Management*. Wiley, 2007.

[64] F. J. Fabozzi, S. M. Focardi, and P. N. Kolm. *Financial Modeling of the Equity Market: from CAPM to Cointegration*, volume 146. John Wiley & Sons, 2006.

[65] F. J. Fabozzi, S. M. Focardi, and P. N. Kolm. *Quantitative Equity Investing: Techniques and Strategies*. Wiley, 2010.

[66] E. F. Fama and K. R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465, 1992.

[67] E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[68] E. F. Fama and K. R. French. Size and book-to-market factors in earnings and returns. *Journal of Finance*, 50(1):131–155, 1995.

[69] E. F. Fama and K. R. French. Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51(1):55–84, 1996.

[70] E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, 18:25–46, 2004.

[71] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[72] J. Fan, L. Qi, and D. Xiu. Quasi-maximum likelihood estimation of garch models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, 32(2):178–191, 2014.

[73] J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012.

[74] B. Fastrich, S. Paterlini, and P. Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, pages 1–18, 2013.

[75] B. Fastrich, S. Paterlini, and P. Winker. Cardinality versus q-norm constraints for index tracking. *Quantitative Finance*, 14(11):2019–2032, 2014.

[76] Y. Feng and D. P. Palomar. SCRIP: Successive convex optimization methods for risk parity portfolios design. *IEEE Transactions on Signal Processing*, 63(19):5285–5300, Oct. 2015.

[77] Y. Feng, D. P. Palomar, and F. Rubio. Robust order execution under box uncertainty sets. In *Proceedings of the Asilomar Conference on Signals Systems, and Computers*, pages 44–48, Pacific Grove, CA, Nov. 2013.

[78] Y. Feng, D. P. Palomar, and F. Rubio. Robust optimization of order execution. *IEEE Transactions on Signal Processing*, 63(4):907–920, Feb. 2015.

[79] Y. Feng, F. Rubio, and D. P. Palomar. Optimal order execution for algorithmic trading: A CVaR approach. In *Proceedings of the IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 480–484, Jun. 2012.

[80] C. Floros. Modelling volatility using high, low, open and closing prices: evidence from four S&P indices. *International Research Journal of Finance and Economics*, 28:198–206, 2009.

[81] G. Frahm. *Generalized elliptical distributions: theory and applications*. PhD thesis, Universität zu Köln, 2004.

[82] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[83] W. Fung and D. A. Hsieh. Measuring the market impact of hedge funds. *Journal of Empirical Finance*, 7(1):1–36, 2000.

[84] M. B. Garman and M. J. Klass. On the estimation of security price volatilities from historical data. *Journal of Business*, pages 67–78, 1980.

[85] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.

[86] D. Goldfarb and G. Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[87] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[88] B. Graham and D. L. Dodd. *Security Analysis: Principles and Technique*. McGraw-Hill, 1934.

[89] B. Graham, J. Zweig, and W. E. Buffett. *The Intelligent Investor: A Book of Practical Counsel*. Harper & Row, 1973.

[90] T. Griveau-Billion, J.-C. Richard, and T. Roncalli. A fast algorithm for computing high-dimensional risk parity portfolios. *arXiv preprint arXiv:1311.4057*, 2013.

[91] R. G. Hagstrom. *The Warren Buffett Way: Investment Strategies of the World's Greatest Investor*. John Wiley & Sons, 1997.

[92] M. Harlacher. Cointegration based statistical arbitrage. *Department of Mathematics, Swiss Federal Institute of Technology, Zurich, Switzerland*, 2012.

[93] R. I. D. Harris. *Using Cointegration Analysis in Econometric Modelling*. Harvester Wheatsheaf, Prentice Hall, 1995.

[94] J. Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics and Econometrics of Securities Trading*. Oxford University Press, USA, 2007.

[95] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

[96] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[97] N. Hautsch. *Econometrics of Financial High-Frequency Data*. Springer Science & Business Media, 2011.

[98] S. Haykin and B. Van Veen. *Signals and Systems*. John Wiley & Sons, 2007.

[99] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.

[100] D. Huang, S. Zhu, F. J. Fabozzi, and M. Fukushima. Portfolio selection under distributional uncertainty: A relative robust CVaR approach. *European Journal of Operational Research*, 203(1):185–194, 2010.

[101] P. J. Huber. *Robust Statistics*. Springer, 2011.

[102] G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9(2):165–200, 2005.

[103] J. C. Hull. *Options, Futures, and Other Derivatives*. Pearson Education India, 9th edition, 2014.

[104] T. M. Idzorek. A step-by-step guide to the Black-Litterman model. *Forecasting Expected Returns in the Financial Markets*, page 17, 2002.

