# Signal Processing and Machine Learning for Statistical Arbitrage in Finance: Models, Algorithms, and Analysis

by

Ziping ZHAO

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in the Department of Electronic and Computer Engineering

August 2019, Hong Kong

## Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Ziping ZHAO

August 2019

Signal Processing and Machine Learning for Statistical Arbitrage in Finance: Models, Algorithms, and Analysis

by

Ziping ZHAO

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Daniel P. PALOMAR (Thesis Supervisor)

Prof. Bertram E. SHI (Department Head)

Thesis Examination Committee

| | |
|---|---|
| 1. Prof. Daniel P. PALOMAR (Supervisor) | Department of Electronic and Computer Engineering |
| 2. Prof. Ross D. MURCH | Department of Electronic and Computer Engineering |
| 3. Prof. Roger S. CHENG | Department of Electronic and Computer Engineering |
| 4. Prof. Xinghua ZHENG | Department of Information Systems, Business Statistics and Operations Management |
| 5. Prof. Chee Wei TAN (External Examiner) | Department of Computer Science<br>City University of Hong Kong |

The Department of Electronic and Computer Engineering

August 2019

*To My Beloved Parents*

# Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Daniel P. Palomar, for his invaluable support and supervision over the past five years. It is him who introduced me to the world of research, provided me with the knowledge on optimization, and taught me the way of critical thinking. His outstanding attributes such as enthusiasm, dedication, as well as the amazing wisdom and technical insight have motivated me to a great extent. His warm encouragement and thoughtful guidance also led me through all the obstacles to the completion of my Ph.D. studies. It is really fortunate for me to be a student of such an excellent professor.

I would like to also thank Prof. Mingyi Hong for the helpful discussions and inspirations in our collaborations. I have always been impressed by his endless stream of ideas.

I am grateful to Prof. Ross D. Murch, Prof. Roger S. Cheng, Prof. Xinghua Zheng, and Prof. Chee Wei Tan for serving as my thesis examination committee members. I appreciate their precious time to read my thesis and to attend the thesis defense. Thanks also go to Prof. Danny H. Tsang for his help with my thesis proposal.

Special thanks go to our convex group members, Yang Yang, Mengyi Zhang, Yiyong Feng, Ying Sun, Junxiao Song, Konstantinos Benidis, Zhongju Wang, Tianyu Qiu, Licheng Zhao, Linlong Wu, Junyan Liu, Sandeep Kumar, Rui Zhou, Jiaxi Ying, Irtaza Khan, and Xiwen Wang. Thanks so much for their encouragement, care, and help. We have experienced a lot of wonderful times and we are like a big family.

My appreciation also goes to all my great friends in Lab 3116 for their friendship and help. They are Shibo Chen, Bin Qian, Lu Yang, Xi Peng, Liusha Yang, Baojian Zhou, Runfa Zhou, Lin Zhang, Xiaokang Wang, Xianghao Yu, Shuqi Chai, Lixiang Lian, Haobo Liang, and Nicolas Auguin.

I also wish to thank Prof. Gesualdo Scutari for his help and advice during my Ph.D. study. My gratitude also extends to my collaborators and friends which I met when I was visiting in the U.S. and they are Prof. Xiao Fu, Prof. Kejun Huang, Prof. Tianyi Chen, Prof. Yanning Shen, Prof. Hoi-To Wai, Dr. Songtao Lu, Dr. Gang Wang, Mr. Haoran Sun, Mr. Xiangyi Chen, and Prof. Lianbi Li.

Last but not least, I would like to thank my parents Xiaozhong Zhao and Fangcui Wang, and my whole family for their unconditional love and support.

*Ziping Zhao*

*August 2019*

# Table of Contents

# List of Figures

# List of Tables

Signal Processing and Machine Learning for Statistical Arbitrage in Finance: Models, Algorithms, and Analysis

by Ziping ZHAO

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

# Abstract

Statistical arbitrage, also known as pairs trading, is an important investment strategy in the financial markets. It is usually involved in a serial stages, say, assets selection, model estimation, portfolio design, and mean reversion trading. In this thesis, we will focus on the model estimation and portfolio design parts.

In econometrics and finance, the vector error correction model, or more generally the reduced-rank regression (RRR) model, is an important regression model for cointegration analysis, which is used to estimate the long-run equilibrium variable relationships. In this thesis, we will first study the efficient estimation of sparse RRR models. The traditional analysis and estimation methodologies assume the underlying Gaussian distribution but, in practice, heavy-tailed data and outliers can lead to the inapplicability of these methods. In this thesis, we propose a robust model estimation method based on the Cauchy distribution to deal with the robust estimation of VECM. Efficient algorithms based on the majorization-minimization method is applied to solve these proposed nonconvex problems. The performance of this algorithm is shown through numerical simulations.

The RRR in fact defines a cointegration space for the investment. After finding such a cointegration space, the following step is how to find an investment portfolio within this space. This thesis also considers the mean-reverting portfolio design problem arising from statistical arbitrage. This problem is formulated by optimizing some criteria characterizing the mean-reversion strength of the portfolio and taking into consideration the variance of the portfolio and an investment budget constraint or an investment leverage constraint at the same time. To deal with these problems, efficient algorithms based on nonconvex optimization methods like the majorization-minimization method or the successive convex approximation method are proposed. Numerical results show that our proposed mean-reverting portfolio design methods can show superior performance in the markets.

# Abbreviations

| | |
|---|---|
| RRR | reduced-rank regression |
| SRRR | sparse reduced-rank regression |
| AltMin | alternating minimization |
| EM | expectation-maximization |
| MM | majorization-minimization |
| VECM | vector error correction model |
| MRP | mean-reverting portfolio |
| S&P 500 | Standard & Poor's 500 |
| ETF | exchange-traded fund |
| DJIA | Dow Jones industrial average |
| OLS | ordinary least squares |
| SDR | semidefinite relaxation |
| QCQP | quadratic constrained quadratic programming |
| QCLP | quadratic constrained linear programming |
| GEVP | generalized eigenvalue problem |
| GTRS | generalized trust region subproblem |
| P&L | profit and loss |
| ROI | return on investment |
| SR | Sharp ratio |
| BP | basis point |
| SCA | successive convex approximation |
| MR | mean-reversion |
| ADMM | alternating direction method of multipliers |

# Notations

| | |
|---|---|
| $a$, $A$ | scalar |
| $\mathbf{a}$ | vector |
| $\mathbf{A}$ | matrix |
| $(\cdot)^T$ | transpose |
| $(\cdot)^*$ | conjugate |
| $(\cdot)^H$ | conjugate transpose |
| $(\cdot)^\dagger$ | Moore-Penrose pseudoinverse |
| $a_i$ | $i$-th entry of $\mathbf{a}$ |
| $a_{i,j}$ | ($i$-th, $j$-th) element of $\mathbf{A}$ |
| $\mathbf{A}(i)$, $\mathbf{a}_i$ | $i$-th column vector of $\mathbf{A}$ |
| $\mathbf{A} \succeq \mathbf{B}$ ($\mathbf{A} \succ \mathbf{B}$) | $\mathbf{A} - \mathbf{B}$ is a positive semidefinite (positive definite) matrix |
| rank$(\cdot)$ | rank |
| $\lvert \cdot \rvert$ | absolute value or cardinality of a set |
| tr$(\cdot)$ | trace |
| $\otimes$ | Kronecker product |
| vec$(\cdot)$ | vectorization |
| $\mathbf{1}$ | all-one vector |
| $\mathbf{I}$ | identity matrix |
| $\lVert\cdot\rVert_2$ | spectrum norm |
| $\lVert\cdot\rVert_F$ | Frobenius norm |
| $\mathbb{E}(\cdot)$ | expectation |
| $\mathcal{O}(\cdot)$ | big-O notation |
| Pr$(\cdot)$ | probability |
| $\mathbb{Z}$ | non-negative integers |
| $\mathbb{N}$ | natural numbers |
| $\mathbb{R}$ | real numbers |
| $\mathbb{R}_+$ | nonnegative real numbers |
| $\mathbb{R}^N$ | $N$-dimensional real vectors |
| $\mathbb{C}$ | complex numbers |
| $\mathbb{S}^K$ | $K \times K$-dimensional symmetric matrices |

# Chapter 1

# Introduction

The main purpose of this dissertation is on signal processing and machine learning methods for statistical arbitrage in finance. Nowadays, financial engineering has been an active research area which receives extensive attention and interest and statistical arbitrage as a risk neutral strategy becomes more and more popular in the financial industry.

## 1.1 Motivation

Statistical arbitrage, also known as pairs trading, is an important investment strategy in the financial markets. It is usually involved in a series of stages, say, assets selection, model estimation, portfolio design and mean reversion trading. In this thesis, we will focus on the model estimation and portfolio design parts.

### 1.1.1 Efficient Estimation of Sparse Reduced-Rank Regression Model

In this chapter, the estimation problem for sparse reduced-rank regression (SRRR) model is considered. The SRRR model is widely used for dimension reduction and variable selection with applications in signal processing, econometrics, etc. The problem is formulated to minimize the least squares loss with a sparsity-inducing penalty considering an orthogonality constraint. Convex sparsity-inducing functions have been used for SRRR in literature. In this work, a nonconvex function is proposed for better sparsity inducing. An efficient algorithm is developed based on the alternating minimization (or projection) method to solve the nonconvex optimization problem. Numerical simulations show that the proposed algorithm is much more efficient compared to the benchmarks and the nonconvex function can result in a better estimation accuracy.

### 1.1.2 Robust Estimation of Sparse Vector Error Correction Model

In econometrics and finance, the vector error correction model (VECM) is an important time series model for cointegration analysis, which is used to estimate the long-run equilibrium variable relationships. The traditional analysis and estimation methodologies assume the underlying Gaussian distribution but, in practice, heavy-tailed data and outliers can lead to the inapplicability of these methods. In this thesis, we propose a robust model estimation method based on the Cauchy distribution to tackle this issue. In addition, sparse cointegration relations are considered to realize feature selection and dimension reduction.

### 1.1.3 Mean-Reverting Portfolio Design With A Budget Constraint

This chapter considers the mean-reverting portfolio (MRP) design problem arising from statistical arbitrage (a.k.a. pairs trading) in the financial markets. It aims at designing a portfolio of underlying assets by optimizing the mean reversion strength of the portfolio, while taking into consideration the portfolio variance and an investment budget constraint. Several specific design problems are considered based on different mean reversion criteria. Efficient algorithms are proposed to solve the problems. Numerical results on both synthetic and market data show that the proposed MRP design methods can generate consistent profits and outperform the traditional design methods and the benchmark methods in the literature.

### 1.1.4 Mean-Reverting Portfolio Design With A Leverage Constraint

In this chapter, the optimal MRP design problem is studied under an investment leverage constraint representing the total investment positions on the underlying assets. A general problem formulation is proposed by considering the design targets subject to a leverage constraint. To solve the problem, a unified optimization framework based on the successive convex approximation method is developed. The superior performance of the proposed formulation and the algorithms are verified through numerical simulations on both synthetic data and real market data.

## 1.2 Outline of the Dissertation

In general terms, the focus of this dissertation is on signal processing and machine learning methods for statistical arbitrage in finance.

- Chapter 1, the present chapter, gives the motivation of each work, outline, and contributions of this dissertation.
- Chapter 2 considers the efficient estimation of sparse reduced-rank regression model via nonconvex optimization.
- Chapter 3 considers the robust maximum likelihood estimation of vector error correction model problem.
- Chapter 4 studies the mean-reverting portfolio design problem with a budget constraint.
- Chapter 5 studies the mean-reverting portfolio design problem with a leverage constraint.
- At last, Chapter 6 summarizes the main obtained results and highlights the future lines of work to conclude this dissertation.

## 1.3 Research Contributions

Detailed contributions are listed as follows.

### Chapter 2

This chapter considers the the efficient estimation of sparse reduced-rank regression model via nonconvex optimization.

The main output of this work is one journal paper and one conference paper:

- Ziping Zhao and Daniel P. Palomar, “Efficient Sparse Reduced Rank Regression via Nonconvex Optimization,” in preparation.
- Ziping Zhao and Daniel P. Palomar, “Sparse Reduced-Rank Regression with Nonconvex Regularization,” in Proc. of the 20th IEEE Statistical Signal Processing Workshop (SSP 2018), Freiburg, Germany, June 10-13, 2018.

## Chapter 3

This chapter considers the robust maximum likelihood estimation of vector error correction model problem.

The main output of this work is one journal paper and one conference paper:

- Ziping Zhao and Daniel P. Palomar, “Robust MLE of Reduced-Rank Regression Model With Simultaneous Factor and Feature Learning,” in preparation.

- Ziping Zhao and Daniel P. Palomar, “Robust Maximum Likelihood Estimation of Sparse Vector Error Correction Model,” in Proc. of the 5th IEEE Global Conference on Signal and Information Processing (GlobalSIP 2017), Montreal, QC, Canada, Nov. 14-16, 2017.

## Chapter 4

This chapter studies the mean-reverting portfolio design problem with a budget constraint.

The main output of this work is one journal paper and one conference paper:

- Ziping Zhao and Daniel P. Palomar, “Mean-Reverting Portfolio With Budget Constraint,” IEEE Transactions on Signal Processing, vol. 66, no. 9, pp. 2342-2357, May 2018.

- Ziping Zhao and Daniel P. Palomar, “Mean-Reverting Portfolio Design via Majorization-Minimization Method,” in Proc. of the 50th Asilomar Conference on Signals, Systems, and Computers (ACSSC/Asilomar 2016), Pacific Grove, CA, USA, Nov. 6-9, 2016.

## Chapter 5

This chapter studies the mean-reverting portfolio design problem with a leverage constraint.

The main output of this work is one journal paper and one conference paper:

- Ziping Zhao, Rui Zhou, and Daniel P. Palomar, “Optimal Mean-Reverting Portfolio With Leverage Constraint for Statistical Arbitrage in Finance,” IEEE Transactions on Signal Processing, vol. 67, no. 7, pp. 1681-1695, April 2019.

- Ziping Zhao, Rui Zhou, Zhongju Wang, and Daniel P. Palomar, “Optimal Portfolio Design for Statistical Arbitrage in Finance,” in Proc. of the 20th IEEE Statistical Signal Processing Workshop (SSP 2018), Freiburg, Germany, June 10-13, 2018.

# Chapter 2

# Efficient Estimation of Sparse Reduced-Rank Regression Model

In this chapter, the estimation problem for sparse reduced-rank regression (SRRR) model is considered. The SRRR model is widely used for dimension reduction and variable selection with applications in signal processing, econometrics, etc. The problem is formulated to minimize the least squares loss with a sparsity-inducing penalty considering an orthogonality constraint. Convex sparsity-inducing functions have been used for SRRR in literature. In this work, a nonconvex function is proposed for better sparsity inducing. An efficient algorithm is developed based on the alternating minimization (or projection) method to solve the nonconvex optimization problem. Numerical simulations show that the proposed algorithm is much more efficient compared to the benchmarks and the nonconvex function can result in a better estimation accuracy.

## 2.1 Introduction

Reduced-Rank Regression (RRR) [1], [2] is a multivariate linear regression model, where the coefficient matrix has a low-rank property. The name of “reduced-rank regression” was first brought up by Izenman [3]. Denote the response (or dependent) variables by $\mathbf{y}_t \in \mathbb{R}^P$ and predictor (or independent) variables by $\mathbf{x}_t \in \mathbb{R}^Q$, a general RRR model is given as follows:

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{A}\mathbf{B}^T\mathbf{x}_t + \boldsymbol{\varepsilon}_t \tag{2.1.1}$$

where the regression parameters are $\boldsymbol{\mu} \in \mathbb{R}^P$, $\mathbf{A} \in \mathbb{R}^{P \times r}$ and $\mathbf{B} \in \mathbb{R}^{Q \times r}$ and $\boldsymbol{\varepsilon}_t$ is the model innovation. Matrix $\mathbf{A}$ is often called sensitivity (or exposure) matrix and $\mathbf{B}$ is called factor matrix with the linear combinations $\mathbf{B}^T \mathbf{x}_t$ called latent factors. The "low-rank structure" formed by $\mathbf{A}\mathbf{B}^T$ essentially reduces the parameter dimension and improves explanatory ability of the model. The RRR model is widely used in situations when the response variables are believed to depend on a few linear combinations of the predictor variables, or when such linear combinations are of special interest.

The RRR model has been used in many signal processing problems, e.g., array signal processing [4], state space modeling [5], filter design [6], channel estimation and equalization for wireless communication [7], [8], [9], etc. It is also widely applied in econometrics and financial economics. Problems in econometrics were also the motivation for the pioneering work on the RRR estimation problem [1]. In financial economics, it can be used when modeling a group of economic indices by the lagged values of a set of economic variables. It is also widely used to model the relationship between financial asset returns and some related explanatory variables. Several asset pricing theories have been proposed for testing the efficiency of portfolios [10] and empirical verification using asset returns data on industry portfolios has been made through tests for reduced-rank regression [11]. The RRR model is also closely related the vector error correction model [12] in time series modeling and the latent factors can be used for statistical arbitrage [13] in finance. More applications on the RRR model can be found in, e.g., [14].

Like the low-rank structure for factor extraction, row-wise group sparsity on matrix $\mathbf{B}$ can also be considered to further realize predicting variable selection, which leads to the sparse RRR (SRRR) model [15]. Since $\mathbf{B}^T \mathbf{x}_t$ can be interpreted as the linear factors linking the response variables and the predictors, the SRRR can generate factors only with a subset of all the predictors. Variable selection is very important target in data analytics since it can help with model interpretability and improve estimation and forecasting accuracy.

In [15], the authors first considered the SRRR estimation problem, where the group sparsity was induced via the group lasso penalty [16]. An algorithm based on the alternating minimization (AltMin) method [17] was proposed. However, the proposed algorithm has a double loop where subgradient or variational method is used for the inner problem solving. Such an algorithm can be very slow in practice due to the double-loop nature where lots of iterations may be necessary to get an accurate enough solution at each iteration. Apart from that, besides the convex function for sparsity inducing, it is generally acknowledged that a nonconvex sparsity-inducing function can attain a better performance [18] which is proposed to use for sparsity estimation in this chapter.

In this chapter, the objective of the SRRR estimation problem is given as the ordinary least squares loss with a sparsity-inducing penalty. An orthogonality constraint is added for model identification purpose [15]. To solve this problem, an efficient AltMin-based single-loop algorithm is proposed. In order to pursue low-cost updating steps, a majorization-minimization method [19] and a nonconvexity redistribution method [20] are further adopted making the variable updates become two closed-form projections. Numerical simulations show that the proposed algorithm is more efficient compared to the benchmarks and the nonconvex function can attain a better estimation accuracy.

## 2.2 Sparse Reduced-Rank Regression

The SRRR estimation problem is formulated as follows:

$$\begin{array}{ll} \underset{\mathbf{A},\mathbf{B}}{\mathsf{minimize}} & F\left(\mathbf{A},\mathbf{B}\right) \triangleq L\left(\mathbf{A},\mathbf{B}\right) + R\left(\mathbf{B}\right) \\ \mathsf{subject\ to} & \mathbf{A}^T\mathbf{A} = \mathbf{I}, \end{array} \tag{2.2.1}$$

where $L\left(\mathbf{A},\mathbf{B}\right)$ is sample loss function and $R\left(\mathbf{B}\right)$ is the row-wise group sparsity regularizer. The constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ is added for identification purpose to deal with the unitary invariance of the parameters [15]. We further assume a sample path $\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^N$ $(N \geq \max(P, Q))$ is available from (2.1.1).

The least squares loss $L\left(\mathbf{A},\mathbf{B}\right)$ for the RRR model is obtained by minimizing a sample $\ell_2$-norm loss as follows[1]:

$$\begin{aligned} L\left(\mathbf{A},\mathbf{B}\right) &= \tfrac{1}{2}\textstyle\sum_{t=1}^N \left\|\mathbf{y}_t - \mathbf{A}\mathbf{B}^T\mathbf{x}_t\right\|_2^2 \\ &= \tfrac{1}{2}\left\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T\mathbf{X}\right\|_F^2, \end{aligned} \tag{2.2.2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$.

Sparse optimization [21] has become the focus of much research interest as a way to realize the variable selection (e.g., the group lasso method). For a vector $\mathbf{x} \in \mathbb{R}^K$, the sparsity

[1] In this chapter, the intercept term has been omitted without loss of generality as in [15], since it can always be removed by assuming that the response and predictor variables have zero mean.

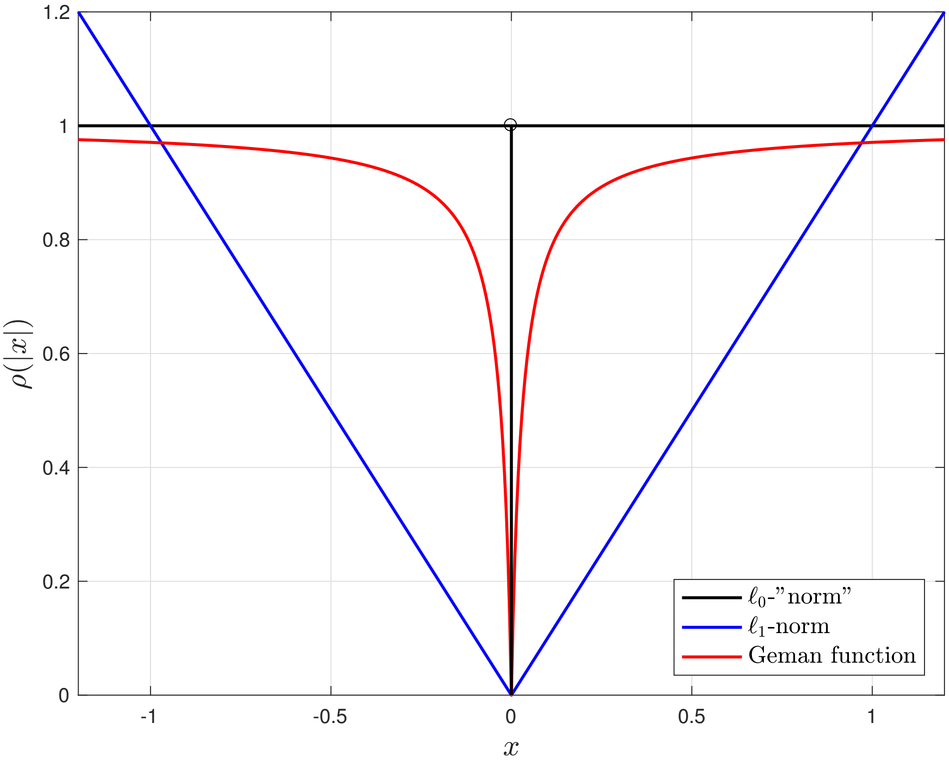


Figure 2.1: Nonsmooth sparsity-inducing function $\rho\left(x\right)$

level is usually measured by the $\ell_0$-norm, i.e., $\|\mathbf{x}\|_0 = \sum_{i=1}^{K} \operatorname{sgn}\left(|x_i|\right)$. Practically, the $\ell_1$-norm is used as the tightest convex relaxation to approximate it as in [15]. Although it is easy for optimization and has been shown to favor sparser solutions, the $\ell_1$-norm can lead to biased estimation with solutions not as accurate and sparse as desired and produce inferior prediction performance [18]. Nonconvex regularizers sacrifice convexity but can have a tighter approximation performance and are proposed for sparsity inducing which outperform the convex $\ell_1$-norm. In this chapter, two nonsmooth sparsity-inducing functions denoted by $\rho\left(|x|\right)$ are considered: the nonconvex Geman function [22] and the convex $\ell_1$-norm. Then, the row-wise group sparsity regularizer $R\left(\mathbf{B}\right)$ induced by $\rho\left(|x|\right)$ is given as follows:

$$R\left(\mathbf{B}\right) = \sum_{i=1}^{Q} \xi_i \rho\left(\|\mathbf{b}_i\|_2\right), \tag{2.2.3}$$

where $\mathbf{b}_i$ denotes the $i$th row of $\mathbf{B}$ and $\rho\left(|x|\right)$ is from $\rho_{\mathrm{GM}}\left(|x|\right) = \frac{|x|}{\theta+|x|}$ $(\theta > 0)$ and $\rho_{\ell_1}\left(|x|\right) = |x|$, which are shown in Figure 2.1.