[105] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

[106] R. Jansen and R. Van Dijk. Optimal benchmark tracking with small portfolios. *The Journal of Portfolio Management*, 28(2):33–39, 2002.

[107] S. Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pages 1551–1580, 1991.

[108] S. Johansen. Likelihood-based inference in cointegrated vector autoregressive models. *Oxford University Press Catalogue*, 1995.

[109] I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

[110] P. Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(03):279–292, 1986.

[111] T. Kailath. *Linear Systems*, volume 1. Prentice-Hall Englewood Cliffs, NJ, 1980.

[112] A. Kammerdiner, A. Sprintson, E. Pasiliao, and V. Boginski. Optimization of discrete broadcast under uncertainty using conditional value-at-risk. *Optimization Letters*, 8(1):45–59, 2014.

[113] J. T. Kent and D. E. Tyler. Maximum likelihood estimation for the wrapped cauchy distribution. *Journal of Applied Statistics*, 15(2):247–254, 1988.

[114] Masaaki Kijima. *Stochastic Processes with Applications to Finance*. CRC Press, 2013.

[115] R. Kissell, M. Glantz, R. Malamut, and N.A. Chriss. *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk*. Amacom, 2003.

[116] G. M. Koop. Forecasting with medium and large bayesian VARs. *Journal of Applied Econometrics*, 28(2):177–203, 2013.

[117] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254, 2009.

[118] C. Lam, Q. Yao, and N. Bathia. Factor modeling for high dimensional time series. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 203–207. Springer, 2011.

[119] Z. M. Landsman and E. A. Valdez. Tail conditional expectations for elliptical distributions. *The North American Actuarial Journal*, 7(4):55–71, 2003.

[120] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

[121] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[122] O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[123] J. Li and P. Stoica. *Robust Adaptive Beamforming*. Wiley, 2006.

[124] W.-L. Li, Y. Zhang, A. M.-C. So, and M. Z. Win. Slow adaptive OFDMA systems through chance constrained programming. *IEEE Transactions on Signal Processing*, 58(7):3858–3869, 2010.

[125] Y.-X. Lin, M. McCrae, and C. Gulati. Loss protection in pairs trading through minimum profit bounds: A cointegration approach. *Advances in Decision Sciences*, 2006.

[126] R. B. Litterman. Forecasting with bayesian vector autoregressions–five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

[127] M. S. Lobo and S. Boyd. The worst-case risk of a portfolio. Technical report, 2000.

[128] D. G. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

[129] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2007.

[130] J. G. MacKinnon. Critical values for cointegration tests. Technical report, Queen's Economics Department Working Paper, 2010.

[131] S. Maillard, T. Roncalli, and J. Teïletche. The properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36(4):60–70, 2010.

[132] B. G. Malkiel. *A Random Walk Down Wall Street: The Time-tested Strategy for Successful Investing*. WW Norton & Company, 9th edition, 2007.

[133] D. G. Manolakis, V. K. Ingle, and S. M. Kogon. *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, volume 46. Artech House Norwood, 2005.

[134] D. Maringer and O. Oyewumi. Index tracking with constrained portfolios. *Intelligent Systems in Accounting, Finance and Management*, 15(1-2):57–71, 2007.

[135] H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[136] H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3(1-2):111–133, 1956.

[137] H. M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press, 1968.

[138] H. M. Markowitz, G. P. Todd, and W. F. Sharpe. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, volume 66. Wiley, 2000.

[139] R. A. Maronna. Robust M-Estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 01 1976.

[140] R. A. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Chichester., 2006.

[141] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2005.

[142] F. W. Meng, J. Sun, and M. Goh. Stochastic optimization problems with CVaR risk measure and their sample average approximation. *Journal of Optimization Theory and Applications*, 146(2):399–418, 2010.

[143] A. Meucci. *Risk and Asset Allocation*. Springer Science & Business Media, 2009.

[144] A. Meucci. Quant nugget 2: Linear vs. compounded returns–common pitfalls in portfolio management. *GARP Risk Professional*, pages 49–51, 2010.

[145] S. Moazeni, T. F. Coleman, and Y. Li. Optimal portfolio execution strategies and sensitivity to price impact parameters. *SIAM Journal on Optimization*, 20(3):1620–1654, 2010.

[146] S. Moazeni, T. F. Coleman, and Y. Li. Regularized robust optimization: the optimal portfolio execution case. *Computational Optimization and Applications*, 55(2):341–377, 2013.

[147] S. Moazeni, T. F. Coleman, and Y. Li. Smoothing and parametric rules for stochastic mean-CVaR optimal execution strategy. *Annals of Operations Research*, pages 1–22, 2013.

[148] D. Monderer and L. S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[149] R. A. Monzingo and T. W. Miller. *Introduction to Adaptive Arrays*. SciTech Publishing, 1980.

[150] MOSEK. The MOSEK optimization toolbox for MATLAB manual. Technical report, 2013.

[151] P. Nath. High frequency pairs trading with US treasury securities: Risks and rewards for hedge funds. *Available at SSRN 565441*, 2003.

[152] W. B. Nicholson, J. Bien, and D. S. Matteson. Hierarchical vector autoregression. *arXiv preprint arXiv:1412.5250*, 2014.

[153] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer Verlag, second edition, 2006.

[154] C. O'Cinneide, B. Scherer, and X. Xu. Pooling trades in a quantitative investment process. *Journal of Portfolio Management*, 32(4):33–43, 2006.

[155] K. J. Oh, T. Y. Kim, and S. Min. Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications*, 28(2):371–379, 2005.

[156] M. O'Hara. *Market Microstructure Theory*, volume 108. Blackwell Cambridge, MA, 1995.

[157] E. Ollila and D. E. Tyler. Regularized m-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070, Nov 2014.

[158] F. Pascal, Y. Chitour, and Y. Quek. Generalized robust shrinkage estimator and its application to stap detection problem. *IEEE Transactions on Signal Processing*, 62(21):5640–5651, 2014.

[159] A. F. Perold. The implementation shortfall: Paper versus reality. *Journal of Portfolio Management*, 14(3):4–9, 1988.

[160] A. Pole. *Statistical Arbitrage: Algorithmic Trading Insights and Techniques*, volume 411. John Wiley & Sons, 2011.

[161] H. Puspaningrum. *Pairs Trading Using Cointegration Approach*. PhD thesis, 2012.

[162] E. Qian. Risk parity portfolios: Efficient portfolios through true diversification. *Panagora Asset Management*, Sept. 2005.

[163] E. Qian. On the financial interpretation of risk contribution: Risk budgets do add up. *Journal of Investment Management*, 4(4):41, 2006.

[164] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[165] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[166] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[167] T. Roncalli. *Introduction to Risk Parity and Budgeting*. CRC Press, 2013.

[168] T. Roncalli and G. Weisang. Risk parity portfolios with risk factors. *Available at SSRN 2155159*, 2012.

[169] A. Roy, T. S. McElroy, and P. Linton. Estimation of causal invertible varma models. *arXiv preprint arXiv:1406.4584*, 2014.

[170] F. Rubio, X. Mestre, and D. P. Palomar. Performance analysis and optimal selection of large minimum variance portfolios under estimation risk. *IEEE Journal of Selected Topics in Signal Processing*, 6(4):337–350, 2012.

[171] D. Ruppert. *Statistics and Data Analysis for Financial Engineering*. Springer, 2010.

[172] S. Sarykalin, G. Serraino, and S. Uryasev. Value-at-risk vs. conditional value-at-risk in risk management and optimization. *Tutorials in Operations Research. INFORMS, Hanover, MD*, 2008.

[173] S. E. Satchell and B. Scherer. Fairness in trading: A microeconomic interpretation. *Journal of Trading*, 5:40–47, 2010.

[174] S. E. Satchell and A. Scowcroft. A demystification of the black–litterman model: Managing quantitative and traditional portfolio construction. *Journal of Asset Management*, 1(2):138–150, 2000.

[175] M. W. P. Savelsbergh, R. A. Stubbs, and D. Vandenbussche. Multi-portfolio optimization: A natural next step. In *Handbook of Portfolio Construction*, pages 565–581. Springer, 2010.

[176] L. L. Scharf. *Statistical Signal Processing*, volume 98. Addison-Wesley Reading, MA, 1991.

[177] Andrea Scozzari, Fabio Tardella, Sandra Paterlini, and Thiemo Krink. Exact and heuristic approaches for the index tracking problem with ucits constraints. *Annals of Operations Research*, 205(1):235–250, 2013.

[178] G. Scutari, F. Facchinei, Peiran Song, D. P. Palomar, and Jong-Shi Pang. Decomposition by partial linearization: Parallel optimization of multi-agent systems. *IEEE Transactions on Signal Processing*, 62(3):641–656, Feb. 2014.

[179] W. F. Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, pages 169–185, 1998.

[180] L. Shi and L. Xie. Optimal sensor power scheduling for state estimation of Gauss–Markov systems over a packet-dropping network. *IEEE Transactions on Signal Processing*, 60(5):2701–2705, May 2012.

[181] L. Shi and H. Zhang. Scheduling two Gauss–Markov systems: An optimal solution for remote state estimation under bandwidth constraint. *IEEE Transactions on Signal Processing*, 60(4):2038–2042, Apr. 2012.