Based on $L\left(\mathbf{A},\mathbf{B}\right)$ and $R\left(\mathbf{B}\right)$, the problem in (2.2.1) is a nonconvex nonsmooth optimization problem due to the nonconvex nonsmooth objective and the nonconvex constraint set.

## 2.3 Problem Solving Based on Alternating Minimization

The objective function in problem (2.2.1) has two variable blocks $(\mathbf{A},\mathbf{B})$. In this section, an alternating minimization (a.k.a. two-block coordinate descent) algorithm [17] will be proposed to solve it. At the $(k+1)$th iteration, this algorithm updates the variables according to the following two steps:

$$\begin{cases} \mathbf{A}^{(k+1)} \leftarrow \arg \min\limits_{\mathbf{A}:\mathbf{A}^T\mathbf{A}=\mathbf{I}} F\left(\mathbf{A};\mathbf{B}^{(k)}\right) \\ \mathbf{B}^{(k+1)} \leftarrow \arg \min\limits_{\mathbf{B}} F\left(\mathbf{B};\mathbf{A}^{(k+1)}\right), \end{cases} \tag{2.3.1}$$

where $\left(\mathbf{A}^{(k)},\mathbf{B}^{(k)}\right)$ are updates generated at the $k$th iteration.

First, let us start with the minimization step w.r.t. variable $\mathbf{A}$ when $\mathbf{B}$ is fixed at $\mathbf{B}^{(k)}$, the problem becomes[2]

$$\begin{array}{ll} \underset{\mathbf{A}}{\mathsf{minimize}} & F\left(\mathbf{A}\right) \simeq \frac{1}{2}\left\|\mathbf{Y}-\mathbf{A}\mathbf{B}^{(k)T}\mathbf{X}\right\|_F^2 \\ \mathsf{subject\ to} & \mathbf{A}^T\mathbf{A}=\mathbf{I}, \end{array} \tag{2.3.2}$$

where the "$\simeq$" means "equivalence" up to additive constants. This nonconvex problem is the classical orthogonal Procrustes problem (projection) [23], which has a closed-form solution given in the following lemma.

**Lemma 1.** *[23] The orthogonal Procrustes problem in* (2.3.2) *can be equivalently reformulated into the following form:*

$$\begin{array}{ll} \underset{\mathbf{A}}{\mathsf{minimize}} & \left\|\mathbf{A}-\mathbf{P}_A^{(k)}\right\|_F^2 \\ \mathsf{subject\ to} & \mathbf{A}^T\mathbf{A}=\mathbf{I}, \end{array}$$

*where* $\mathbf{P}_A^{(k)} \triangleq \mathbf{Y}\mathbf{X}^T\mathbf{B}^{(k)}$*. Let the thin singular value decomposition (SVD) be* $\mathbf{P}_A = \mathbf{U}\mathbf{S}\mathbf{V}^T$*, where* $\mathbf{U}\in\mathbb{R}^{Q\times r}$ *and* $\mathbf{S},\mathbf{V}\in\mathbb{R}^{r\times r}$*, then the optimal update* $\mathbf{A}^{(k+1)}$ *is given by*

$$\mathbf{A}^{(k+1)} = \mathbf{U}\mathbf{V}^T. \tag{2.3.3}$$

[2] For simplicity, $F\left(\mathbf{A};\mathbf{B}^{(k)}\right)$ is written as $F\left(\mathbf{A}\right)$ and likewise the fixed variables $\mathbf{A}^{(k)}$ and/or $\mathbf{B}^{(k)}$ in other functions will also be reduced in the following.

Then, when fixing $\mathbf{A}$ with $\mathbf{A}^{(k+1)}$, the problem for $\mathbf{B}$ is

$$\begin{aligned}\underset{\mathbf{B}}{\mathsf{minimize}} \quad F(\mathbf{B}) = \tfrac{1}{2}\left\|\mathbf{Y} - \mathbf{A}^{(k+1)}\mathbf{B}^T\mathbf{X}\right\|_F^2 \\ + \textstyle\sum_{i=1}^{Q} \xi_i \rho\left(\|\mathbf{b}_i\|_2\right),\end{aligned} \tag{2.3.4}$$

which is a penalized multivariate regression problem. It has no analytical solution but standard nonconvex optimization algorithms or solvers can be applied to solve it. However, using such methods will lead to an iterative process, which could be undesirable in terms of efficiency. In addition, since the nonconvexity of this problem, if no guarantee for the solution quality can be claimed, the overall convergence for the alternating algorithm in general is not guaranteed.

In this chapter, the $\mathbf{B}$-subproblem is solved via a simple update rule while guaranteeing convergence of the overall algorithm. We propose to update $\mathbf{B}$ by solving a majorized surrogate problem for problem (2.3.4) [19], [24] written as

$$\mathbf{B}^{(k+1)} \leftarrow \arg\min_{\mathbf{B}} \overline{F}\left(\mathbf{B}; \mathbf{A}^{(k+1)}, \mathbf{B}^{(k)}\right), \tag{2.3.5}$$

where $\overline{F}\left(\mathbf{B}; \mathbf{A}^{(k+1)}, \mathbf{B}^{(k)}\right)$ or simply $\overline{F}(\mathbf{B})$ is the majorizing function of $F(\mathbf{B})$ at $(\mathbf{A}^{(k+1)}, \mathbf{B}^{(k)})$. To get $\overline{F}(\mathbf{B})$, we need the following results.

**Proposition 1.** *[19] Let $\mathbf{A} \in \mathbb{S}^K$, then at any point $\mathbf{x}^{(k)} \in \mathbb{R}^K$, $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is majorized as follows:*

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} \leq & \mathbf{x}^{(k)T}\mathbf{A}\mathbf{x}^{(k)} + 2\mathbf{x}^{(k)T}\mathbf{A}\left(\mathbf{x} - \mathbf{x}^{(k)}\right) \\ & + \psi(\mathbf{A})\left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_2^2,\end{aligned}$$

*where $\psi(\mathbf{A}) \geq \lambda_{\max}(\mathbf{A})$ is a pre-specified constant.*

Observing that the first part in $F\left(\mathbf{B}; \mathbf{A}^{(k+1)}\right)$, i.e., the least squares loss $L\left(\mathbf{B}; \mathbf{A}^{(k+1)}\right)$, is quadratic in $\mathbf{B}$, based on Proposition 1, we can have the following result.

**Lemma 2.** *The function $L\left(\mathbf{B}; \mathbf{A}^{(k+1)}\right)$ can be majorized at $\left(\mathbf{A}^{(k+1)}, \mathbf{B}^{(k)}\right)$ by*

$$\overline{L}(\mathbf{B}) \simeq \frac{1}{2}\psi(\mathbf{G}^{(k)})\left\|\mathbf{B} - \mathbf{P}_B^{(k)}\right\|_F^2,$$

*where $\mathbf{G}^{(k)} \triangleq \mathbf{A}^{(k+1)T}\mathbf{A}^{(k+1)} \otimes \mathbf{X}\mathbf{X}^T$, $\psi\left(\mathbf{G}^{(k)}\right) \geq \lambda_{\max}\left(\mathbf{G}^{(k)}\right)$, and $\mathbf{P}_B^{(k)} \triangleq \psi^{-1}\left(\mathbf{G}^{(k)}\right)\mathbf{X}\mathbf{Y}^T\mathbf{A}^{(k+1)} - \psi^{-1}\left(\mathbf{G}^{(k)}\right)\mathbf{X}\mathbf{X}^T\mathbf{B}^{(k)}\mathbf{A}^{(k+1)T}\mathbf{A}^{(k+1)} + \mathbf{B}^{(k)}$.*

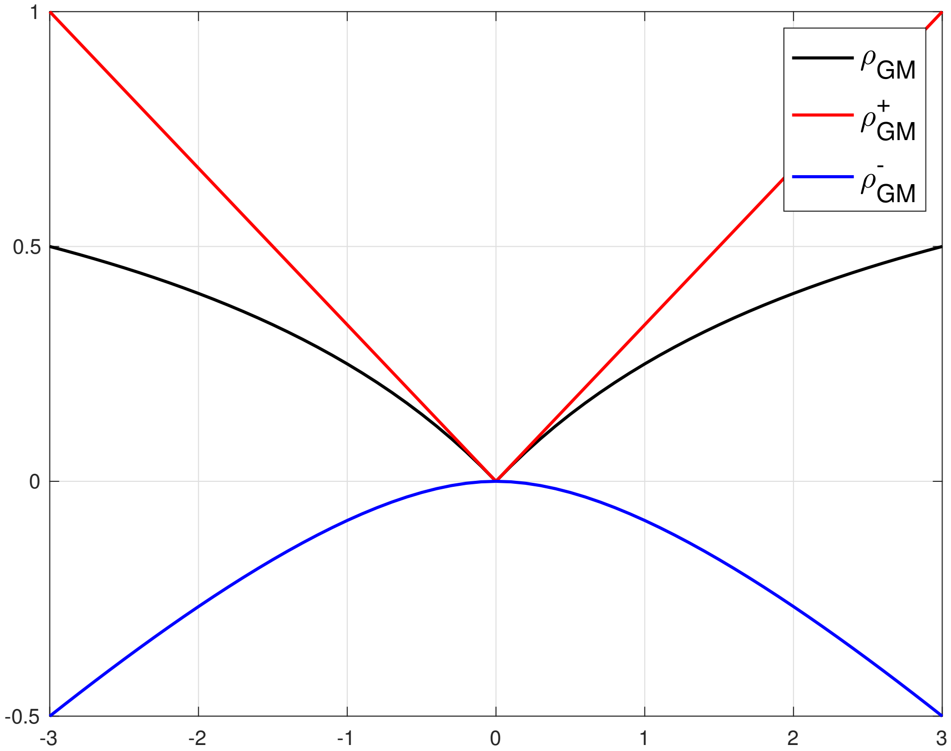


Figure 2.2: Nonconvexity Redistribution Method for $\rho_{\text{GM}}(|x|)$

Likewise, the majorization method can also be applied to the regularizer $R(\mathbf{B})$. But we first need the following result.

**Proposition 2.** *[20] The nonsmooth sparsity-inducing function $\rho(|x|)$ can be decomposed as*

$$\rho(|x|) = \kappa|x| + \rho(|x|) - \kappa|x|,$$

*where $\rho(|x|) - \kappa|x|$ is a smooth and concave function when $\kappa \triangleq \rho'(0^+)$. Specifically, for $\rho_{\ell_1}(|x|)$, $\kappa = 1$; and for $\rho_{\text{GM}}(|x|)$, $\kappa = 1/\theta$.*

An illustrating example for Proposition 2 is given in Figure 2.2. And based on Proposition 2, we can accordingly decompose the row-wise group sparsity regularizer $R(\mathbf{B})$ as

$$R(\mathbf{B}) = R^{+}(\mathbf{B}) + R^{-}(\mathbf{B}), \tag{2.3.6}$$

where $R^{+}(\mathbf{B}) = \kappa \sum_{i=1}^{Q} \xi_i \|\mathbf{b}_i\|_2$ which exactly takes the form of classical group lasso and $R^{-}(\mathbf{B}) = R(\mathbf{B}) - R^{+}(\mathbf{B})$. For $R^{-}(\mathbf{B})$, we can have the following majorization result.

**Lemma 3.** *The function $R^{-}(\mathbf{B})$ can be majorized at $\mathbf{B}^{(k)}$ by*

$$\overline{R}^{-}(\mathbf{B}) \simeq \operatorname{Tr}\left(\mathbf{K}^{(k)T}\mathbf{B}\right),$$

*where $\mathbf{K}^{(k)} = R^{-\prime}\left(\mathbf{B}^{(k)}\right)$ with $R^{-\prime}\left(\mathbf{B}^{(k)}\right)$ to be the gradient of $R^{-}(\mathbf{B})$ at point $\mathbf{B}^{(k)}$ and specifically*

$$\mathbf{k}_i^{(k)} \triangleq \xi_i \left[\rho'\left(\left\|\mathbf{b}_i^{(k)}\right\|_2\right) - \kappa\right] \frac{\mathbf{b}_i^{(k)}}{\left\|\mathbf{b}_i^{(k)}\right\|_2},$$

*where $\mathbf{k}_i^{(k)}$ denotes the $i$th column of $\mathbf{K}^{(k)}$.*

Based on $\overline{L}(\mathbf{B})$ in Lemma 2 and $\overline{R}^{-}(\mathbf{B})$ in Lemma 3, we can finally have the majorization function for $F(\mathbf{B})$ given as

$$\begin{aligned}\overline{F}(\mathbf{B}) &= \overline{L}(\mathbf{B}) + R^{+}(\mathbf{B}) + \overline{R}^{-}(\mathbf{B}) \\ &\simeq \tfrac{1}{2}\psi\left(\mathbf{G}^{(k)}\right)\left\|\mathbf{B} - \mathbf{P}_{B,R}^{(k)}\right\|_F^2 + R^{+}(\mathbf{B}),\end{aligned} \tag{2.3.7}$$

where $\mathbf{P}_{B,R}^{(k)} \triangleq \mathbf{P}_B^{(k)} - \psi^{-1}\left(\mathbf{G}^{(k)}\right)\mathbf{K}^{(k)}$. The result by using Lemma 2 and Lemma 3 is that we shift the nonconvexity associated with the nonconvex regularizer to the loss function, and transform the nonconvex regularizer to the familiar convex group lasso regularizer. It is easy to observe that the algorithm derivation above can be easily applied to the classical group lasso and at that case $\mathbf{K}^{(k)} = \mathbf{0}$.

Finally, the majorizing problem for the $\mathbf{B}$-subproblem is given in the following form:

$$\underset{\mathbf{B}}{\mathsf{minimize}} \quad \tfrac{1}{2}\psi\left(\mathbf{G}^{(k)}\right)\left\|\mathbf{B} - \mathbf{P}_{B,R}^{(k)}\right\|_F^2 + \tfrac{1}{\theta}\textstyle\sum_{i=1}^{Q} \xi_i \left\|\mathbf{b}_i\right\|_2, \tag{2.3.8}$$

which becomes separable among the rows of matrix $\mathbf{B}$. The resulting separable problems can be efficiently solved using the proximal algorithms [25] and have closed-form solutions which are given in the following lemma.

**Lemma 4.** *[25] The problem in* (2.3.8) *has a closed-form proximal update which is given by*

$$\mathbf{b}_i^{(k+1)} = \left[1 - \tfrac{1}{\theta}\tfrac{\xi_i}{\psi\left(\mathbf{G}^{(k)}\right)}\tfrac{1}{\left\|\mathbf{p}_i^{(k)}\right\|_2}\right]^{+} \mathbf{p}_i^{(k)},$$

*where $[x]^{+} = \max(x, 0)$, and $\mathbf{p}_i^{(k)}$ is the $i$th row of $\mathbf{P}_{B,R}^{(k)}$.*

### 2.3.1 AltMin-MM: Algorithm for SRRR Estimation

Based on the alternating minimization algorithm together with the majorization and nonconvex redistribution methods, to solve the original SRRR estimation problem (2.2.1), we just need to update the variables with closed-form solutions alternatingly until convergence.

The overall algorithm is summarized in the following.

**Algorithm 2.1** AltMin-MM: Algorithm for SRRR Estimation

```
Require: $\mathbf{X}$, $\mathbf{Y}$ and $\xi_i$ with $i = 1, \ldots, r$.
Set $k = 0$, $\mathbf{A}^{(0)}$ and $\mathbf{B}^{(0)}$.
repeat
    Compute $\mathbf{P}_A^{(k)}$
    Update $\mathbf{A}^{(k+1)}$ in closed-form solution (Lemma 1)
    Compute $\mathbf{G}^{(k)}$, $\psi(\mathbf{G}^{(k)})$ and $\mathbf{P}_{B,R}^{(k)}$
    Update $\mathbf{B}^{(k+1)}$ in closed-form solution (Lemma 4)
    $k \leftarrow k + 1$
until convergence
```

## 2.4 Numerical Simulations

In order to test the performance of the problem model and proposed algorithm. Numerical simulations are considered in this section. An SRRR $(P = 7,\ Q = 5,\ r = 3)$ with underlying group sparse structure for $\mathbf{B}$ is specified firstly. Then a sample path $\{\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\varepsilon}_t\}_{t=1}^N$ is generated.

We first examine the efficiency of our proposed AltMin-MM algorithm when the sparsity regularizer is the group lasso penalty, i.e., $\rho\left(|x|\right) = |x|$ which is adopted in [15]. We compare our algorithm with the AltMin-based algorithms with subproblem solved by subgradient method (AltMin-SubGrad) and by variational inequality method (AltMin-VarIneq) for the proposed problem in (2.2.1). The convergence result of the objective function value ($N = 100$) is shown in Fig. 2.3. It is easy to see that our proposed algorithm can have a faster convergence. It should be mentioned that although the first descent step can attain a better solution in the benchmark methods, since a lot of iterations can be required to get a accuracy enough solution, they show a slower convergence in general.

We further test the case when the regularizer is based on nonconvex Geman function, i.e., $\rho\left(|x|\right) = \frac{|x|}{\theta+|x|}$ ($\theta = 0.05$). Since there is no benchmark in the literature, our proposed algorithm AltMin-MM is compared with a benchmark where the convex $\mathbf{B}$-subproblem is derived to be a tight majorized problem of the original problem by just majorizing the nonconvex term $R^{-}(\mathbf{B})$ and is solved using `CVX`. The objective function convergence result is shown in Fig. 2.4.

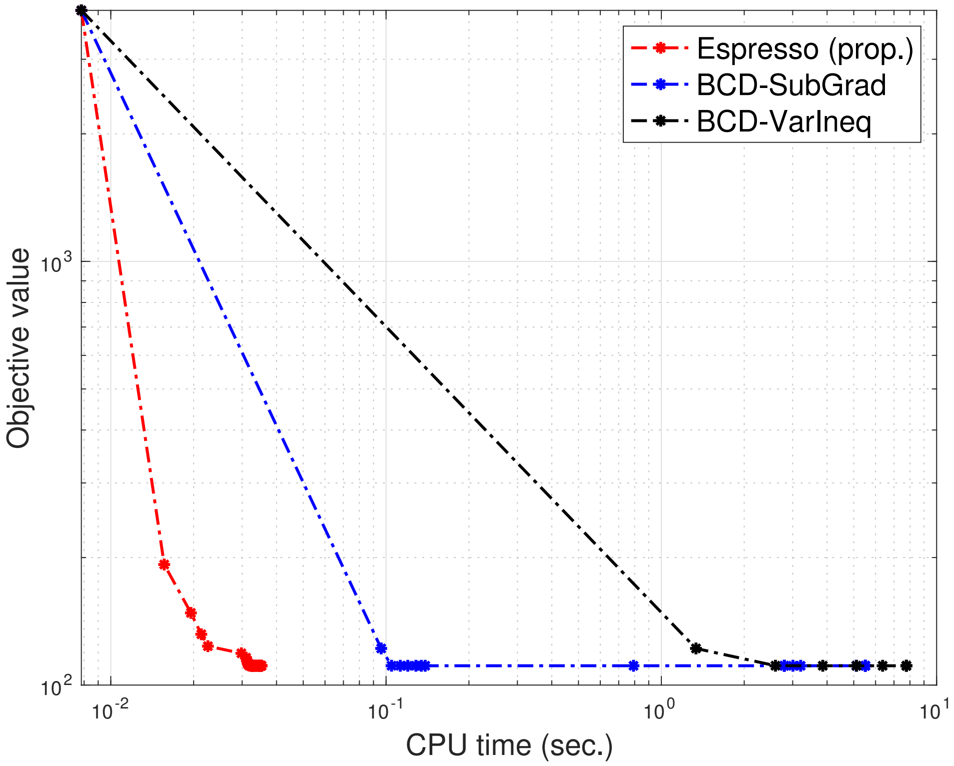


Figure 2.3: Convergence comparison for objective function value.