[182] A. Silvennoinen and T. Teräsvirta. Multivariate GARCH models. In *Handbook of Financial Time Series*, pages 201–229. Springer, 2009.

[183] N. Y. Soltani, S.-J. Kim, and G. B. Giannakis. Chance-constrained optimization of OFDMA cognitive radio uplinks. *IEEE Transactions on Wireless Communications*, 12(3):1098–1107, 2013.

[184] I. Song. *New Quantitative Approaches to Asset Selection and Portfolio Construction*. PhD thesis, Columbia University, 2014.

[185] J. Song, P. Babu, and D. P. Palomar. Sparse generalized eigenvalue problem via smooth optimization. *IEEE Transactions on Signal Processing*, 63(7):1627–1642, April 2015.

[186] S. Song and P. J. Bickel. Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 2011.

[187] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.

[188] J. H. Stock and M. W. Watson. Testing for common trends. *Journal of the American statistical Association*, 83(404):1097–1107, 1988.

[189] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1-4):625–653, 1999.

[190] Y. Sun, P. Babu, and D. P. Palomar. Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Transactions on Signal Processing*, 62(19):5143–5156, 2014.

[191] Y. Sun, P. Babu, and D. P. Palomar. Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions. *IEEE Transactions on Signal Processing*, 63(12):3096–3109, June 2015.

[192] K. S. Tatsuoka and D. E. Tyler. On the uniqueness of S-functionals and m-functionals under nonelliptical distributions. *The Annals of Statistics*, pages 1219–1243, 2000.

[193] E. O. Thorp and S. T. Kassouf. *Beat the Market: A Scientific Stock Market System*. Random House New York, 1967.

[194] K.-C. Toh, M. J. Todd, and R. Tütüncü. On the implementation and usage of SDPT3–a MATLAB software package for semidefinite-quadratic-linear programming, version 4.0. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 715–754. Springer, 2012.

[195] K. Triantafyllopoulos and G. Montana. Dynamic modeling of mean-reverting spreads for statistical arbitrage. *Computational Management Science*, 8(1-2):23–49, 2011.

[196] R. S. Tsay. *Analysis of Financial Time Series*, volume 543. Wiley-Interscience, 3rd edition, 2010.

[197] R. S. Tsay. *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, 2013.

[198] D. N. C. Tse. Optimal power allocation over parallel Gaussian broadcast channels. In *Proceedings of the International Symposium on Information Theory*, page 27, 1997.

[199] A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information theory*, 1(1):1–182, 2004.

[200] R. H. Tütüncü and M. Koenig. Robust asset allocation. *Annals of Operations Research*, 132(1):157–187, 2004.

[201] D. E. Tyler. A distribution-free m-estimator of multivariate scatter. *The Annals of Statistics*, pages 234–251, 1987.

[202] D. E. Tyler. Statistical analysis for the angular central gaussian distribution on the sphere. *Biometrika*, 74(3):579–589, 1987.

[203] G. Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*, volume 217. John Wiley & Sons, 2004.

[204] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo. Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem. *IEEE Transactions on Signal Processing*, 51(2):313–324, 2003.

[205] S. A. Vorobyov, A. B. Gershman, Z. Q. Luo, and N. Ma. Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity. *Signal Processing Letters, IEEE*, 11(2):108–111, 2004.

[206] C. Wells. *The Kalman Filter in Finance*, volume 32. Springer Science & Business Media, 1996.

[207] A. Wiesel. Unified framework to regularized covariance estimation in scaled gaussian models. *IEEE Transactions on Signal Processing*, 60(1):29–38, 2012.

[208] C. Yang and L. Shi. Deterministic sensor data scheduling under limited communication resource. *IEEE Transactions on Signal Processing*, 59(10):5050–5056, Oct. 2011.

[209] D. Yang and Q. Zhang. Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, 73(3):477–492, 2000.

[210] Y. Yang, F. Rubio, G. Scutari, and D. P. Palomar. Multi-portfolio optimization: A potential game approach. *IEEE Transactions on Signal Processing*, 61(22):5590–5602, Nov. 2013.

[211] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

[212] M. Zhang, F. Rubio, and D. P. Palomar. Improved calibration of high-dimensional precision matrices. *IEEE Transactions on Signal Processing*, 61(6):1509–1519, 2013.

[213] M. Zhang, F. Rubio, D. P. Palomar, and X. Mestre. Finite-sample linear filter optimization in wireless communications and financial systems. *IEEE Transactions on Signal Processing*, 61(20):5014–5025, 2013.

[214] X. Zhang, H. V. Poor, and M. Chiang. Optimal power allocation for distributed detection over MIMO channels in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(9):4124–4140, Sept. 2008.