We also examine the estimation accuracy of the proposed formulation and algorithm. It is evaluated by computing the angle between the estimated factor matrix space $\hat{\mathbf{B}}^{(m)}$ and the true space $\mathbf{B}$ denoted by $\theta^{(m)}(\hat{\mathbf{B}}^{(m)}, \mathbf{B})$ for the $m$th Monte-Carlo simulation, with $m = 1, \ldots, M$ and $M = 500$. The angle $\theta^{(m)}(\hat{\mathbf{B}}^{(m)}, \mathbf{B})$ is computed as follows [2]. First, compute the QR decompositions $\hat{\mathbf{B}}^{(m)} = \mathbf{Q}_m \mathbf{R}_m$ and $\mathbf{B} = \mathbf{QR}$. Next, compute the SVD of $\mathbf{Q}_m^T \mathbf{Q} = \mathbf{U}_Q \mathbf{S}_Q \mathbf{V}_Q$ where the diagonal elements of $\mathbf{S}_Q$ is written as $s_1 \geq \ldots \geq s_r$. Then, the maximum angle is given by $\theta^{(m)}(\hat{\mathbf{B}}^{(m)}, \mathbf{B}) = \arccos(s_r)$. The averaged angle for $M$ Monte-Carlo runs is given by

$$\theta(\hat{\mathbf{B}}, \mathbf{B}) = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \theta^{(m)}(\hat{\mathbf{B}}^{(m)}, \mathbf{B}),$$

where it can take values from $0$ (identical subspaces) to $\frac{\pi}{2}$ (orthogonal subspaces). We compared three cases which are RRR estimation (without sparsity), SRRR estimation with convex sparsity-inducing function $\rho_{\ell_1}(|x|)$, and SRRR estimation with nonconvex sparsity-inducing

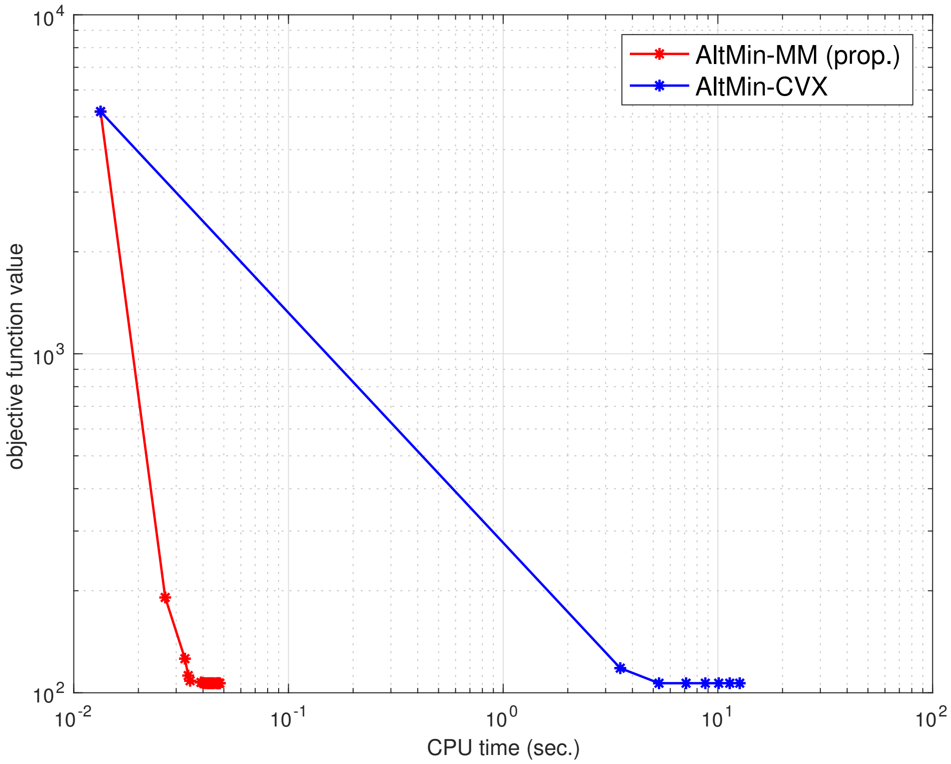


Figure 2.4: Convergence comparison for objective function value.

function $\rho_{\text{GM}}(|x|)$. It is easy to say that, the SRRR problem formulation can really exploit the group sparsity structure in $\mathbf{B}$ and the nonconvex function $\rho_{\text{GM}}(|x|)$ shows a better performance over the convex one.

## 2.5 Chapter Summary and Conclusions

The SRRR model estimation problem has been considered in this chapter. It has been formulated to minimize the least squares loss with a nonconvex nonsmooth group sparsity penalty by considering an orthogonality constraint. An efficient and flexible algorithm has been proposed which exploits the problem structure and has a low computational complexity. Numerical simulations have shown the efficiency of the proposed algorithm. It should be noted that the proposed algorithm is also general enough to be extended to many other sparse and structured problems in machine learning and signal processing.

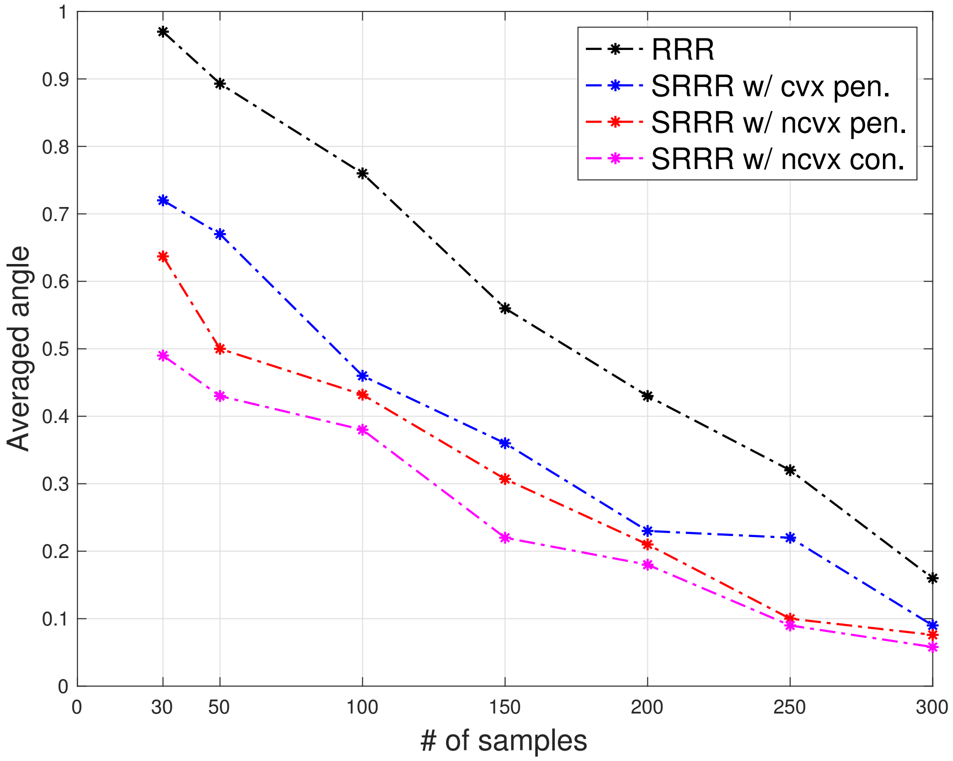


Figure 2.5: Estimation accuracy based on averaged angle.

# Chapter 3

# Robust Estimation of Sparse Vector Error Correction Model

## 3.1 Introduction

The vector error correction model (VECM) [26] is very important in cointegration analysis to estimate and test for the long-run cointegrated equilibriums. It is widely used in time series modeling for financial returns and macroeconomic variables. In [27], [28], Engle and Granger first proposed the concept of "cointegration" to describe the linear stationary relationships in the nonstationary time series. Later, Johansen studied the statistical estimation and inference problem in time series cointegration modeling [29], [30], [31]. A VECM for $\mathbf{y}_t \in \mathbb{R}^K$ is given as follows:

$$\Delta\mathbf{y}_t = \boldsymbol{\nu} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \textstyle\sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta\mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \tag{3.1.1}$$

where $\Delta$ is the first difference operator, i.e., $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$, $\boldsymbol{\nu}$ denotes the drift, $\boldsymbol{\Pi}$ determines the long-run equilibriums, $\boldsymbol{\Gamma}_i$ $(i = 1, \ldots, p-1)$ contains the short-run effects, and $\boldsymbol{\varepsilon}_t$ is the innovation with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. Matrix $\boldsymbol{\Pi}$ has a reduced cointegration rank $r$, i.e., $\mathrm{rank}\,(\boldsymbol{\Pi}) = r < K$, and it can be written as $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$ $(\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{K\times r})$. Accordingly, $\mathbf{y}_t$ is said to be cointegrated with rank $r$, and $\boldsymbol{\beta}^T\mathbf{y}_t$ gives the long-run stationary time series defined by the cointegration matrix $\boldsymbol{\beta}$. Such long-run equilibriums are often implied by economic theory and can be used for statistical arbitrage [32].

It is well-known that financial returns and macroeconomic variables exhibit heavy-tails and are often associated with outliers due to external factors, like political and regulatory changes, as well as data corruption, like faulty observations and wrongly processed data [33]. These stylized features contradict the popular Gaussian noise assumption typically made in the theoretical analysis and estimation procedures with adverse effects in the estimated models. Cointegration analysis is particularly sensitive to these issues. Papers [34], [35], [36] discussed the properties of the Dickey-Fuller test and the Johansen test in the presence of outliers. Lucas studied such issues both from a theoretical and an empirical point of view [37], [38], [39]. To deal with the heavy-tails and outliers in time series modeling, simple and effective estimation methods are needed. In [40], the pseudo maximum likelihood estimators were introduced for VECM. In this paper, based on [40], we formulate the estimation problem based on the log-likelihood function of the Cauchy distribution as a conservative representative of the heavy-tailed distributions to better fit the heavy-tails and dampen the influence of outliers.

Sparse optimization [21] has become the focus of much research interest as a way to realize feature selection and dimension reduction (e.g., lasso [41]). In [42], element-wise sparsity was imposed on $\boldsymbol{\beta}$ in VECM modeling. As indicated by [15], [43], to realize the feature selection purpose, group sparsity is better since it can simultaneously reduce the same variable in all cointegration relations and naturally keep the geometry of the low-rank parameter space. In this paper, instead of imposing the group sparsity on $\boldsymbol{\beta}$, we equivalently put group sparsity on $\boldsymbol{\Pi}$ and add a rank constraint for it, which can realize the same target without the ahead factorization $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$. For sparsity pursuing, i.e., approximating the $\ell_0$-“norm”, rather than the popular $\ell_1$-norm, we use a nonconvex Geman-type function [22] which has a better approximation power. A smoothed counterpart is also firstly proposed to reduce the “singularity issue” in optimization, based on which the group sparsity regularizer of $\boldsymbol{\beta}$ is attained.

Robust estimation is somewhat underrated in financial applications due to the complex computations that are time and resource intensive. By considering the robust loss and the regularizer, a nonconvex optimization problem is finally formulated. The expectation-maximization (EM) is usually used to solve the robust losses (e.g., [44]). However, EM cannot be applied for our formulation. To deal with it, an efficient algorithm based on the majorization-minimization (MM) method is proposed with estimation performance numerically shown.

## 3.2 Robust Estimation of Sparse VECM

Suppose a sample path $\{\mathbf{y}_t\}_{t=1}^N$ $(N > K)$ and the needed pre-sample values are available, then the VECM (3.1.1) can be written into a matrix form as follows:

$$\Delta\mathbf{Y} = \mathbf{\Pi}\mathbf{Y}_{-1} + \mathbf{\Gamma}\Delta\mathbf{X} + \mathbf{E}, \tag{3.2.1}$$

where $\mathbf{\Gamma} = [\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_{p-1}, \boldsymbol{\nu}]$, $\Delta\mathbf{Y} = [\Delta\mathbf{y}_1, \ldots, \Delta\mathbf{y}_N]$, $\mathbf{Y}_{-1} = [\mathbf{y}_0, \ldots, \mathbf{y}_{N-1}]$, $\Delta\mathbf{X} = [\Delta\mathbf{x}_1, \ldots, \Delta\mathbf{x}_N]$ with $\Delta\mathbf{x}_t = \left[\begin{array}{c} \Delta\mathbf{y}_{t-1}^T, \ldots, \Delta\mathbf{y}_{t-p+1}^T, 1 \end{array}\right]^T$, and $\mathbf{E} = [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N]$.

### 3.2.1 Robustness Pursued by Cauchy Log-likelihood Loss

The robustness is pursued by a multivariate Cauchy distribution. Assume the innovations $\boldsymbol{\varepsilon}_t$'s in (3.1.1) follow Cauchy distribution, i.e., $\boldsymbol{\varepsilon}_t \sim \text{Cauchy}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \in \mathbb{S}_{++}^K$, then the probability density function is given by

$$g_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}_t) = \frac{\Gamma\left(\frac{1+K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)(\nu\pi)^{\frac{K}{2}}} [\det(\mathbf{\Sigma})]^{-\frac{1}{2}} \left(1 + \boldsymbol{\varepsilon}_t^T \mathbf{\Sigma}^{-1} \boldsymbol{\varepsilon}_t\right)^{-\frac{1+K}{2}}.$$

The negative log-likelihood loss function of the Cauchy distribution for $N$ samples from (3.1.1) is written as follows:

$$\begin{aligned} L(\boldsymbol{\theta}) = \tfrac{N}{2} \log\det(\mathbf{\Sigma}) + \tfrac{1+K}{2} \textstyle\sum_{i=1}^N \log\Big(1+ \\ \left\|\mathbf{\Sigma}^{-\frac{1}{2}}(\Delta\mathbf{y}_i - \mathbf{\Pi}\mathbf{y}_{i-1} - \mathbf{\Gamma}\Delta\mathbf{x}_{i-1})\right\|_2^2\Big), \end{aligned} \tag{3.2.2}$$

where the constants are dropped and $\boldsymbol{\theta} \triangleq \{\mathbf{\Pi}(\boldsymbol{\alpha}, \boldsymbol{\beta}), \mathbf{\Gamma}, \mathbf{\Sigma}\}$.

### 3.2.2 Group Sparsity Pursued by Nonconvex Regularizer

For a vector $\mathbf{x} \in \mathbb{R}^K$, the sparsity level is usually measured by the $\ell_0$-"norm" (or $\text{sgn}(|x|)$) as $\|\mathbf{x}\|_0 = \sum_{i=1}^K \text{sgn}(|x_i|) = k$, where $k$ is the number of nonzero entries in $\mathbf{x}$. Generally, applying the $\ell_0$-"norm" to different groups of variables can enforce group sparsity in the solutions. The $\ell_0$-"norm" is not convex and not continuous, which makes it computationally difficult and leads to intractable NP-hard problems. So, $\ell_1$-norm as the tightest convex relaxation is usually used to approximate the $\ell_0$-"norm" in practice, which is easier for optimization and still favors sparse solutions.

Tighter nonconvex sparsity-inducing functions can lead to better performance [21]. In this paper, to better pursue the sparsity and to remove the "singularity issue", i.e., when using nonsmooth functions, the variable may get stuck at a nonsmooth point [45], a smooth nonconvex function based on the rational (Geman) function in [22] is used given as follows:

$$\mathrm{rat}_p^{\epsilon}(x) = \begin{cases} \frac{px^2}{2\epsilon(p+\epsilon)^2}, & |x| \leq \epsilon \\ \frac{|x|}{p+|x|} - \frac{2\epsilon^2+p\epsilon}{2(p+\epsilon)^2}, & |x| > \epsilon \end{cases}.$$

In order to attain feature selection in VECM, i.e., sparse cointegration relations, according to [15], [43], we can impose the row-wise group sparsity on matrix $\boldsymbol{\beta}$. In fact, due to $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$, the row-wise sparsity imposed on $\boldsymbol{\beta}$ can also be realized by directly estimating $\boldsymbol{\Pi}$ through imposing the column-wise group sparsity on $\boldsymbol{\Pi}$ and constraining its rank. Then we have the sparsity regularizer of matrix $\boldsymbol{\Pi}$ which is given by

$$R(\boldsymbol{\Pi}) = \sum_{i=1}^{K} \mathrm{rat}_p^{\epsilon}(\|\boldsymbol{\pi}_i\|_2), \tag{3.2.3}$$

where $\boldsymbol{\pi}_i$ $(i = 1, \ldots, K)$ denotes the $i$th column of $\boldsymbol{\Pi}$. The grouping effect is achieved by taking the $\ell_2$-norm of each group, and then applying the group regularization.

### 3.2.3 Problem Formulation

By combining the robust loss function (3.2.2) and the sparsity regularizer (3.2.3), we attain a penalized maximum likelihood estimation formulation which is specified as follows:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}=\{\boldsymbol{\Pi},\boldsymbol{\Gamma},\boldsymbol{\Sigma}\}}{\mathsf{minimize}} & F(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}) + \xi R(\boldsymbol{\Pi}) \\ \mathsf{subject\ to} & \mathrm{rank}(\boldsymbol{\Pi}) \leq r,\ \boldsymbol{\Sigma} \succeq \mathbf{0}. \end{array} \tag{3.2.4}$$

This is a constrained smooth nonconvex problem due to the nonconvexity of the objective function and the constraint set.

## 3.3 Problem Solving via The MM Method

### 3.3.1 Majorization for the Robust Loss Function $L(\boldsymbol{\theta})$

Instead of using the EM method [44], in this paper, we derive the majorizing function for $L(\boldsymbol{\theta})$ from an MM perspective.

**Lemma 5.** *At any point* $x^{(k)} \in \mathbb{R}$, $\log(1+x) \leq \log\left(1+x^{(k)}\right) + \frac{1}{1+x^{(k)}}\left(x - x^{(k)}\right)$*, with the equality attained at* $x = x^{(k)}$.

Based on Lemma 5, at the iterate $\boldsymbol{\theta}^{(k)}$, the loss function $L(\boldsymbol{\theta})$ can be majorized by the following function:

$$\begin{aligned}&\overline{L}_1\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right) \simeq \\ &\tfrac{N}{2}\log\det(\boldsymbol{\Sigma}) + \tfrac{1}{2}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1} - \boldsymbol{\Gamma}\Delta\bar{\mathbf{X}}\right)\right\|_F^2,\end{aligned}$$

where "$\simeq$" means "equivalence" up to additive constants, $\Delta\bar{\mathbf{Y}} = \Delta\mathbf{Y}\text{diag}\left(\sqrt{\mathbf{w}^{(k)}}\right)$, $\bar{\mathbf{Y}}_{-1} = \mathbf{Y}_{-1}\text{diag}\left(\sqrt{\mathbf{w}^{(k)}}\right)$, and $\Delta\bar{\mathbf{X}} = \Delta\mathbf{X}\text{diag}\left(\sqrt{\mathbf{w}^{(k)}}\right)$ with $\mathbf{w}^{(k)} \in \mathbb{R}^N$ and the element

$$w_t^{(k)} = \frac{1+K}{1+\left\|\boldsymbol{\Sigma}^{-\frac{(k)}{2}}\left(\Delta\mathbf{y}_t - \boldsymbol{\Pi}^{(k)}\mathbf{y}_{t-1} - \boldsymbol{\Gamma}^{(k)}\Delta\mathbf{x}_{t-1}\right)\right\|_2^2},\ t = 1\ldots N.$$

By taking the partial derivatives for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$, and defining the projection matrix $\bar{\mathbf{M}} = \mathbf{I}_N - \Delta\bar{\mathbf{X}}^T\left(\Delta\bar{\mathbf{X}}\Delta\bar{\mathbf{X}}^T\right)^{-1}\Delta\bar{\mathbf{X}}$, the majorizing function $\overline{L}_1\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)$ is minimized when

$$\begin{aligned}&\boldsymbol{\Gamma}(\boldsymbol{\Pi}) = \left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1}\right)\Delta\bar{\mathbf{X}}^T\left(\Delta\bar{\mathbf{X}}\Delta\bar{\mathbf{X}}^T\right)^{-1}, \\ &\boldsymbol{\Sigma}(\boldsymbol{\Pi}) = \tfrac{1}{N}\left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1}\right)\bar{\mathbf{M}}\left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1}\right)^T.\end{aligned}$$

Substituting these equations back into $\overline{L}_1\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)$, we have

$$\begin{aligned}&\overline{L}_1\left(\boldsymbol{\Pi}, \boldsymbol{\theta}^{(k)}\right) \simeq \\ &\tfrac{N}{2}\log\det\left[\left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1}\right)\bar{\mathbf{M}}\left(\Delta\bar{\mathbf{Y}} - \boldsymbol{\Pi}\bar{\mathbf{Y}}_{-1}\right)^T\right].\end{aligned}$$

Then we introduce the following useful lemma.

**Lemma 6.** *At any point* $\mathbf{R}^{(k)} \in \mathbb{S}_{++}^K$, $\log\det(\mathbf{R}) \leq \text{Tr}\left(\mathbf{R}^{-(k)}\mathbf{R}\right) + \log\det\left(\mathbf{R}^{(k)}\right) - K$*, with the equality attained at* $\mathbf{R} = \mathbf{R}^{(k)}$.

Based on Lemma 6, $\overline{L}_1\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$ is further majorized by

$$\overline{L}_2\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)\simeq\frac{1}{2}\left\|\mathbf{\Sigma}^{-\frac{(k)}{2}}\left(\Delta\bar{\mathbf{Y}}-\mathbf{\Pi}\bar{\mathbf{Y}}_{-1}\right)\bar{\mathbf{M}}\right\|_F^2.$$

Finally, after majorization, $\overline{L}_2\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$ becomes a quadratic function in $\mathbf{\Pi}$.

### 3.3.2 Majorization for the Sparsity Regularizer $R\left(\mathbf{\Pi}\right)$

In this section, we introduce the majorization trick to deal with the nonconvex sparsity regularizer $R\left(\mathbf{\Pi}\right)$.

**Lemma 7.** *At any given point $x^{(k)}$, $\mathrm{rat}_p^{\epsilon}\left(x\right)\leq\frac{q^{(k)}}{2}x^2+c^{(k)}$, with the equality attained at $x=x^{(k)}$, the coefficient $q^{(k)}=p\left[\max\left\{\epsilon,\left|x^{(k)}\right|\right\}\left(p+\max\left\{\epsilon,\left|x^{(k)}\right|\right\}\right)^2\right]^{-1}$, and constant $c^{(k)}=\frac{p\max\left\{\epsilon,\left|x^{(k)}\right|\right\}+2\left(\max\left\{\epsilon,\left|x^{(k)}\right|\right\}\right)^2}{2\left(p+\max\left\{\epsilon,\left|x^{(k)}\right|\right\}\right)^2}-\frac{p\epsilon+2\epsilon^2}{2(p+\epsilon)^2}$.*

The majorization in Lemma 7 is pictorially shown in Fig. 3.1. Then at $\boldsymbol{\theta}^{(k)}$, the regularizer $R\left(\mathbf{\Pi}\right)$ can be majorized by

$$\overline{R}\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)\simeq\frac{1}{2}\mathrm{vec}\left(\mathbf{\Pi}\right)^T\left[\mathrm{diag}\left(\mathbf{q}^{(k)}\right)\otimes\mathbf{I}_K\right]\mathrm{vec}\left(\mathbf{\Pi}\right),$$

where $\mathbf{q}^{(k)}\in\mathbb{R}^K$ with the $i$th ($i=1,\ldots,K$) element

$$q_i^{(k)}=p\left[\max\left\{\epsilon,\left\|\boldsymbol{\pi}_i^{(k)}\right\|_2\right\}\left(p+\max\left\{\epsilon,\left\|\boldsymbol{\pi}_i^{(k)}\right\|_2\right\}\right)^2\right]^{-1}.$$

### 3.3.3 The Majorized Subproblem in MM

By combining $\overline{L}_2\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$ and $\overline{R}\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$, we can get the majorizing function for $\overline{F}\left(\boldsymbol{\theta}\right)$ which is given as follows:

$$\begin{aligned}\overline{F}_1\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)&\simeq\overline{L}_2\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)+\xi\overline{R}\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)\\&\simeq\frac{1}{2}\mathrm{vec}\left(\mathbf{\Pi}\right)^T\mathbf{G}^{(k)}\mathrm{vec}\left(\mathbf{\Pi}\right)-\mathrm{vec}\left(\mathbf{H}^{(k)}\right)^T\mathrm{vec}\left(\mathbf{\Pi}\right),\end{aligned}$$

where $\mathbf{G}^{(k)}=\bar{\mathbf{Y}}_{-1}\bar{\mathbf{M}}\bar{\mathbf{Y}}_{-1}^T\otimes\mathbf{\Sigma}^{-(k)}+\xi\mathrm{diag}\left(\mathbf{q}^{(k)}\right)\otimes\mathbf{I}_K$, and $\mathbf{H}^{(k)}=\mathbf{\Sigma}^{-(k)}\Delta\bar{\mathbf{Y}}\bar{\mathbf{M}}\bar{\mathbf{Y}}_{-1}^T$. Although $\overline{F}_1\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$ is a quadratic function in $\mathbf{\Pi}$, together with the nonconvex rank constraint on $\mathbf{\Pi}$ in (3.2.4), the problem is still hard to solve.

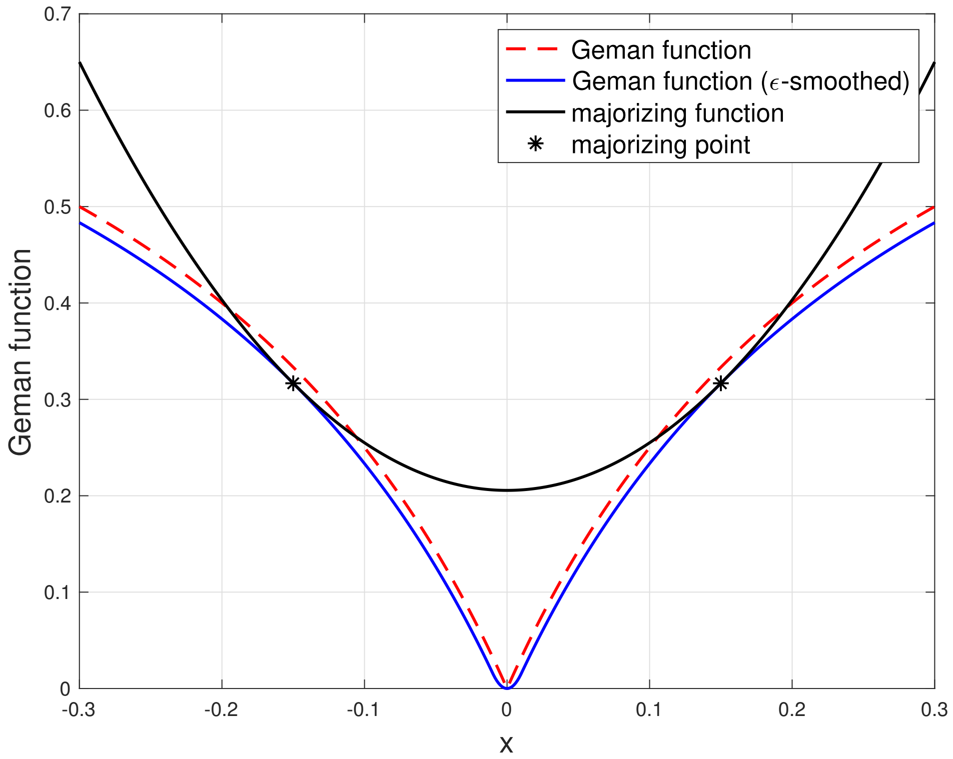


Figure 3.1: Majorization for smoothed sparsity-inducing function.

**Lemma 8.** *Let* $\mathbf{A},\mathbf{B}\in\mathbb{S}^K$ *and* $\mathbf{B}\succeq\mathbf{A}$*, then at any point* $\mathbf{x}^{(k)}\in\mathbb{R}^K$*,* $\mathbf{x}^T\mathbf{A}\mathbf{x}\leq\mathbf{x}^T\mathbf{B}\mathbf{x}+2\mathbf{x}^{(k)T}\left(\mathbf{A}-\mathbf{B}\right)\mathbf{x}+\mathbf{x}^{(k)T}\left(\mathbf{B}-\mathbf{A}\right)\mathbf{x}^{(k)}$ *with the equality attained at* $\mathbf{x}=\mathbf{x}^{(k)}$.

Based on Lemma 8 and noticing $\psi_{\mathbf{G}}^{(k)}\mathbf{I}_{K^2}\succeq\mathbf{G}^{(k)}$ for any $\psi_{\mathbf{G}}^{(k)}$ satisfying $\psi_{\mathbf{G}}^{(k)}\geq\lambda_{\max}\left(\mathbf{G}^{(k)}\right)$, $\overline{F}_1\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)$ can be further majorized by the following function:

$$\overline{F}_2\left(\mathbf{\Pi},\boldsymbol{\theta}^{(k)}\right)\simeq\tfrac{1}{2}\psi_{\mathbf{G}}^{(k)}\left\|\mathbf{\Pi}-\mathbf{P}^{(k)}\right\|_F^2,$$

where $\mathbf{P}^{(k)}=\mathbf{\Pi}^{(k)}-\psi_{\mathbf{G}}^{-(k)}\mathbf{\Sigma}^{-(k)}\mathbf{\Pi}^{(k)}\bar{\mathbf{Y}}_{-1}\bar{\mathbf{M}}\bar{\mathbf{Y}}_{-1}^T-\xi\psi_{\mathbf{G}}^{-(k)}\mathbf{\Pi}^{(k)}\mathrm{diag}\left(\mathbf{q}^{(k)}\right)+\psi_{\mathbf{G}}^{-(k)}\mathbf{H}^{(k)}$.

Finally, the majorized subproblem for problem (3.2.4) is

$$\underset{\mathbf{\Pi}}{\mathsf{minimize}}\ \left\|\mathbf{\Pi}-\mathbf{P}^{(k)}\right\|_F^2\ \mathsf{subject\ to}\ \mathrm{rank}\left(\mathbf{\Pi}\right)\leq r. \tag{3.3.1}$$

This problem has a closed form solution. Let the singular value decomposition for $\mathbf{P}$ be $\mathbf{P}=\mathbf{U}\mathbf{S}\mathbf{V}^T$, the optimal $\mathbf{\Pi}$ is $\mathbf{\Pi}^\star=\mathbf{U}\mathbf{S}_r\mathbf{V}^T$, where $\mathbf{S}_r$ is obtained by thresholding the smallest $(P-r)$ diagonal elements in $\mathbf{S}$ to be zeros. Accordingly, parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be factorized by $\mathbf{\Pi}^\star=\boldsymbol{\alpha}^\star\boldsymbol{\beta}^{\star T}$.

### 3.3.4 The MM-RSVECM Algorithm

Based on the MM method, to solve the original problem (3.2.4), we just need to iteratively solve a low-rank approximation problem (3.3.1) with a closed form solution at each iteration. The overall algorithm is summarized in Algorithm 3.1.

**Algorithm 3.1** MM-RSVECM - Robust MLE of Sparse VECM

```
Input: $\{\mathbf{y}_i\}_{i=1}^{N}$ and needed pre-sampled values.
Initialization: $\boldsymbol{\Pi}^{(0)}\left(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}\right)$, $\boldsymbol{\Gamma}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$ and $k = 1$.
Repeat
    1. Compute $\mathbf{w}^{(k)}$, $\mathbf{q}^{(k)}$, $\mathbf{G}^{(k)}$, $\mathbf{H}^{(k)}$, $\psi_{\mathbf{G}}^{(k)}$ and $\mathbf{P}^{(k)}$;
    2. Update $\boldsymbol{\Pi}^{(k)}$ by solving (3.3.1) and $\boldsymbol{\Gamma}^{(k)}$, $\boldsymbol{\Sigma}^{(k)}$;
    3. $k = k + 1$;
Until $\boldsymbol{\Pi}^{(k)}$, $\boldsymbol{\Gamma}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ satisfy a termination criterion.
Output: $\hat{\boldsymbol{\Pi}}\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right)$, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$.
```

## 3.4 Numerical Simulations

Numerical simulations are considered in this section. A VECM $(K = 5,\ r = 3,\ N = 1000)$ with underlying group sparse structure for $\boldsymbol{\Pi}$ is specified firstly. Then a time series sample path is generated with innovations distributed to Student $t$-distribution with degree of freedom $p = 3$. We first compare our algorithm (MM-RSVECM) with the gradient descent algorithm (GD-RSVECM) for the proposed nonconvex problem formulation in (3.2.4). The convergence result of the objective function value is shown in Fig. 3.2.

Based on the MM method, MM-RSVECM obtains a faster convergence than GD-RSVECM. This may be because the algorithm based on the MM method better exploits the structure of the original problem.

Then the proposed problem formulation based on Cauchy log-likelihood loss function is further validated by comparing the parameter estimation accuracy under student $t$-distributions with different degree of freedom $p$. The estimation accuracy is measure by the normalized mean squared error (NMSE):

$$\text{NMSE}\left(\mathbf{\Pi}\right)=\frac{\mathbb{E}\left[\left\|\hat{\mathbf{\Pi}}-\mathbf{\Pi}_{\text{true}}\right\|_{F}^{2}\right]}{\left\|\mathbf{\Pi}_{\text{true}}\right\|_{F}^{2}}.$$

In Fig. 3.3, we show the simulation results for NMSE $(\mathbf{\Pi})$ by using three estimation methods, which are based on Gaussian innovation assumption, true Student $t$-distribution, and the proposed Cauchy innovation assumption.

From Fig. 3.3, we can see the parameter estimated from Cauchy assumption using the MM-VECM algorithm consistently has a lower parameter estimation error compared to the estimation results from Gaussian assumption and even using the true Student $t$-distribution. Based on this, the proposed problem formulation is validated.

## 3.5 Chapter Summary and Conclusions

This chapter has considered the robust and sparse VECM estimation problem. The problem has been formulated by considering a robust Cauchy log-likelihood loss function and a nonconvex group sparsity regularizer. An efficient algorithm based on the MM method has been proposed with the efficiency of the algorithm and the estimation accuracy validated through numerical simulations.

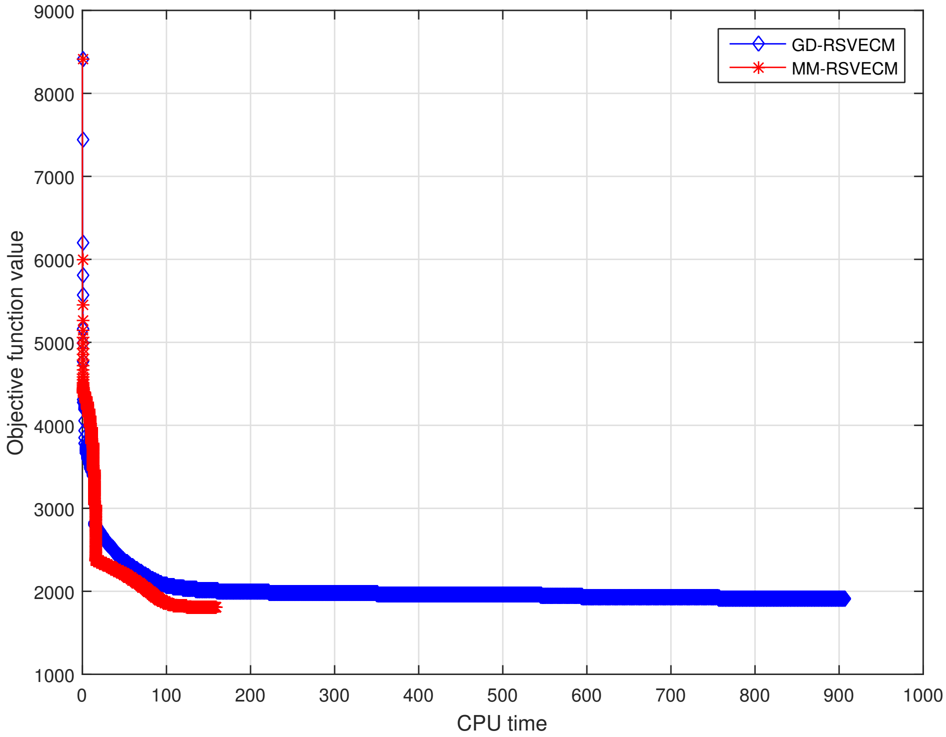


Figure 3.2: Convergence comparison for objective function value.

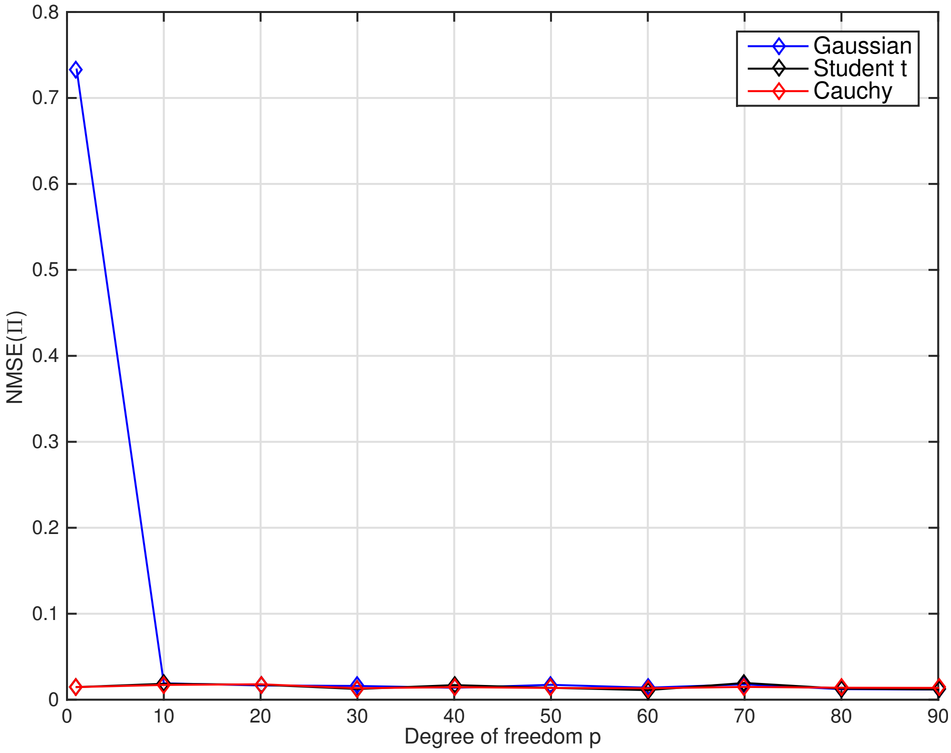


Figure 3.3: NMSE $(\mathbf{\Pi})$ vs degree of freedom $p$ for $t$-distributions.

# Chapter 4

# Mean-Reverting Portfolio Design With A Budget Constraint

This chapter considers the mean-reverting portfolio (MRP) design problem arising from statistical arbitrage (a.k.a. pairs trading) in the financial markets. It aims at designing a portfolio of underlying assets by optimizing the mean reversion strength of the portfolio, while taking into consideration the portfolio variance and an investment budget constraint. Several specific design problems are considered based on different mean reversion criteria. Efficient algorithms are proposed to solve the problems. Numerical results on both synthetic and market data show that the proposed MRP design methods can generate consistent profits and outperform the traditional design methods and the benchmark methods in the literature.

## 4.1 Introduction

Pairs trading [46], [47], [48], [49], also known as spread trading [50], [51], [52], [53], is a famous investment and trading strategy pioneered by scientists Gerry Bamberger and David Shaw, as well as the quantitative trading group led by Nunzio Tartaglia at Morgan Stanley in the mid 1980s. As indicated by the name, it is a trading strategy that focuses on a pair of assets at the same time rather than a single one. Investors or arbitrageurs embracing this strategy do not need to forecast the absolute price of every single asset within one trading pair, which by nature is difficult, but only the relative price of this pair. As a contrarian investment strategy, in order to arbitrage from the market, investors should buy the under-priced asset and short-sell the over-priced one. Profits will be locked in after the trading positions are unwound when the relative mispricing of the pair corrects itself in the future.

More generally, pairs trading with only two trading assets falls into the umbrella of statistical arbitrage [32], [54], [55], [56], also referred to as *stat. arb.*, where the underlying trading basket could consist of three or more financial assets of many kinds such as equities, options, bonds, futures, commodities, etc. Statistical arbitrage opportunities exist as a result of the market inefficiency. Since such strategies can hedge the overall market or systematic risk, and profits do not depend on the movements and conditions of the general financial markets, it is also a kind of market neutral strategy [57], [58]. Nowadays, statistical arbitrage is widely used by many parties in the financial markets, e.g., institutional investors, hedge funds, proprietary trading firms, and individual investors [59].

There are many ways to construct a trading basket, where the cointegration-based method is a prominent one. In [27], [28], the authors first came up with the concept of "cointegration" to describe the linear stationary and hence mean-reverting relationship of the underlying nonstationary time series which are named to be cointegrated. Later, the cointegrated vector autoregressive model was proposed to incorporate such cointegration relations in time series modeling [12], [30]. Empirical and technical analyses show that such relations exist in different financial markets and can be used to get arbitrage opportunities [60], [61], [62], [63], [64]. Taking the prices of common stocks for example, it is generally known that the stock price can be modeled as a nonstationary random walk process which is hard to predict. However, since companies in the same financial sectors or industries usually share similar fundamental characteristics, their stock prices very often move in company with each other under the same trend, and cointegration relations can be established therefrom for arbitrage. Illustrative examples are the two American famous consumer staple companies Coca-Cola and PepsiCo and the two energy companies Ensco and Noble Corporation. Examples for other financial assets, to name a few, are the future contract prices of E-mini S&P 500 and E-mini Dow, the ETF prices of SPDR S&P 500 and SPDR DJIA, the US dollar foreign exchange rates for different countries, the swap rates for US interest rates of different maturities, and so on.

Mean reversion is a classic indicator of predictability in financial markets. Assets in one cointegration relation can be used to form a portfolio or basket and traded based upon the mean reversion property therein. We call such a designed portfolio a mean-reverting portfolio (MRP) or sometimes a long-short portfolio which is also named a "spread". An asset that naturally shows stationarity is a spread as well, e.g., the option implied volatility for stocks. The profits of statistical arbitrage come directly from trading on the mean reversion of a spread around its long-run equilibrium. MRPs in practice are usually constructed using heuristic or statistical methods. Traditional statistical methods are the Engle-Granger ordinary least squares (OLS) method [28] and the Johansen model-based method [30]. In practice, inherent correlations may exist among different spreads. For example, the spreads estimated from the Johansen method which essentially forms a "cointegration subspace". When having multiple MRPs, instead of trading them separately neglecting the possible connections, a natural and interesting question is whether we can design an optimized MRP based on the underlying spreads which could outperform every single one. In this chapter, this issue is addressed.

Designing one MRP by choosing proportions of various assets falls within the umbrella of portfolio optimization or asset allocation problem [65]. Portfolio optimization is important in portfolio management as well as in algorithmic trading in the financial industry. The seminal chapter [66] by Markowitz in 1952 laid on the foundations of what is now popularly referred to as the mean-variance portfolio and the modern portfolio theory. Given a collection of financial assets, the mean-variance portfolio design problem is aimed at finding a tradeoff between the expected return and the risk. Different from that, to design a mean-reverting portfolio, there are two main factors to consider: i) the designed MRP should exhibit a strong mean reversion indicating that it has frequent mean-crossing points and hence bring in trading opportunities, and ii) the designed MRP should exhibit sufficient but controlled variance so that each trade can provide enough profit while controlling the probability that the expected mean reversion equilibrium does not break down. In [67], the author first proposed to design an MRP by optimizing a criterion characterizing the mean reversion strength, and portfolios for swaps and foreign exchange rates were designed. Later, authors in [68], [69] realized that solving the MRP design problem in [67] could result in a portfolio with very low variance, then the variance control was taken into consideration and also new desirable mean reversion criteria were proposed with portfolios for option implied volatilities designed.

The methods proposed in [67], [68], [69] are general and tractable for MRP design. However, they are all carried out by imposing an $\ell_2$-norm constraint on the portfolio weights. The $\ell_2$-norm has a physical meaning of power constraint in wireless communications and used as a similarity constraint in radar signal processing, but its practical significance in financial applications is unclear since the $\ell_2$-norm on portfolio weights do not carry a physical meaning in a financial context. In practice, for portfolio design, the constraint on portfolio weights should represent the investment policy and allocation [70]. So, in this chapter, we propose to use the investment budget constraints which explicitly represent the budget allocation for different assets.

In [67], [68], semidefinite programming relaxation (SDR) methods were used to solve the nonconvex MRP design problems. SDR also has the drawback of squaring the number of variables, which lifts the problem to much higher dimension. Besides that, not every proposed problem formulation in [68] has a tight SDR with zero duality gap, which makes it hard to justify the resulting solution properties. After solving an SDR, randomization-based rank reduction methods, e.g., [71], are typically applied in order to recover a rank-1 feasible solution from a tight SDR for the original problem, which are computationally costly in general. To solve our problem formulations, instead of resorting to SDR, more efficient solving algorithms are developed.

To make it clear, the contributions of this chapter are summarized as follows.

- Based on the mean reversion criteria in [68], [69], the MRP design problem is formulated with a variance constraint and an investment budget constraint (not an $\ell_2$-norm constraint). Two commonly used budget constraints are considered, namely, the dollar neutral constraint and the net budget constraint.

- Efficient algorithms are proposed for problem solving. For some problems, after reformulations they can be readily tackled by solving a quadratically constrained quadratic programming (QCQP), specifically, a generalized eigenvalue problem (GEVP) or a generalized trust region subproblem (GTRS) depending on the constraints.

- Other MRP design problems are efficiently solved based on the majorization-minimization (MM) method by solving a sequence of QCQPs, which are named iteratively reweighted GEVP (IRGEVP) or iteratively reweighted GTRS (IRGTRS). Due to the power of MM, more efficient algorithms, named extended IRGEVP (E-IRGEVP) and extended IRGTRS (E-IRGTRS), are also proposed by solving a quadratically constrained linear programming (QCLP) with a closed-form solution at each iteration.

- The complexity per iteration and convergence properties, like monotonic decreasingness and convergence to a stationary point, are analyzed for the MM-based algorithms.

The remaining sections of this chapter are organized as follows. In Section 4.2, we briefly introduce the MRP. In Section 4.3, the MRP design problem is formulated based on some mean reversion criteria and two investment budget constraints. Section 4.4 introduces the GEVP and GTRS algorithms. The MM-based algorithms are elaborated in Section 4.5 with algorithm complexity and convergence analysis given in Section 4.6. The numerical performance is evaluated in Section 4.7 and, finally, the concluding remarks are drawn in Section 4.8.

## 4.2 Mean-Reverting Portfolio and Mean Reversion Trading

For a financial asset, e.g., a common stock, a future contract, an ETF, or a portfolio of them, its price at time index or holding period $t \in \mathbb{N}$ is denoted by $p_t \in \mathbb{R}_+$, and the corresponding logarithmic price or log-price $y_t \in \mathbb{R}$ is computed as $y_t = \log(p_t)$, where $\log(\cdot)$ is the natural logarithm function. An illustrative example of the log-prices for two security assets denoted as $[y_1, y_2]$ is shown in Figure 4.1.

For one single asset, the (cumulative) return at time $t$ for $\tau$ holding periods is defined as

$$r_t(\tau) = \frac{p_t - p_{t-\tau}}{p_{t-\tau}}, \tag{4.2.1}$$

where $\tau$ denotes the period length and is usually omitted when the length is one. Then we can have

$$\begin{aligned} r_t(\tau) &\approx \log(p_t) - \log(p_{t-\tau}) \\ &= y_t - y_{t-\tau}, \end{aligned} \tag{4.2.2}$$

where the approximation follows from $\log(1+x) \approx x$ for small $x$, which is valid for the usual trading intervals. Here, the return $r_t(\tau)$ as a rate of return is used to measure the aggregate amount of profits or losses (in percentage) of an investment strategy on one asset over a time period $\tau$.

In order to make an profitable investment (i.e., with a positive return) in the financial markets, the investors need either to buy an asset before its price is going up or to sell an asset before its price is going down. However, in many cases, the asset price is hard to predict. It is usually difficult for people to decide the time point to make an investment on the asset.

In statistical arbitrage strategy, rather than investing on a single asset, people invest on a portfolio of assets at the same time. Such a portfolio or spread is stationary and thus easy to choose the time for investment. In practice, spreads can be naturally stationary like option implied volatilities, designed using methods like technical or fundamental analysis, or constructed based on statistical models. In Figure 4.1, a spread designed from two security assets is shown.

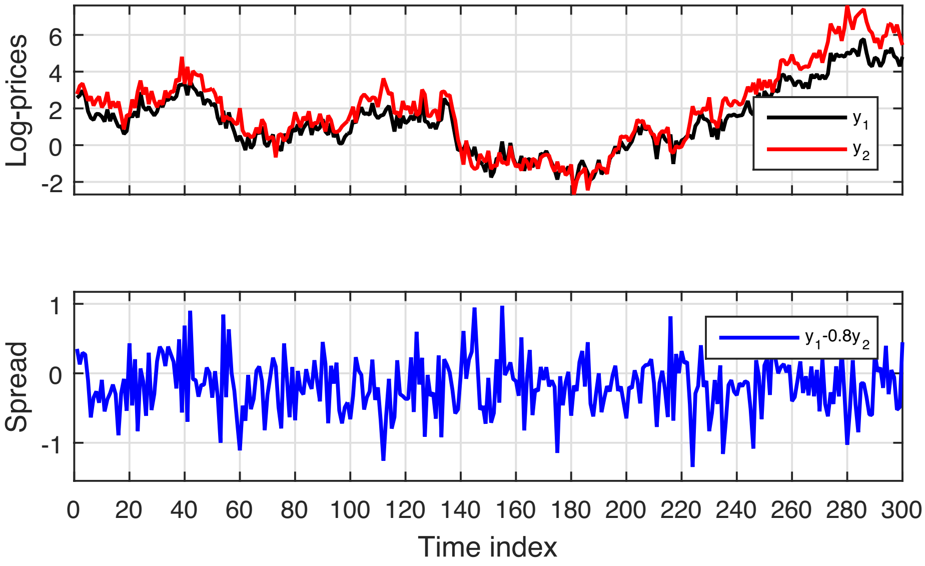


Figure 4.1: An illustrative example of log-prices for two assets and the spread.

### 4.2.1 Mean-Reverting Portfolio (MRP)

Different spreads may possess different mean reversion and variance properties in nature. Our objective is to design an MRP to combine such spreads into an improved overall spread with better properties. Suppose there exist $N$ spreads denoted by $\mathbf{s}_t = [s_{1,t}, s_{2,t}, \ldots, s_{N,t}]^T$. We denote the designed mean-reverting portfolio (MRP) by the portfolio weight or hedge ratio $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$, then the resulting MRP (or spread) is given by

$$z_t = \mathbf{w}^T \mathbf{s}_t = \sum_{n=1}^{N} w_n s_{n,t}, \tag{4.2.3}$$

where vector $\mathbf{w}$ indicates the market value proportion invested on the underlying spreads[1]. For $n = 1, 2, \ldots, N$, $w_n > 0$, $w_n < 0$, and $w_n = 0$ mean a long position (i.e., it is bought),

[1] If the spread is designed based on asset price $p_t$ instead of the log-price, $\mathbf{w}$ indicates the asset amount proportion measured in shares.

a short position (i.e., it is short-sold or, more plainly, borrowed and sold), and no position on the spread, respectively.

When the spread $s_t$ is composed with other underlying financial assets (say, spread from the cointegration model [72]), we can further have the relation between the designed MRP and the underlying financial assets. If a collection of $M$ assets is considered with their log-prices denoted by $\mathbf{y}_t = [y_{1,t}, y_{2,t}, \ldots, y_{M,t}]^T$, and a portfolio is defined by the weights $\mathbf{w}_s = [w_{s,1}, w_{s,2}, \ldots, w_{s,M}]^T$, its (log-price) spread $s_t$ is accordingly given by $s_t = \mathbf{w}_s^T \mathbf{y}_t$. Then if $N$ such spreads are consider as in (4.2.3), we can get the resulting MRP as

$$z_t = \mathbf{w}_p^T \mathbf{y}_t = \sum_{m=1}^{M} w_{p,m} y_{m,t}, \tag{4.2.4}$$

where $\mathbf{w}_p = \mathbf{W}_s \mathbf{w}$ denoting the portfolio weight directly defined on the underlying assets and $\mathbf{W}_s = [\mathbf{w}_{s_1}, \mathbf{w}_{s_2}, \ldots, \mathbf{w}_{s_N}]$.

It is worth noting that an MRP can be interpreted as a synthesized stationary asset. The spread accordingly means the log-price for this MRP, which is much easier to profit from (i.e., to arbitrage) compared to the underlying component assets. The trading strategy to make profits from an MRP is called the mean reversion trading, which is precisely to trade on the mean reversion property of the spread around its equilibrium, i.e., to buy this MRP when the price is lower than its equilibrium and to sell it when the price is higher than its equilibrium.

### 4.2.2 Mean Reversion Trading

In this chapter, we use a simple mean reversion trading strategy where the trading signals, i.e., to buy, to sell, or simply to hold, are designed based on simple event triggers. The trading is carried out on the designed spread $z_t$ which is tested to be unit-root stationary. A trading position (a long position denoted by $1$ and a short position by $-1$) is a state for investment and it is opened when the spread $z_t$ is away from its equilibrium $\mu_z$ by a predefined trading threshold $\Delta$ and closed (denoted by $0$) when $z_t$ crosses its equilibrium $\mu_z$. (A common variation is to close the position after the spread crosses the equilibrium by more than another threshold $\Delta'$.) The time period from position opening to position closing is defined as a trading period.

In order to get a standard trading rule, we use the *z-score*, which is a normalized spread measuring the distance to the spread equilibrium in units of standard deviations as follows:

$$\tilde{z}_t = \frac{z_t - \mu_z}{\sigma_z}, \tag{4.2.5}$$

where $\mu_z$ and $\sigma_z$ are the mean and the standard deviation of the spread $z_t$ and computed over an in-sample look-back period in practice. For $\tilde{z}_t$, we have $\mathsf{E}\left[\tilde{z}_t\right] = 0$ and $\mathsf{Std}\left[\tilde{z}_t\right] = 1$. Then, we can define the threshold as $\Delta = d \times \sigma_z$, for some value of $d$ (e.g., $d = 1$).

In the trading stage, based on the trading position and observed (normalized) spread value at holding period $t$, we can get the trading actions at the next consecutive holding period $t+1$. The mean reversion trading strategy is summarized in Table 4.1 and a simple trading example based on this strategy is illustrated in Figure 4.2.

Table 4.1: Trading Positions, Normalized Spread, and Trading Actions of a Mean Reversion Trading Strategy

| Trading Position at $t$ | Normalized Spread $\tilde{z}_t$ | Action(s) Taken within Holding Period $t+1$ | Trading Position at $t+1$ |
|---|---|---|---|
| 1 | $+d \leq \tilde{z}_t$ | Close the long position & Open a short position | -1 |
| | $0 \leq \tilde{z}_t < +d$ | Close the long position | 0 |
| | $\tilde{z}_t < 0$ | No action | 1 |
| 0 | $+d \leq \tilde{z}_t$ | Open a short position | -1 |
| | $-d < \tilde{z}_t < +d$ | No action | 0 |
| | $\tilde{z}_t \leq -d$ | Open a long position | 1 |
| -1 | $0 < \tilde{z}_t$ | No action | -1 |
| | $-d < \tilde{z}_t \leq 0$ | Close the short position | 0 |
| | $\tilde{z}_t \leq -d$ | Close the short position & Open a long position | 1 |

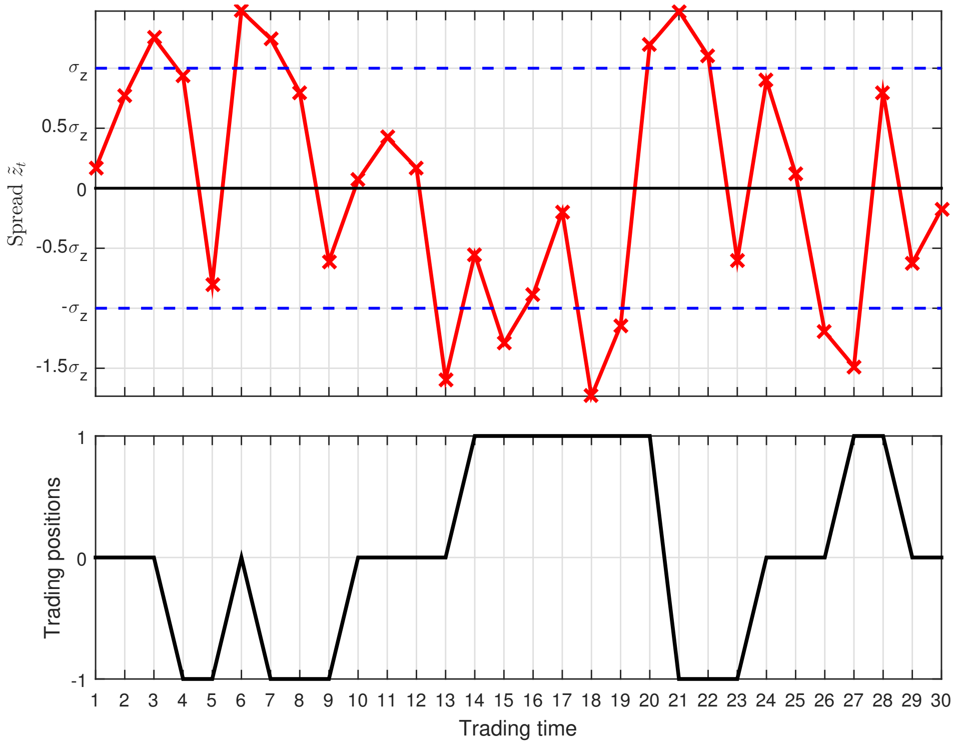


Figure 4.2: An example for mean reversion trading (trading threshold $\Delta = \sigma_z$).

Based on this trading scheme, we can get the profit and loss (P&L) for the MRP which measures the payoff and is also the amount of profits or losses (in units of dollars) of an investment on the portfolio for some holding periods. Within one trading period, if a long position is opened on an MRP at time $t_o$ and closed at time $t_c$, then the multi-period P&L of this MRP at time $t$ ($t_o \leq t \leq t_c$) accumulated from $t_o$ is computed as $\mathrm{P\&L}_t(\tau) = \mathbf{w}_p^T \mathbf{r}_t(\tau) = \mathbf{w}_p^T \mathbf{r}_t(t - t_o)$,[2] where $\tau = t - t_o$ denotes the length of the holding period, and $\mathbf{r}_t(\tau) = [r_{1,t}(\tau), r_{2,t}(\tau), \ldots, r_{M,t}(\tau)]^T$ is the return vector. More generally, the cumulative P&L of this MRP at time $t$ for $\tau$ ($0 \leq \tau \leq t - t_o$) holding periods is defined as

$$\mathrm{P\&L}_t(\tau) = \mathbf{w}_p^T \mathbf{r}_t(t - t_o) - \mathbf{w}_p^T \mathbf{r}_{t-\tau}(t - \tau - t_o), \tag{4.2.6}$$

where we define $\mathbf{r}_t(0) = \mathbf{0}$. Then we have the single-period P&L (e.g., daily P&L, monthly P&L) denoted by $\mathrm{P\&L}_t$ at time $t$ (i.e., $\tau = 1$) is computed as

$$\mathrm{P\&L}_t = \mathbf{w}_p^T \mathbf{r}_t(t - t_o) - \mathbf{w}_p^T \mathbf{r}_{t-1}(t - 1 - t_o). \tag{4.2.7}$$

[2] Here $\mathbf{w}_p$ defines the real dollar values for the underlying assets, which is the portfolio weights scaled up by the investment budget.

If, instead, a short position is opened on this MRP, then multi-period P&L is $\mathrm{P\&L}_t\left(\tau\right) = \mathbf{w}_p^T \mathbf{r}_{t-\tau}\left(t-\tau-t_o\right) - \mathbf{w}_p^T \mathbf{r}_t\left(t-t_o\right)$ and the single-period P&L is $\mathrm{P\&L}_t = \mathbf{w}_p^T \mathbf{r}_{t-1}\left(t-1-t_o\right) - \mathbf{w}_p^T \mathbf{r}_t\left(t-t_o\right)$. About the portfolio P&L calculation within the trading periods, we have the following lemma.

**Lemma 9.** *Within one trading period, if the price change of every asset in an MRP is small enough, then the P&L in* (4.2.6) *can be approximately calculated by the change of the log-price spread* $z_t$*. Specifically,*

*1) for a long position on the MRP,* $\mathrm{P\&L}_t\left(\tau\right) \approx z_t - z_{t-\tau}$*; and*

*2) for a short position on the MRP,* $\mathrm{P\&L}_t\left(\tau\right) \approx z_{t-\tau} - z_t$*.*

**Proof 1.** *See Appendix 4.9.*

In fact, Lemma 9 reveals the philosophy behind the MRP design problem and also the mean reversion trading by showing the connection between the log-price spread value and the computation of the portfolio return.

## 4.3 Problem Formulation for MRP Design

The traditional mean-variance portfolio which is based on the Nobel prize-winning Markowitz portfolio theory [66], [73] aims at finding a desired trade-off between return and risk, with the latter being measured by the variance. For the mean-reverting portfolio design, we formulate the problem by optimizing a mean reversion criterion quantifying the mean reversion strength [68], [69], while controlling its variance and imposing an investment budget constraint.

### 4.3.1 Mean Reversion Criteria

In this section, we introduce several mean reversion criteria that can characterize the mean reversion strength of the designed spread $z_t$. We start by defining the $i$th order (lag-$i$) auto-covariance matrix for a stochastic process $\mathbf{s}_t$ as

$$\mathbf{M}_i = \mathsf{Cov}\Big(\mathbf{s}_t, \mathbf{s}_{t+i}\Big) = \mathsf{E}\left[\Big(\mathbf{s}_t - \mathsf{E}\left[\mathbf{s}_t\right]\Big)\Big(\mathbf{s}_{t+i} - \mathsf{E}\left[\mathbf{s}_{t+i}\right]\Big)^T\right],$$

where $i \in \mathbb{N}$. Specifically, when $i = 0$, $\mathbf{M}_0$ stands for the (positive definite) covariance matrix of $\mathbf{y}_t$.

Since for any random process $\mathbf{s}_t$, we can always get its centered form as $\tilde{\mathbf{s}}_t = \mathbf{s}_t - \mathsf{E}[\mathbf{s}_t]$, without loss of generality, we use $\mathbf{s}_t$ to denote its centered counterpart $\tilde{\mathbf{s}}_t$ in the following.

#### 4.3.1.1 Predictability Statistics $\mathrm{pre}(\mathbf{w})$

Consider a centered univariate stationary autoregressive process $z_t = \hat{z}_{t-1} + \epsilon_t$, where $\hat{z}_{t-1}$ is the prediction of $z_t$ based on the information up to time $t-1$, and $\epsilon_t$ denotes a white noise independent from $\hat{z}_{t-1}$. The predictability statistics [74] is defined as

$$\mathrm{pre} = \frac{\sigma_{\hat{z}}^2}{\sigma_z^2}, \tag{4.3.1}$$

where $\sigma_z^2 = \mathsf{E}[z_t^2]$ and $\sigma_{\hat{z}}^2 = \mathsf{E}\left[\hat{z}_{t-1}^2\right]$. If we define $\sigma_\epsilon^2 = \mathsf{E}[\epsilon_t^2]$, then we have $\sigma_z^2 = \sigma_{\hat{z}}^2 + \sigma_\epsilon^2$ in the denominator. When $\mathrm{pre}$ is small, the variance of $\epsilon_t$ dominates that of $\hat{z}_{t-1}$, and $z_t$ behaves like a white noise; when $\mathrm{pre}$ is large, the variance of $\hat{z}_{t-1}$ dominates that of $\epsilon_t$, and $z_t$ can be well predicted by $\hat{z}_{t-1}$. The predictability statistics is usually used to measure how close a random process is to a white noise.

Based on this criterion, in order to design a spread $z_t$ as close as possible to a white noise process, we need to minimize $\mathrm{pre}$ in (4.3.1). For $z_t = \mathbf{w}^T\mathbf{s}_t$, we assume the spread $\mathbf{s}_t$ follows a centered vector autoregressive model of order $1$ (VAR(1)) as $\mathbf{s}_t = \mathbf{A}\mathbf{s}_{t-1} + \mathbf{e}_t$, where $\mathbf{A}$ is the autoregressive coefficient and $\mathbf{e}_t$ denotes a white noise independent from $\mathbf{s}_{t-1}$. Then we can get $\mathbf{A} = \mathbf{M}_1^T\mathbf{M}_0^{-1}$. Premultiplying the VAR(1) by $\mathbf{w}$ and defining $\hat{z}_{t-1} = \mathbf{w}^T\mathbf{A}\mathbf{s}_{t-1}$ and $\epsilon_t = \mathbf{w}^T\mathbf{e}_t$, we have $\sigma_z^2 = \mathbf{w}^T\mathbf{M}_0\mathbf{w}$ and $\sigma_{\hat{z}}^2 = \mathbf{w}^T\mathbf{T}\mathbf{w}$ with $\mathbf{T} = \mathbf{A}\mathbf{M}_0\mathbf{A}^T = \mathbf{M}_1^T\mathbf{M}_0^{-1}\mathbf{M}_1$. This also applies to high order models VAR($p$) ($p > 1$) through proper reparametrization [26]. Then the predictability statistics for $z_t$ is computed as

$$\mathrm{pre}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{T}\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}. \tag{4.3.2}$$

#### 4.3.1.2 Portmanteau Statistics $\mathrm{por}(p, \mathbf{w})$

The portmanteau statistics of order $p$ [75] for a centered univariate stationary process $z_t$ is defined as

$$\mathrm{por}(p) = \sum_{i=1}^{p} \rho_i^2, \tag{4.3.3}$$

where $\rho_i$ is the $i$th order (lag-$i$) autocorrelation of $z_t$ defined as $\rho_i = \mathsf{E}[z_t z_{t+i}] / \mathsf{E}[z_t^2]$. The portmanteau statistics is used to test whether a random process is close to a white noise.

From (4.3.3), we have $\operatorname{por}(p) \geq 0$ and the minimum is attained by a white noise, i.e., the portmanteau statistics for a white noise process is 0 for any $p$.

Based on this criterion, in order to get a spread $z_t$ close to a white noise process, we need to minimize $\operatorname{por}_z(p)$ for a prespecified order $p$. For an MRP $z_t = \mathbf{w}^T\mathbf{s}_t$, the $\rho_i = \mathbf{w}^T\mathsf{E}\left[\mathbf{s}_t\mathbf{s}_{t+i}^T\right]\mathbf{w}/\mathbf{w}^T\mathsf{E}\left[\mathbf{s}_t\mathbf{s}_t^T\right]\mathbf{w} = \mathbf{w}^T\mathbf{M}_i\mathbf{w}/\mathbf{w}^T\mathbf{M}_0\mathbf{w}$. Then we can get the expression for $\operatorname{por}(p, \mathbf{w})$ as

$$\operatorname{por}(p, \mathbf{w}) = \sum_{i=1}^{p}\left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2. \tag{4.3.4}$$

#### 4.3.1.3 Crossing Statistics $\operatorname{cro}(\mathbf{w})$ and Penalized Crossing Statistics $\operatorname{pcro}(p, \mathbf{w})$

Crossing statistics (zero-crossing rate) of a centered stationary process $z_t$ is defined as $\operatorname{zcr} = 1/(T-1)\sum_{t=2}^{T}\mathbf{1}_E(z_t)$, where the indicator function $\mathbf{1}_E(z_t) = \begin{cases} 1, & z_t \in E \\ 0, & z_t \notin E \end{cases}$ with event $E = \{z_t z_{t-1} \leq 0\}$. It is used to test the probability that a process crosses its mean per unit of time. According to [76], [77], for a centered stationary Gaussian process, $\operatorname{zcr} = 1/\pi \arccos(\rho_1)$.

Based on this criterion, in order to get a spread $z_t$ having many zero-crossings, we can minimize $\rho_1$. So for a spread $z_t = \mathbf{w}^T\mathbf{s}_t$, we define the crossing statistics as

$$\operatorname{cro}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}. \tag{4.3.5}$$

In [68], besides minimizing $\operatorname{cro}(\mathbf{w})$, it is also proposed to ensure the absolute high order autocorrelations $|\rho_i|$'s ($i = 2, \ldots, p$) are small which can result in good performance. In this chapter, we denote this criterion as the penalized crossing statistics of order $p$ as

$$\operatorname{pcro}(p, \mathbf{w}) = \frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} + \eta\sum_{i=2}^{p}\left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2, \tag{4.3.6}$$

where $\eta$ is a positive prespecified penalization factor.

### 4.3.2 Investment Budget Constraint

In this chapter, two types of budget constraints are considered, namely, dollar neutral constraint and net budget constraint.

The dollar neutral constraint, denoted by $\mathcal{W}_0$, means the net investment or net portfolio position is zero; in other words, all the long positions are financed by the short positions, commonly termed self-financing.[3] The portfolio in this case is called zero-cost portfolio. It is represented mathematically by

$$\mathcal{W}_0 = \left\{\mathbf{1}^T\mathbf{w} = 0\right\}. \tag{4.3.7}$$

The net budget constraint, denoted by $\mathcal{W}_1$, means the net investment or net portfolio position is nonzero and equal to the current budget which is normalized to one.[4] It is represented mathematically by

$$\mathcal{W}_1 = \left\{\mathbf{1}^T\mathbf{w} = 1\right\}. \tag{4.3.8}$$

It is worth noting that the two trading spreads defined by $\mathbf{w}^T\mathbf{y}_t$ and $-\mathbf{w}^T\mathbf{y}_t$ are naturally the same, because in statistical arbitrage the actual investment not only depends on $\mathbf{w}$, which defines a spread, but also on whether a long or short position is taken on this spread later in the trading stage.

### 4.3.3 General MRP Design Problem Formulation

To make the illustration for the MRP design problem clear in the following, we denote the mean reversion criterion in a compact form as $F(\mathbf{w})$ that takes all the aforementioned criteria into account as follows:

$$F(\mathbf{w}) = \xi\frac{\mathbf{w}^T\mathbf{H}\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} + \zeta\left(\frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2 + \eta\sum_{i=2}^{p}\left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2, \tag{4.3.9}$$

which particularizes to i) $\mathrm{pre}(\mathbf{w})$, when $\xi = 1$, $\mathbf{H} = \mathbf{T}$, and $\zeta = \eta = 0$; ii) $\mathrm{por}(p, \mathbf{w})$, when $\xi = 0$, and $\zeta = \eta = 1$; iii) $\mathrm{cro}(\mathbf{w})$, when $\xi = 1$, $\mathbf{H} = \mathbf{M}_1$, and $\zeta = \eta = 0$; and iv) $\mathrm{pcro}(p, \mathbf{w})$, when $\xi = 1$, $\mathbf{H} = \mathbf{M}_1$, $\zeta = 0$, and $\eta > 0$. The matrices $\mathbf{M}_i$'s in (4.3.9) are assumed symmetric without loss of generality since they can always be symmetrized. As mentioned before, the variance of the spread should be controlled to a certain level which is represented as $\mathsf{Var}\left[\mathbf{w}^T\mathbf{s}_t\right] = \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu$ with $\nu$ being a predefined positive constant. Due to this variance constraint, the denominators in $F(\mathbf{w})$ can be reduced. Denoting the portfolio

[3] The dollar neutral constraint generally cannot be satisfied by the traditional design methods, like methods in [28] and [30], and the methods in [68].

[4] The net portfolio position can be positive or negative under net budget constraint. Since the problem formulation in (4.3.10) is invariant to the sign of $\mathbf{w}$, only the case that budget is normalized to positive 1 is considered.

investment budget constraint by $\mathcal{W}$, the general MRP design problem can be formulated as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\mathsf{mininize}} \quad & \underbrace{\xi \mathbf{w}^T \mathbf{H} \mathbf{w} + \zeta \left(\mathbf{w}^T \mathbf{M}_1 \mathbf{w}\right)^2 + \eta \sum_{i=2}^{p} \left(\mathbf{w}^T \mathbf{M}_i \mathbf{w}\right)^2}_{\triangleq f(\mathbf{w})} \\ \mathsf{subject\ to} \quad & \mathbf{w}^T \mathbf{M}_0 \mathbf{w} = \nu \\ & \mathbf{w} \in \mathcal{W}_i, \ (i = 0, 1) . \end{aligned} \tag{4.3.10}$$

The MRP design problem (4.3.10) is a nonconvex smooth constrained optimization problem [78] with highly nonconvex (quartic or quadratic) objective function and nonconvex constraint set. To solve the problem, efficient, effective, and convergent algorithms are designed in the following sections.

## 4.4 Problem Solving via GEVP and GTRS Algorithms

In this section, solving methods for the MRP design problem formulations using $\mathrm{pre}\left(\mathbf{w}\right)$ and $\mathrm{cro}\left(\mathbf{w}\right)$ (i.e., (4.3.10) with $\zeta = \eta = 0$) are introduced.

### 4.4.1 GEVP: Solving Algorithm for MRP Design Using $\mathrm{pre}\left(\mathbf{w}\right)$ and $\mathrm{cro}\left(\mathbf{w}\right)$ with $\mathbf{w} \in \mathcal{W}_0$

We recast the relevant problems in (4.3.10) as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\mathsf{minimize}} \quad & \mathbf{w}^T \mathbf{H} \mathbf{w} \\ \mathsf{subject\ to} \quad & \mathbf{w}^T \mathbf{M}_0 \mathbf{w} = \nu \\ & \mathbf{1}^T \mathbf{w} = 0, \end{aligned} \tag{4.4.1}$$

By rewriting the constraint $\mathbf{1}^T \mathbf{w} = 0$ as $\mathbf{w}^T \mathbf{1} \mathbf{1}^T \mathbf{w} = 0$ (since the problem is invariant to a sign change in $\mathbf{w}$) and using the matrix lifting technique (i.e., $\mathbf{W} = \mathbf{w}\mathbf{w}^T$), the above

**Algorithm 4.1** GEVP - Algorithm for MRP design problems using $\mathrm{pre}(\mathbf{w})$ and $\mathrm{cro}(\mathbf{w})$ with $\mathbf{w} \in \mathcal{W}_0$.

```
Require: $\mathbf{N}$, $\mathbf{N}_0$, and $\nu$.
Set $k = 0$ and $\mathbf{x}^{(0)} \in \{\mathbf{x} \mid \mathbf{x}^T\mathbf{N}_0\mathbf{x} = \nu\}$;
repeat
    $R(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)T}\mathbf{N}\mathbf{x}^{(k)}/\mathbf{x}^{(k)T}\mathbf{N}_0\mathbf{x}^{(k)}$;
    $\mathbf{d}^{(k)} = \mathbf{N}\mathbf{x}^{(k)} - R(\mathbf{x}^{(k)})\mathbf{N}_0\mathbf{x}^{(k)}$;
    $\hat{\mathbf{x}} = \mathbf{x}^{(k)} + \tau\mathbf{d}^{(k)}$ with $\tau$ minimizing $R(\mathbf{x}^{(k)} + \tau\mathbf{d}^{(k)})$;
    $\mathbf{x}^{(k+1)} = \sqrt{\nu}\hat{\mathbf{x}}/\sqrt{\hat{\mathbf{x}}^T\mathbf{N}_0\hat{\mathbf{x}}}$;
    $k = k + 1$;
until convergence
```

problem can be solved by the following SDR:

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \quad & \mathrm{Tr}\left(\mathbf{H}\mathbf{W}\right) \\ \text{subject to} \quad & \mathrm{Tr}\left(\mathbf{M}_0\mathbf{W}\right) = \nu \\ & \mathrm{Tr}\left(\mathbf{1}\mathbf{1}^T\mathbf{W}\right) = 0 \\ & \mathbf{W} \succeq \mathbf{0}. \end{aligned} \tag{4.4.2}$$

Although problem (4.4.1) is nonconvex, it has no duality gap [79], [80]. In other words, by solving the SDR (4.4.2), a rank-1 solution for $\mathbf{W}$ always exists which is a feasible global optimal solution for (4.4.1).

As an alternative to the SDR procedure, the optimal solution for (4.4.1) can be efficiently solved by reformulating it as a nonconvex QCQP. Considering $\mathbf{w} = \mathbf{F}\mathbf{x}$, where $\mathbf{F}$ is a left-invertible matrix that lies on the null space of $\mathbf{1}^T$ (i.e., $\mathbf{1}^T\mathbf{F} = \mathbf{0}$), we define $\mathbf{N} = \mathbf{F}^T\mathbf{H}\mathbf{F}$ and $\mathbf{N}_0 = \mathbf{F}^T\mathbf{M}_0\mathbf{F}$, then problem (4.4.1) is equivalent to the following one:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{x}^T\mathbf{N}\mathbf{x} \\ \text{subject to} \quad & \mathbf{x}^T\mathbf{N}_0\mathbf{x} = \nu. \end{aligned} \tag{4.4.3}$$

This QCQP problem is also known as a generalized eigenvalue problem (GEVP) [81] which can be efficiently solved by tailored algorithms. We choose the steepest descent algorithm in [82] to solve it, which is summarized in Algorithm 4.1.

### 4.4.2 GTRS: Solving Algorithm for MRP Design Using $\mathrm{pre}(\mathbf{w})$ and $\mathrm{cro}(\mathbf{w})$ with $\mathbf{w} \in \mathcal{W}_1$

The relevant problems in (4.3.10) can be rewritten as

$$\begin{array}{ll} \underset{\mathbf{w}}{\mathsf{minimize}} & \mathbf{w}^T\mathbf{H}\mathbf{w} \\ \mathsf{subject\ to} & \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu \\ & \mathbf{1}^T\mathbf{w} = 1, \end{array} \tag{4.4.4}$$

Again, problem (4.4.4) can be solved by the SDR. It can also be efficiently solved as a QCQP. Considering $\mathbf{w} = \mathbf{F}\mathbf{x} + \mathbf{w}_0$ where $\mathbf{F}$ is defined as before and $\mathbf{w}_0$ is (any) particular solution of $\mathbf{1}^T\mathbf{w} = 1$, and defining $\mathbf{N} = \mathbf{F}^T\mathbf{H}\mathbf{F}$, $\mathbf{p} = \mathbf{F}^T\mathbf{H}\mathbf{w}_0$, $b = \mathbf{w}_0^T\mathbf{H}\mathbf{w}_0$, $\mathbf{N}_0 = \mathbf{F}^T\mathbf{M}_0\mathbf{F}$, $\mathbf{p}_0 = \mathbf{F}^T\mathbf{M}_0\mathbf{w}_0$, and $b_0 = \mathbf{w}_0^T\mathbf{M}_0\mathbf{w}_0$, the problem (4.4.4) is equivalent to the following nonconvex QCQP:

$$\begin{array}{ll} \underset{\mathbf{x}}{\mathsf{minimize}} & \mathbf{x}^T\mathbf{N}\mathbf{x} + 2\mathbf{p}^T\mathbf{x} + b \\ \mathsf{subject\ to} & \mathbf{x}^T\mathbf{N}_0\mathbf{x} + 2\mathbf{p}_0^T\mathbf{x} + b_0 = \nu. \end{array} \tag{4.4.5}$$

This QCQP is specially named generalized trust region subproblem (GTRS) [83], [84]. Such problem is usually nonconvex but possesses necessary and sufficient optimality conditions. Efficient solving algorithms for global optimal solution based on the matrix pencil technique can be derived. According to Theorem 3.2 in [83], the optimality conditions for the primal and dual variables $(\mathbf{x}^\star, \xi^\star)$ of problem (4.4.5) are given as follows:

$$\begin{cases} \left(\mathbf{N} + \xi^\star\mathbf{N}_0\right)\mathbf{x}^\star + \mathbf{p} + \xi^\star\mathbf{p}_0 = 0, \\ \mathbf{x}^{\star T}\mathbf{N}_0\mathbf{x}^\star + 2\mathbf{p}_0^T\mathbf{x}^\star + b_0 - \nu = 0, \\ \mathbf{N} + \xi^\star\mathbf{N}_0 \succeq \mathbf{0}. \end{cases} \tag{4.4.6}$$

Assuming $\mathbf{N} + \xi\mathbf{N}_0 \succ \mathbf{0}$,[5] the optimal solution is given by

$$\mathbf{x}(\xi) = -\left(\mathbf{N} + \xi\mathbf{N}_0\right)^{-1}\left(\mathbf{p} + \xi\mathbf{p}_0\right), \tag{4.4.7}$$

[5] The limiting case $\mathbf{N} + \xi\mathbf{N}_0$ being singular (i.e., $\xi = -\lambda_{\min}(\mathbf{N}, \mathbf{N}_0)$) can be treated separately. The assumption here is reasonable since the case when $\xi = -\lambda_{\min}(\mathbf{N}, \mathbf{N}_0)$ is very rare to occur theoretically and practically.

**Algorithm 4.2** GTRS - Algorithm for MRP design problems using $\mathrm{pre}\,(\mathbf{w})$ and $\mathrm{cro}\,(\mathbf{w})$ with $\mathbf{w} \in \mathcal{W}_1$.

```
Require: $\mathbf{N}$, $\mathbf{N}_0$, $\mathbf{p}$, $\mathbf{p}_0$, $b_0$, and $\nu$.
Compute $\lambda_{\min}(\mathbf{N}, \mathbf{N}_0)$.
Set $k = 0$ and $\xi^{(0)} \in (-\lambda_{\min}(\mathbf{N}, \mathbf{N}_0), \infty)$;
repeat
    $\mathbf{x}^{(k)} = -(\mathbf{N} + \xi^{(k)}\mathbf{N}_0)^{-1}(\mathbf{p} + \xi^{(k)}\mathbf{p}_0)$;
    $\phi(\xi^{(k)}) = \mathbf{x}^{(k)T}\mathbf{N}_0\mathbf{x}^{(k)} + 2\mathbf{p}_0^T\mathbf{x}^{(k)} + b_0 - \nu$;
    Update $\xi^{(k+1)}$ by a line search algorithm;
    $k = k + 1$;
until convergence
```

and $\xi$ is the unique solution for equation $\phi\left(\xi\right) = 0$, where

$$\phi\left(\xi\right) = \mathbf{x}\left(\xi\right)^T \mathbf{N}_0 \mathbf{x}\left(\xi\right) + 2\mathbf{p}_0^T \mathbf{x}\left(\xi\right) + b_0 - \nu, \tag{4.4.8}$$

and $\xi \in \mathcal{I}$. The interval $\mathcal{I} = \{\xi \mid \mathbf{N} + \xi\mathbf{N}_0 \succ \mathbf{0}\}$, which implies $\mathcal{I} = \left(-\lambda_{\min}\left(\mathbf{N}, \mathbf{N}_0\right), \infty\right)$, where $\lambda_{\min}\left(\mathbf{N}, \mathbf{N}_0\right)$ is the minimum generalized eigenvalue of matrix pair $\left(\mathbf{N}, \mathbf{N}_0\right)$. According to Theorem 5.2 in [83], the function $\phi\left(\xi\right)$ is strictly decreasing on $\mathcal{I}$. So based on this property, a one dimensional search method (e.g., bisection algorithm) can be used to find the optimal $\xi$ over $\mathcal{I}$. The algorithm for solving problem (4.4.5) is summarized in Algorithm 4.2.

## 4.5 Problem Solving via MM-Based Algorithms

In this section, we first discuss the MM method briefly, and then two solving algorithms for the MRP design problems using $\mathrm{por}\,(p, \mathbf{w})$ (i.e., (4.3.10) with $\xi = 0$ and $\zeta = \eta = 1$) and $\mathrm{pcro}\,(p, \mathbf{w})$ (i.e., (4.3.10) with $\xi = 1$, $\mathbf{H} = \mathbf{M}_1$, $\zeta = 0$ and $\eta > 0$) are derived based on the MM method together with the GEVP and GTRS algorithms in Section 4.4.

### 4.5.1 The MM Method

The MM method [19], [85] refers to majorization-minimization for minimization problems or minorization-maximization for maximization problems. It is also known as the successive upper bound minimization method [24], [86].

For an optimization problem given as follows:

$$\begin{array}{ll} \underset{\mathbf{x}}{\mathsf{minimize}} & f(\mathbf{x}) \\ \mathsf{subject\ to} & \mathbf{x} \in \mathcal{X}, \end{array} \tag{4.5.1}$$

where the constraint set $\mathcal{X} \subseteq \mathbb{R}^N$ and no assumption is on the convexity of $f(\mathbf{x})$ and $\mathcal{X}$, instead of dealing with the original problem which could be difficult to tackle directly, the MM method solves a series of simple subproblems with surrogate functions that majorize the original objective function $f(\mathbf{x})$ over the set $\mathcal{X}$.

More specifically, starting from an initial feasible point $\mathbf{x}^{(0)}$, the MM method produces a sequence $\left\{\mathbf{x}^{(k)}\right\}$ according to the following update rule:

$$\mathbf{x}^{(k)} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \overline{f}\left(\mathbf{x}, \mathbf{x}^{(k-1)}\right), \tag{4.5.2}$$

where $\mathbf{x}^{(k-1)}$ is the point generated by the update rule at the $(k-1)$th iteration and $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ is called the majorizing function of $f(\mathbf{x})$ at $\mathbf{x}^{(k)}$.

As to claim convergence for the MM method, the function $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ should satisfy the following assumptions:

$$\begin{aligned} &\text{A1) } \overline{f}\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k)}\right) = f\left(\mathbf{x}^{(k)}\right), \ \forall \mathbf{x}^{(k)} \in \mathcal{X}, \\ &\text{A2) } \overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right) \geq f(\mathbf{x}), \ \forall \mathbf{x}, \mathbf{x}^{(k)} \in \mathcal{X}, \\ &\text{A3) } \overline{f}'\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k)}; \mathbf{d}\right) = f'\left(\mathbf{x}^{(k)}; \mathbf{d}\right), \ \forall \mathbf{d} \text{ s.t.} \mathbf{x}^{(k)} + \mathbf{d} \in \mathcal{X}, \end{aligned} \tag{4.5.3}$$

where $f'\left(\mathbf{x}^{(k)}; \mathbf{d}\right)$ stands for the directional derivative of $f(\mathbf{x})$ at $\mathbf{x}^{(k)}$ along the direction $\mathbf{d}$, i.e.,

$$f'\left(\mathbf{x}^{(k)}; \mathbf{d}\right) = \liminf_{t\to 0} \frac{f\left(\mathbf{x}^{(k)} + t\mathbf{d}\right) - f\left(\mathbf{x}^{(k)}\right)}{t};$$

similarly, $\overline{f}'\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k)}; \mathbf{d}\right)$ is the directional derivative for $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ at $\mathbf{x}^{(k)}$ along $\mathbf{d}$; and $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ is assumed continuous in both $\mathbf{x}$ and $\mathbf{x}^{(k)}$.[6] For convex $\mathcal{X}$, the proof of convergence to a d(irectional)-stationary point is established in [86], i.e., the limit point $\mathbf{x}^{(\infty)}$ of $\left\{\mathbf{x}^{(k)}\right\}$ satisfies

$$f'\left(\mathbf{x}^{(\infty)}; \mathbf{d}\right) \geq 0, \ \forall \mathbf{d} \text{ s.t. } \mathbf{x}^{(\infty)} + \mathbf{d} \in \mathcal{X}. \tag{4.5.4}$$

[6] Note that if $f(\mathbf{x})$ and $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ are both continuously differentiable, then A1) and A2) imply A3).

For a nonconvex set $\mathcal{X}$, to claim stationarity convergence, the A3) in (4.5.3) should be modified as

$$\mathrm{A3'})\ \overline{f}'\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k)}; \mathbf{d}\right) = f'\left(\mathbf{x}^{(k)}; \mathbf{d}\right), \forall \mathbf{d} \in \mathcal{T}_{\mathcal{X}}\left(\mathbf{x}^{(k)}\right), \tag{4.5.5}$$

where in this case $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ and $f(\mathbf{x})$ are defined on the whole space $\mathbb{R}^N$ and $\mathcal{T}_{\mathcal{X}}\left(\mathbf{x}^{(k)}\right)$ means the Bouligand tangent cone of $\mathcal{X}$ at $\mathbf{x}^{(k)} \in \mathcal{X}$. Then, the limit point $\mathbf{x}^{(\infty)}$ of $\left\{\mathbf{x}^{(k)}\right\}$ can be proved to be a B(ouligand)-stationary point satisfying

$$f'\left(\mathbf{x}^{(\infty)}; \mathbf{d}\right) \geq 0, \ \forall \mathbf{d} \in \mathcal{T}_{\mathcal{X}}\left(\mathbf{x}^{(\infty)}\right), \tag{4.5.6}$$

where the expression $\mathbf{d} \in \mathcal{T}_{\mathcal{X}}\left(\mathbf{x}^{(\infty)}\right)$ means there exist a sequence of points $\left\{\mathbf{x}^{(k)}\right\} \in \mathcal{X}$ converging to $\mathbf{x}^{(\infty)}$ and a sequence of positive scalars $\left\{\tau^{(k)}\right\}$ converging to $0$ such that $\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{x}^{(k)}-\mathbf{x}^{(\infty)}}{\tau^{(k)}}$. For more details of B-stationarity, please refer to [87], [88].

Although the definition for the majorizing functions $\overline{f}\left(\mathbf{x}, \mathbf{x}^{(k)}\right)$ gives us a great deal of choosing flexibility, they must be properly chosen so as to make the iterative update in (4.5.2) easy to compute while maintaining a fast convergence over the iterations. In the following, we are going to solve the MRP design problem based on the MM method.

### 4.5.2 IRGEVP and IRGTRS: Solving Algorithms for MRP Design Using $\mathrm{por}\,(p, \mathbf{w})$ and $\mathrm{pcro}\,(p, \mathbf{w})$

From (4.3.10), the MRP deign problems using $\mathrm{por}\,(p, \mathbf{w})$ and $\mathrm{pcro}\,(p, \mathbf{w})$ can be written as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\mathsf{minimize}} \quad & f(\mathbf{w}) = \xi \mathbf{w}^T\mathbf{M}_1\mathbf{w} + \zeta\left(\mathbf{w}^T\mathbf{M}_1\mathbf{w}\right)^2 \\ & + \eta \sum_{i=2}^{p}\left(\mathbf{w}^T\mathbf{M}_i\mathbf{w}\right)^2 \\ \mathsf{subject\ to} \quad & \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu \\ & \mathbf{w} \in \mathcal{W}_i,\ (i = 0, 1). \end{aligned} \tag{4.5.7}$$

Problem (4.5.7) is nonconvex with nonconvex quartic objective function, nonconvex quadratic equality constraint and convex linear constraint. In order to solve this problem via MM method, the key step is to find a majorizing function of the objective such that the majorized subproblem is easy to solve.

To compute a majorizing function, the following mathematical manipulations are necessary. We first get the Cholesky decomposition of $\mathbf{M}_0$ which is given as $\mathbf{M}_0 = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements. We further define $\bar{\mathbf{w}} = \mathbf{L}^T\mathbf{w}$, $\bar{\mathbf{M}}_i = \mathbf{L}^{-1}\mathbf{M}_i\mathbf{L}^{-T}$, $\bar{\mathbf{W}} = \bar{\mathbf{w}}\bar{\mathbf{w}}^T$, and the set $\mathcal{W}$ is mapped to $\bar{\mathcal{W}}$ under the linear transformation $\mathbf{L}$. Then using $\bar{\mathbf{w}}^T\mathbf{A}\bar{\mathbf{w}} = \text{Tr}\left(\mathbf{A}\bar{\mathbf{W}}\right)$, problem (4.5.7) can be rewritten as

$$\begin{aligned}
\underset{\bar{\mathbf{w}},\bar{\mathbf{W}}}{\text{minimize}} \quad & \xi\text{Tr}\left(\bar{\mathbf{M}}_1\bar{\mathbf{W}}\right) + \zeta\Big(\text{Tr}\left(\bar{\mathbf{M}}_1\bar{\mathbf{W}}\right)\Big)^2 \\
& + \eta\sum_{i=2}^{p}\Big(\text{Tr}\left(\bar{\mathbf{M}}_i\bar{\mathbf{W}}\right)\Big)^2 \\
\text{subject to} \quad & \bar{\mathbf{W}} = \bar{\mathbf{w}}\bar{\mathbf{w}}^T \\
& \bar{\mathbf{w}}^T\bar{\mathbf{w}} = \nu \\
& \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i, \ (i = 0, 1) \,.
\end{aligned} \tag{4.5.8}$$

Since $\text{Tr}\left(\bar{\mathbf{M}}_i\bar{\mathbf{W}}\right) = \text{vec}\left(\bar{\mathbf{M}}_i\right)^T \text{vec}\left(\bar{\mathbf{W}}\right)$ ($\mathbf{M}_i$'s are assumed symmetric), problem (4.5.8) can be reformulated as follows:

$$\begin{aligned}
\underset{\bar{\mathbf{w}},\bar{\mathbf{W}}}{\text{minimize}} \quad & \xi\text{vec}\left(\bar{\mathbf{M}}_1\right)^T \text{vec}\left(\bar{\mathbf{W}}\right) + \text{vec}\left(\bar{\mathbf{W}}\right)^T \bar{\mathbf{M}}\text{vec}\left(\bar{\mathbf{W}}\right) \\
\text{subject to} \quad & \bar{\mathbf{W}} = \bar{\mathbf{w}}\bar{\mathbf{w}}^T \\
& \bar{\mathbf{w}}^T\bar{\mathbf{w}} = \nu \\
& \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i, \ (i = 0, 1) \,,
\end{aligned} \tag{4.5.9}$$

where in the objective function

$$\bar{\mathbf{M}} \triangleq \zeta\text{vec}\left(\bar{\mathbf{M}}_1\right)\text{vec}\left(\bar{\mathbf{M}}_1\right)^T + \eta\sum_{i=2}^{p}\text{vec}\left(\bar{\mathbf{M}}_i\right)\text{vec}\left(\bar{\mathbf{M}}_i\right)^T. \tag{4.5.10}$$

Specifically, we have the expressions for portmanteau statistics $\text{por}\,(p, \mathbf{w})$ (i.e., $\zeta = 1$ and

$\eta = 1$) and penalized crossing statistics $\mathrm{pcro}\,(p, \mathbf{w})$ (i.e., $\zeta = 0$ and $\eta > 0$) as follows:[7]

$$\bar{\mathbf{M}} = \begin{cases} \sum_{i=1}^{p} (\mathbf{L} \otimes \mathbf{L})^{-1} \operatorname{vec}(\mathbf{M}_i) \operatorname{vec}(\mathbf{M}_i)^T (\mathbf{L} \otimes \mathbf{L})^{-T}, \\ \qquad\qquad\qquad\qquad \text{for } \mathrm{por}\,(p, \mathbf{w}); \\ \eta \sum_{i=2}^{p} (\mathbf{L} \otimes \mathbf{L})^{-1} \operatorname{vec}(\mathbf{M}_i) \operatorname{vec}(\mathbf{M}_i)^T (\mathbf{L} \otimes \mathbf{L})^{-T}, \\ \qquad\qquad\qquad\qquad \text{for } \mathrm{pcro}\,(p, \mathbf{w}). \end{cases}$$

The objective function in problem (4.5.9) becomes quadratic in variable $\bar{\mathbf{W}}$; however, this problem is still hard to solve due to the rank-1 constraint $\bar{\mathbf{W}} = \bar{\mathbf{w}}\bar{\mathbf{w}}^T$. We then consider applying the MM idea on this problem based on the following result. Let $\mathbf{A} \in \mathbb{S}^K$ and $\mathbf{B} \in \mathbb{S}^K$ such that $\mathbf{B} \succeq \mathbf{A}$. At any point $\mathbf{x}_0 \in \mathbb{R}^K$, the quadratic function $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is majorized by $\mathbf{x}^T\mathbf{B}\mathbf{x} + 2\mathbf{x}_0^T (\mathbf{A} - \mathbf{B})\,\mathbf{x} + \mathbf{x}_0^T (\mathbf{B} - \mathbf{A})\,\mathbf{x}_0$. It follows from $(\mathbf{x} - \mathbf{x}_0)^T (\mathbf{B} - \mathbf{A})(\mathbf{x} - \mathbf{x}_0) \geq 0$, when $\mathbf{B} \succeq \mathbf{A}$ for any $\mathbf{x}_0$. According to Lemma 4.5.2, at the $(k+1)$th iteration with point $\bar{\mathbf{W}}^{(k)}$, the second term (quadratic in $\bar{\mathbf{W}}$) in the objective function of problem (4.5.9) can be majorized by the following function:

$$\begin{aligned} u_1\left(\bar{\mathbf{W}}, \bar{\mathbf{W}}^{(k)}\right) &= \psi\big(\bar{\mathbf{M}}\big)\mathrm{vec}\Big(\bar{\mathbf{W}}\Big)^T \mathrm{vec}\Big(\bar{\mathbf{W}}\Big) \\ &+ 2\mathrm{vec}\Big(\bar{\mathbf{W}}^{(k)}\Big)^T \Big(\bar{\mathbf{M}} - \psi\big(\bar{\mathbf{M}}\big)\mathbf{I}\Big)\mathrm{vec}\Big(\bar{\mathbf{W}}\Big) \\ &+ \mathrm{vec}\Big(\bar{\mathbf{W}}^{(k)}\Big)^T \Big(\psi\big(\bar{\mathbf{M}}\big)\mathbf{I} - \bar{\mathbf{M}}\Big)\mathrm{vec}\Big(\bar{\mathbf{W}}^{(k)}\Big), \end{aligned} \tag{4.5.11}$$

where $\psi\big(\bar{\mathbf{M}}\big)$ only depends on matrix $\bar{\mathbf{M}}$ and satisfies $\psi\big(\bar{\mathbf{M}}\big)\mathbf{I} \succeq \bar{\mathbf{M}}$. On the choice of $\psi\big(\bar{\mathbf{M}}\big)$ in (4.5.11), according to Lemma 4.5.2, it is obvious that $\psi\big(\bar{\mathbf{M}}\big)$ can be chosen as the spectral norm of $\bar{\mathbf{M}}$, i.e., $\left\|\bar{\mathbf{M}}\right\|_2 = \lambda_{\max}\left(\bar{\mathbf{M}}\right)$. In the implementation of the algorithm, although $\lambda_{\max}\left(\bar{\mathbf{M}}\right)$ only needs to be computed once for the whole algorithm, it still may not be computationally easy to get. Then since $\left\|\bar{\mathbf{M}}\right\|_F \geq \left\|\bar{\mathbf{M}}\right\|_2$, we can choose $\psi\big(\bar{\mathbf{M}}\big) = \left\|\bar{\mathbf{M}}\right\|_F$, which is easier for computation.

In the majorizing function, the first term and the last term are just two constants irrelevant of the optimization variable $\bar{\mathbf{W}}$, since the first term $\mathrm{vec}\left(\bar{\mathbf{W}}\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) = \left(\bar{\mathbf{w}}^T\bar{\mathbf{w}}\right)^2 = \nu^2$, and the last term only depends on $\bar{\mathbf{W}}^{(k)}$. After replacing the second term by its majorizing function (4.5.11) in problem 4.5.9 and ignoring the constants, the majorized problem is given

[7] It follows from $\mathrm{vec}\,(\mathbf{ABC}) = \left(\mathbf{C}^T \otimes \mathbf{A}\right) \mathrm{vec}\,(\mathbf{B})$ and $\mathbf{A}^{-1} \otimes \mathbf{B}^{-1} = (\mathbf{A} \otimes \mathbf{B})^{-1}$.

by

$$\begin{aligned} \underset{\bar{\mathbf{w}},\bar{\mathbf{W}}}{\mathsf{minimize}} \quad & \xi \mathrm{vec}\left(\bar{\mathbf{M}}_1\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) \\ & + 2\mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)^T \left(\bar{\mathbf{M}} - \psi(\bar{\mathbf{M}})\mathbf{I}\right)\mathrm{vec}\left(\bar{\mathbf{W}}\right) \\ \mathsf{subject\ to} \quad & \bar{\mathbf{W}} = \bar{\mathbf{w}}\bar{\mathbf{w}}^T \\ & \bar{\mathbf{w}}^T\bar{\mathbf{w}} = \nu \\ & \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i,\ (i = 0, 1)\,, \end{aligned} \tag{4.5.12}$$

where the objective function becomes linear in the variable $\bar{\mathbf{W}}$ rather than quadratic as in (4.5.8) and (4.5.9). Further, by changing variable $\bar{\mathbf{W}}$ back to $\mathbf{w}$, we can get the overall majorizing function for (4.5.7) and the majorized subproblem in $\mathbf{w}$ which is given in the following lemma.

**Lemma 10.** *The final majorizing function of $f(\mathbf{w})$ in (4.5.7) is*

$$\begin{aligned} \overline{f}_1\left(\mathbf{w}, \mathbf{w}^{(k)}\right) &= \mathbf{w}^T\mathbf{H}^{(k)}\mathbf{w} + 2\psi(\bar{\mathbf{M}})\nu^2 \\ &- \zeta\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_1\mathbf{w}^{(k)}\right)^2 - \eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)^2, \end{aligned} \tag{4.5.13}$$

*where $\mathbf{H}^{(k)}$ is defined as follows:*

$$\begin{aligned} \mathbf{H}^{(k)} \triangleq\ & \xi\mathbf{M}_1 + 2\zeta\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_1\mathbf{w}^{(k)}\right)\mathbf{M}_1 \\ & + 2\eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)\mathbf{M}_i \\ & - 2\psi(\bar{\mathbf{M}})\mathbf{M}_0\mathbf{w}^{(k)}\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_0. \end{aligned} \tag{4.5.14}$$

*More specifically, for portmanteau statistics* $\mathrm{por}(p, \mathbf{w})$ *(i.e., $\xi = 0$, $\zeta = 1$ and $\eta = 1$) and penalized crossing statistics* $\mathrm{pcro}(p, \mathbf{w})$ *(i.e., $\xi = 1$, $\zeta = 0$ and $\eta > 0$), we have*

$$\mathbf{H}^{(k)} = \begin{cases} 2\sum_{i=1}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)\mathbf{M}_i & \\ -2\psi(\bar{\mathbf{M}})\mathbf{M}_0\mathbf{w}^{(k)}\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_0, & \textit{for}\ \mathrm{por}(p, \mathbf{w})\,; \\ \mathbf{M}_1 + 2\eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)\mathbf{M}_i & \\ -2\psi(\bar{\mathbf{M}})\mathbf{M}_0\mathbf{w}^{(k)}\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_0, & \textit{for}\ \mathrm{pcro}(p, \mathbf{w})\,. \end{cases}$$

**Algorithm 4.3** IRGEVP and IRGTRS - Algorithms for MRP design problems using $\mathrm{por}\,(p, \mathbf{w})$ and $\mathrm{pcro}\,(p, \mathbf{w})$.

```
Require: $p$, $\mathbf{M}_i$ with $i = 1, \ldots, p$, and $\nu > 0$.
Set $k = 0$ and $\mathbf{w}^{(0)} \in \mathcal{W}$;
Compute $\bar{\mathbf{M}}$ in (4.5.10) and $\psi(\bar{\mathbf{M}})$;
repeat
    Compute $\mathbf{H}^{(k)}$ in (4.5.14);
    Update $\mathbf{w}^{(k+1)}$ by solving
        1) the GEVP in (4.4.3) for $\mathbf{w} \in \mathcal{W}_0$; or
        2) the GTRS in (4.4.5) for $\mathbf{w} \in \mathcal{W}_1$;
    $k = k + 1$;
until convergence
```

*Thus, the majorized problem for problem* (4.5.7) *is given by*

$$\begin{array}{ll} \underset{\mathbf{w}}{\mathsf{minimize}} & \mathbf{w}^T \mathbf{H}^{(k)} \mathbf{w} \\ \mathsf{subject\ to} & \mathbf{w}^T \mathbf{M}_0 \mathbf{w} = \nu \\ & \mathbf{w} \in \mathcal{W}_i,\ (i = 0, 1)\,. \end{array} \tag{4.5.15}$$

**Proof 2.** *See Appendix 4.10.*

Lemma 10 shows that the objective function in the majorized problem (4.5.15) is a quadratic upperbound of that in the original problem (4.5.7). Depending on the specific form of $\mathcal{W}$, subproblem (4.5.15) can be efficiently solved for a global optimal solution by using an GEVP or an GTRS problem discussed in Section 4.4.

Finally, in order to handle the original nonconvex problem (4.5.7) directly, we just need to iteratively solve a sequence of QCQPs (i.e., GEVPs or GTRSs). We name these MM-based algorithms iteratively reweighted GEVP (IRGEVP) and iteratively reweighted GTRS (IRGTRS), respectively, which are summarized in Algorithm 4.3.

### 4.5.3 E-IRGEVP and E-IRGTRS: Solving Algorithms for MRP Design Using $\mathrm{por}\,(p, \mathbf{w})$ and $\mathrm{pcro}\,(p, \mathbf{w})$

In Section 4.5.2, based on algorithms IRGEVP or IRGTRS, the MRP design problems can be efficiently resolved by solving a nonconvex QCQP at every iteration. However, instead of dealing with a QCQP, it would be much desirable if we could get a closed-form solution for the majorized problem at each iteration. In fact, this target can be attained and the whole procedure is discussed in the following.

Instead of introducing the algorithm derivation from the original problem (4.5.7), for simplicity we start from the majorized problem (4.5.15) in Section 4.5.2. Problem (4.5.15) is equivalent to (4.10.4) which is recast as follows:

$$\begin{aligned} \underset{\bar{\mathbf{w}}}{\mathsf{minimize}} \quad & \bar{\mathbf{w}}^T \bar{\mathbf{H}}^{(k)} \bar{\mathbf{w}} \\ \mathsf{subject\ to} \quad & \bar{\mathbf{w}}^T \bar{\mathbf{w}} = \nu \\ & \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i, \ (i = 0, 1) \, . \end{aligned} \tag{4.5.16}$$

Based on Lemma 4.5.2, at the $(k+1)$th iteration with iterate $\bar{\mathbf{w}}^{(k)}$, the objective function in (4.5.16) (quadratic in $\bar{\mathbf{w}}$) can be majorized by the following majorizing function:

$$\begin{aligned} u_2\left(\bar{\mathbf{w}}, \bar{\mathbf{w}}^{(k)}\right) = {} & \psi\big(\bar{\mathbf{H}}^{(k)}\big)\left(\bar{\mathbf{w}}^T \bar{\mathbf{w}}\right) \\ & + 2\big(\bar{\mathbf{w}}^{(k)}\big)^T \left(\bar{\mathbf{H}}^{(k)} - \psi\big(\bar{\mathbf{H}}^{(k)}\big)\mathbf{I}\right) \bar{\mathbf{w}} \\ & + \left(\bar{\mathbf{w}}^{(k)}\right)^T \left(\psi\big(\bar{\mathbf{H}}^{(k)}\big)\mathbf{I} - \bar{\mathbf{H}}^{(k)}\right) \bar{\mathbf{w}}^{(k)}, \end{aligned} \tag{4.5.17}$$

where $\psi\big(\bar{\mathbf{H}}^{(k)}\big)$ can be chosen as $\left\|\bar{\mathbf{H}}^{(k)}\right\|_F$, and the first and the last terms are constants. Dropping the constants in (4.5.17), the majorized problem for (4.5.16) is given as follows:

$$\begin{aligned} \underset{\bar{\mathbf{w}}}{\mathsf{minimize}} \quad & \left(\bar{\mathbf{w}}^{(k)}\right)^T \left(\bar{\mathbf{H}}^{(k)} - \psi\big(\bar{\mathbf{H}}^{(k)}\big)\mathbf{I}\right) \bar{\mathbf{w}} \\ \mathsf{subject\ to} \quad & \bar{\mathbf{w}}^T \bar{\mathbf{w}} = \nu \\ & \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i, \ (i = 0, 1) \, . \end{aligned} \tag{4.5.18}$$

By changing variable $\bar{\mathbf{w}}$ back to $\mathbf{w}$, we can get the overall majorizing function and the majorized subproblem given in the following lemma.

**Lemma 11.** *The two majorization steps in* (4.5.11) *and* (4.5.17) *can be shown as one overall majorization at point* $\mathbf{w}^{(k)}$ *for problem* (4.5.7) *with the majorizing function given as follows:*

$$\begin{aligned} \overline{f}_2\left(\mathbf{w}, \mathbf{w}^{(k)}\right) = {} & 2\big(\mathbf{e}^{(k)}\big)^T \mathbf{w} - \left(\mathbf{w}^{(k)}\right)^T \mathbf{H}^{(k)} \mathbf{w}^{(k)} \\ & + 2\psi\big(\bar{\mathbf{H}}^{(k)}\big)\nu - \zeta\left(\big(\mathbf{w}^{(k)}\big)^T \mathbf{M}_1 \mathbf{w}^{(k)}\right)^2 \\ & - \eta \sum_{i=2}^{p} \left(\big(\mathbf{w}^{(k)}\big)^T \mathbf{M}_i \mathbf{w}^{(k)}\right)^2 + 2\psi\big(\bar{\mathbf{M}}\big)\nu^2, \end{aligned} \tag{4.5.19}$$

*where*

$$\mathbf{e}^{(k)} \triangleq \left(\mathbf{H}^{(k)} - \psi\left(\bar{\mathbf{H}}^{(k)}\right)\mathbf{M}_0\right)\mathbf{w}^{(k)}. \tag{4.5.20}$$

*Thus, the final majorized problem for problem* (4.5.7) *becomes*

$$\begin{array}{ll} \underset{\mathbf{w}}{\mathsf{minimize}} & \left(\mathbf{e}^{(k)}\right)^T\mathbf{w} \\ \mathsf{subject\ to} & \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu \\ & \mathbf{w} \in \mathcal{W}_i,\ (i = 0, 1)\,. \end{array} \tag{4.5.21}$$

**Proof 3.** *See Appendix 4.11.*

Lemma 11 shows after using the MM trick twice, the objective function in problem (4.5.21) becomes a linear upperbound in variable $\mathbf{w}$ of the original problem (4.5.7). By the trick used to get problems (4.4.3) and (4.4.5), we can eliminate the linear constraint in (4.5.21). Then, it becomes a QCLP which has a closed-form solution rather than the QCQP derived from (4.5.15). Based on Lagrange duality, problem (4.5.21) has a closed-form solution. Specifically, for $\mathbf{w} \in \mathcal{W}_0$,

$$\begin{aligned} &\mathbf{w}^\star = \\ &-\left(\frac{\nu}{\left(\mathbf{e}^{(k)}\right)^T\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)}}\right)^{\frac{1}{2}}\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)}, \end{aligned}$$

and for $\mathbf{w} \in \mathcal{W}_1$,

$$\begin{aligned} &\mathbf{w}^\star = \\ &-\left(\frac{\nu - \mathbf{w}_0^T\mathbf{M}_0\mathbf{w}_0 + \mathbf{w}_0^T\mathbf{M}_0\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{M}_0\mathbf{w}_0}{\left(\mathbf{e}^{(k)}\right)^T\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)}}\right)^{\frac{1}{2}} \\ &\times\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)} - \mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{M}_0\mathbf{w}_0 + \mathbf{w}_0, \end{aligned}$$

where $\mathbf{F}$ satisfies $\mathbf{1}^T\mathbf{F} = \mathbf{0}$, and $\mathbf{w}_0$ satisfies $\mathbf{1}^T\mathbf{w}_0 = 1$. See Appendix 4.12. Finally, the MRP design problem (4.5.7) is solved iteratively by a closed-form update at each iteration. Just to make a connection with IRGEVP and IRGTRS, these algorithms are named extended IRGEVP (E-IRGEVP) and extended IRGTRS (E-IRGTRS) which are summarized in Algorithm 4.4.

**Algorithm 4.4** E-IRGEVP and E-IRGTRS - Algorithms for MRP design problems using $\mathrm{por}\,(p, \mathbf{w})$ and $\mathrm{pcro}\,(p, \mathbf{w})$.

```
Require: $p$, $\mathbf{M}_i$ with $i = 1, \ldots, p$, and $\nu > 0$.
Set $k = 0$ and $\mathbf{w}^{(0)} \in \mathcal{W}$;
Compute $\bar{\mathbf{M}}$ in (4.5.10) and $\psi(\bar{\mathbf{M}})$;
repeat
    Compute $\bar{\mathbf{H}}^{(k)}$ in (4.10.3), $\psi(\bar{\mathbf{H}}^{(k)})$, and $\mathbf{e}^{(k)}$ in (4.5.20);
    Update $\mathbf{w}^{(k+1)}$ with a closed-form solution according to Lemma 4.5.3;
    $k = k + 1$;
until convergence
```

## 4.6 Complexity and Convergence Analysis

### 4.6.1 Complexity Analysis

For Algorithms 4.1 and 4.2 (i.e., GEVP and GTRS) in Section 4.4, the per-iteration computational cost mainly comes from the matrix multiplication with complexity of $\mathcal{O}\left(N^3\right)$. The algorithm converges to the global optimal solution of the original problem (4.4.1) or (4.4.4). For the MM-based Algorithms 4.3 and 4.4 (i.e., IRGEVP, IRGTRS, E-IRGEVP and E-IRGTRS) in Section 4.5, the per-iteration computational cost comes from the Cholesky decomposition or matrix multiplication, so the complexity is still of $\mathcal{O}\left(N^3\right)$.

### 4.6.2 Convergence Analysis

The algorithms IRGEVP and IRGTRS given in Algorithm 4.3 and algorithms E-IRGEVP and E-IRGTRS given in Algorithm 4.4 are all based on the general MM method, thus according to Section 4.5.1, we know that the sequence of objective values $\left\{f\left(\mathbf{w}^{(k)}\right)\right\}$ generated by these algorithms is nonincreasing. The original optimization problem (4.5.7) is a constrained minimization problem and the objective function $f$ is bounded below, thus the sequence $\left\{f\left(\mathbf{w}^{(k)}\right)\right\}$ is guaranteed to converge to a finite value. Then based on the B-stationarity defined in Section 4.5.1, we can further give the convergence property for the sequence $\left\{\mathbf{w}^{(k)}\right\}$ generated by the MM-based algorithms in the following result.

**Proposition 3.** *Every limit point, denoted by* $\mathbf{w}^{(\infty)}$*, of the sequence* $\left\{\mathbf{w}^{(k)}\right\}$ *generated by the MM-based algorithms (i.e., Algorithm 4.3 and Algorithm 4.4) is a B-stationary point of problem* (4.5.7).

**Proof 4.** *See Appendix 4.13.*

## 4.7 Numerical Experiments

A statistical arbitrage strategy involves several steps of which the MRP design is a central one. Here, we divide the whole strategy into four sequential steps, namely: assets pool construction, MRP design, unit-root test, and mean reversion trading. In the first step, we select a collection of possibly cointegrated asset candidates to construct an asset pool, on which we will not elaborate in this chapter. In the second step, based on the candidate assets from the asset pool, MRPs are designed using either traditional design methods like Engle-Granger OLS method [28] and Johansen method [29] or the proposed methods in this chapter. In the third step, unit-root test procedures like Augmented Dickey-Fuller test [89] and Phillips-Perron test [90] are applied to test the stationarity or mean reversion property of the designed MRPs. In the fourth step, MRPs passing the unit-root tests will be traded based on a designed mean reversion trading strategy.

In this section, we first illustrate several performance metrics for the portfolio investment. Then the performance of our proposed MRP design methods in Sections 4.4 and 4.5 is evaluated using both synthetic data and real market data are shown accordingly.

### 4.7.1 Performance Metrics

In this chapter, we employ the following performance metrics for the numerical experiments.

#### 4.7.1.1 Portfolio Return Measures

In Section 4.2, we have defined the multi-period P&L $\mathrm{P\&L}_t(\tau)$ and single-period P&L $\mathrm{P\&L}_t$. Since there is no trading conducted between two trading periods, the P&L measures (both the multi-period P&L and single-period P&L) are simply defined to be 0. In the following, based on the P&L definition, we give the following useful portfolio return measures.

**Cumulative P&L** In order to measure the cumulative return performance for an MRP, we define the cumulative P&L (not compounding) in one trading from time $t_1$ to $t_2$ as

$$\mathrm{Cum.\ P\&L}\,(t_1, t_2) = \sum_{t=t_1}^{t_2} \mathrm{P\&L}_t. \tag{4.7.1}$$

**Return On Investment (ROI)** Since different MRPs may have different leverage properties due to $\mathbf{w}_p$, we introduce another portfolio return measure (rate of return) called return on investment (ROI). Within one trading period, the ROI at time $t$ ($t_o \leq t \leq t_c$) is defined to be the single-period P&L at time $t$ normalized by the gross investment deployed which is $\|\mathbf{w}_p\|_1$ (that is the gross investment exposure to the market including the long position investment and the short position investment) written as

$$\mathrm{ROI}_t = \frac{\mathrm{P\&L}_t}{\|\mathbf{w}_p\|_1}. \tag{4.7.2}$$

Like the P&L measures, between two trading periods, $\mathrm{ROI}_t$ is defined to be $0$.

#### 4.7.1.2 Sharpe Ratio (SR)

The Sharpe ratio (SR) [91] is a measure for calculating risk-adjusted return. It describes how much excess return one can receive for the extra volatility (square root of variance). The annualized Sharpe ratio of ROI (or, equivalently, Sharpe ratio of P&L) for a trading stage from time $t_1$ to $t_2$ is defined as follows:

$$\mathrm{SR}_{\mathrm{ROI}}\left(t_1, t_2\right) = \sqrt{252}\frac{\mu_{\mathrm{ROI}}}{\sigma_{\mathrm{ROI}}}, \tag{4.7.3}$$

where $\mu_{\mathrm{ROI}} = 1/\left(t_2 - t_1\right)\sum_{t=t_1}^{t_2}\mathrm{ROI}_t$, $\sigma_{\mathrm{ROI}} = \left[1/\left(t_2 - t_1\right)\sum_{t=t_1}^{t_2}\left(\mathrm{ROI}_t - \mu_{\mathrm{ROI}}\right)^2\right]^{1/2}$, $t$ denotes day, and the factor $\sqrt{252}$ relates the daily SR to the annualized SR (assuming 252 trading days per year). In the computation of the SR, we set the risk-free return to be $0$, in which case it reduces to the information ratio.

#### 4.7.1.3 Transaction Cost

The transaction or trading costs refer to brokerage commissions, stamp fees, bid-ask spreads, financing costs, and so on. In our experiments, we assume the transaction cost to be fixed as 35 basis points (BPs), i.e., 0.35%, per trade when opening or closing a trading position, then the round-trip transaction cost is 70 BPs.

### 4.7.2 Synthetic Data Experiments

For synthetic data experiments, we generate the sample path of log-prices for $M$ financial assets using a multivariate cointegrated system model [26], where there are $r$ long-run cointegration relations and $M - r$ common trends. We divide the sample path into two stages: in-sample training stage and out-of-sample backtesting or trading stage. All the parameters like spread equilibrium $\mu_z$, trading threshold $\Delta$, and portfolio weight $\mathbf{w}$ are decided in the training stage. The out-of-sample performance of our design methods are tested in the trading stage. In the synthetic experiments, we set $M = 6$ and $r = 5$ and only show the performance of the MRP design methods under net budget constraint $\mathcal{W}_1$. We estimate $N = 5$ spreads using the generated sample path by the OLS and the Johansen method. Based on these five spreads, an MRP is designed as $z_t = \mathbf{w}^T \mathbf{s}_t$. The simulated log-prices and the spreads for the trading stage are shown in Figure 4.3.

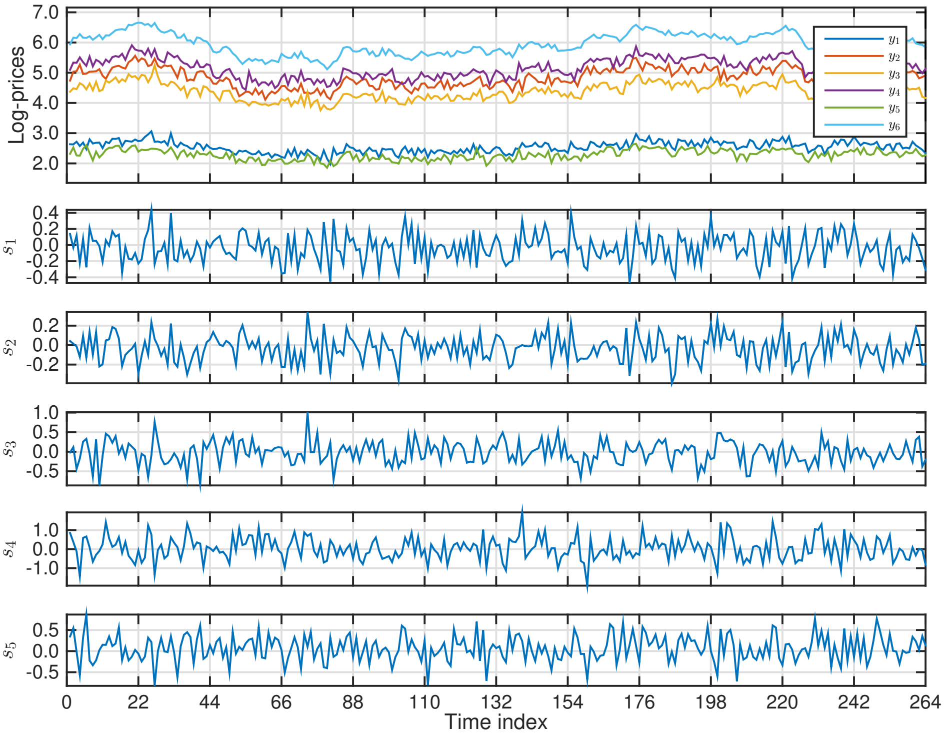


Figure 4.3: Log-prices and five estimated spreads. (The sample length for in-sample training is chosen to be $5 \times 12 \times 22$, and the sample length for out-of-sample trading is $12 \times 22$.)

In Figure 4.4, our proposed problem formulation (denoted as IRGTRS (prop.) and E-IRGTRS (prop.)) is compared to the benchmark formulation in [68] (denoted as SDR (bench.)). To ensure a fair comparison, the net investment budget (i.e., $\mathbf{1}^T \mathbf{w}$) and the variance of the spread (i.e., $\mathbf{w}^T \mathbf{M}_0 \mathbf{w}$) are set to be the same for all the methods. From the simulation results,

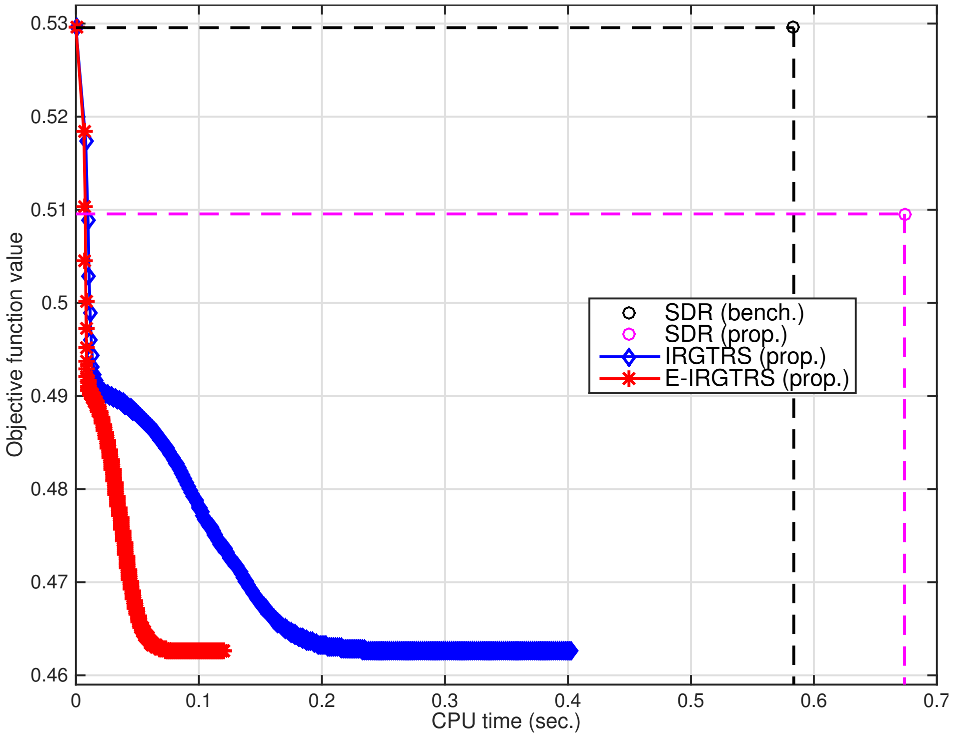


Figure 4.4: Numerical convergence of objective function value for pcro $(5, \mathbf{w})$.

the proposed MRP design problem formulation can attain a lower objective function value. The proposed problem formulation is also solved using the SDR method (denoted as SDR (prop.)) with comparison to the MM-based algorithms (denoted as IRGTRS (prop.) and E-IRGTRS (prop.)) in Figure 4.4. From the convergence results, the MM-based algorithms are better than the SDR methods in terms of converging solution property and the time.

The performance of the MRPs designed using our proposed methods are compared with those of one underlying spread and the method in [68] based on pcro $(5, \mathbf{w})$ and pre $(\mathbf{w})$, which are shown in Figure 4.5 and Figure 4.6. From our simulations, we can conclude that our designed MRPs do generate consistent positive profits. And simulation results also show that our designed portfolios can outperform the underlying spreads and the MRPs designed using methods in [68] with higher Sharpe ratios of ROIs and higher cumulative P&Ls.

### 4.7.3 Market Data Experiments

We also test our methods using real market data from the Standard & Poor's 500 (S&P 500) Index, which is usually considered as one of the best representatives for the U.S. stock markets. The data are retrieved from Google Finance[8] and adjusted daily closing stock prices are

[8] https://www.google.com/finance

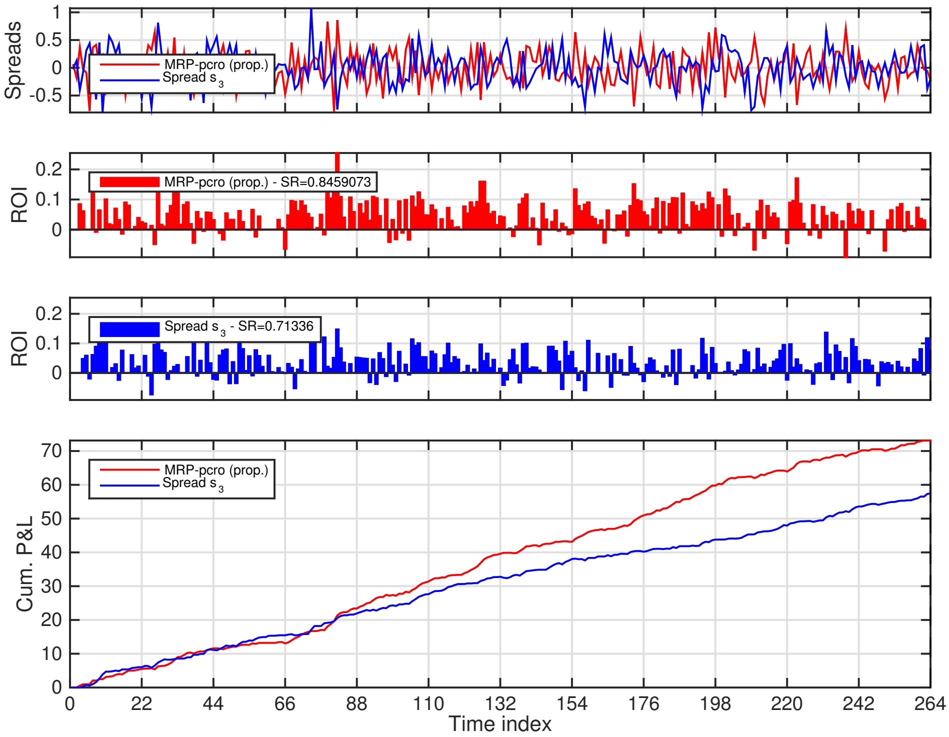


Figure 4.5: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP-pcro (prop.) with one underlying spread denoted as Spread $s_3$.

employed. We first choose stock candidates which are possibly cointegrated to form stock asset pools. One stock pool is $\{\mathsf{APA}, \mathsf{AXP}, \mathsf{CAT}, \mathsf{COF}, \mathsf{FCX}, \mathsf{IBM}, \mathsf{MMM}\}$, where the stocks are denoted by their ticker symbols. Three spreads are constructed from this pool based on the Johansen method. Then MRP design methods are employed and unit-root tests are used to test their tradability. The log-prices of the stocks and the log-prices for the three spreads are shown in Figure 4.7. Based on the mean reversion trading framework mentioned before, one trading experiment is carried out from February 1st, 2012 to June 30th, 2014.

In Figure 4.8, we compare the performance of our designed MRP with the underlying spread $s_1$. The log-prices for the designed spreads, and the out-of-sample performance like ROIs, Sharpe ratios of ROIs, and cumulative P&Ls are reported. It is shown that using our method, the designed MRP can achieve a higher Sharpe ratio and a better final cumulative return. We also compare our proposed design method with the method in [68] based on the mean reversion criterion por $(3, \mathbf{w})$ where the investment budget and the portfolio variance are set to be the same. From Figure 4.9, we can see that our proposed method can outperform the benchmark method through a mean reversion trading design with a higher Sharpe ratio and a higher final cumulative return performance.

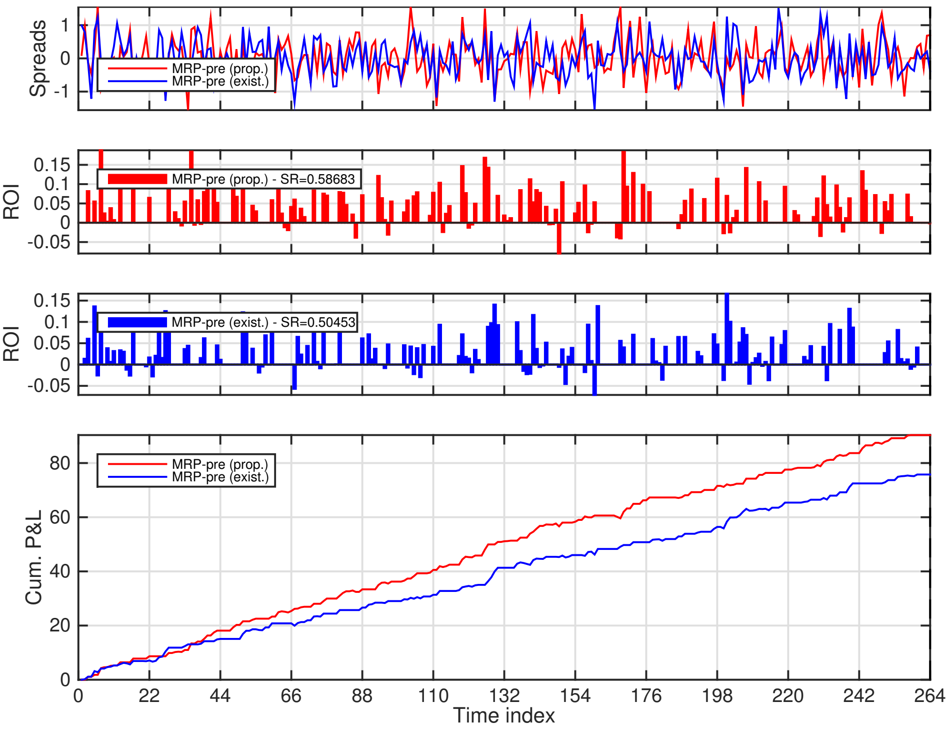


Figure 4.6: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP-pre (prop.) and one existing benchmark method in [68] denoted as MRP-pre (exist.).

## 4.8 Chapter Summary and Conclusions

The mean-reverting portfolio design problem arising from statistical arbitrage has been considered in this chapter. We have formulated the MRP design problem as the optimization of a mean reversion criterion characterizing the mean reversion strength of the portfolio and, at the same time, taking into consideration the variance of the portfolio and an investment budget constraint. Several specific optimization problems have been considered based on the general design idea and efficient algorithms have been derived for problem solving. Numerical results show that our proposed methods are able to generate consistent positive profits and outperform the the design methods in literature.

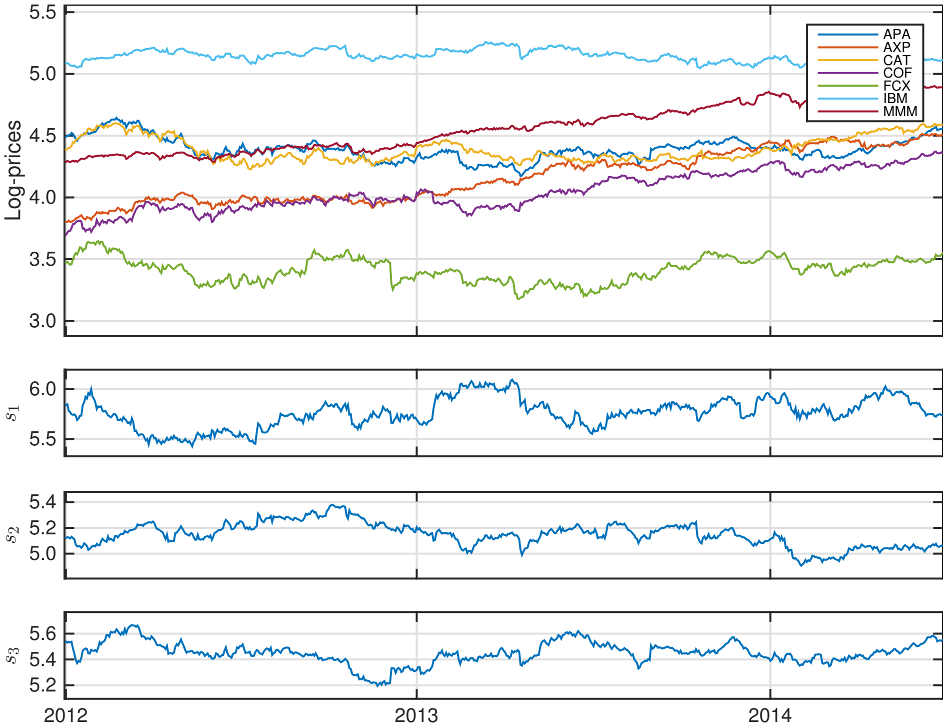


Figure 4.7: Log-prices for {APA, AXP, CAT, COF, FCX, IBM, MMM} and three spreads $s_1$, $s_2$, and $s_3$.

## 4.9 Proof for Lemma 9

Since the spread of an MRP is defined as $z_t = \mathbf{w}_p^T \mathbf{y}_t$, then the multi-period P&L at time $t$ for $\tau$ holding periods for a long position on the MRP is given by

$$\begin{aligned}
&\text{P\&L}_t(\tau) \\
=&\mathbf{w}_p^T \mathbf{r}_t (t - t_o) - \mathbf{w}_p^T \mathbf{r}_{t-\tau} (t - \tau - t_o) \\
=&\sum_{m=1}^{M} \left( w_{p,m} r_{m,t} (t - t_o) - w_{p,m} r_{m,t-\tau} (t - \tau - t_o) \right) \\
=&\sum_{m=1}^{M} \left( w_{p,m} \left( \frac{p_{m,t}}{p_{m,t_o}} - 1 \right) - w_{p,m} \left( \frac{p_{m,t-\tau}}{p_{m,t_o}} - 1 \right) \right) \\
\approx&\sum_{m=1}^{M} w_{p,m} \left[ \log(p_{m,t}) - \log(p_{m,t_o}) \right] \\
&\qquad - \sum_{m=1}^{M} w_{p,m} \left[ \log(p_{m,t-\tau}) - \log(p_{m,t_o}) \right] \\
=&\sum_{m=1}^{M} w_{p,m} y_{m,t} - \sum_{m=1}^{M} w_{p,m} y_{m,t-\tau} \\
=&z_t - z_{t-\tau}.
\end{aligned}$$

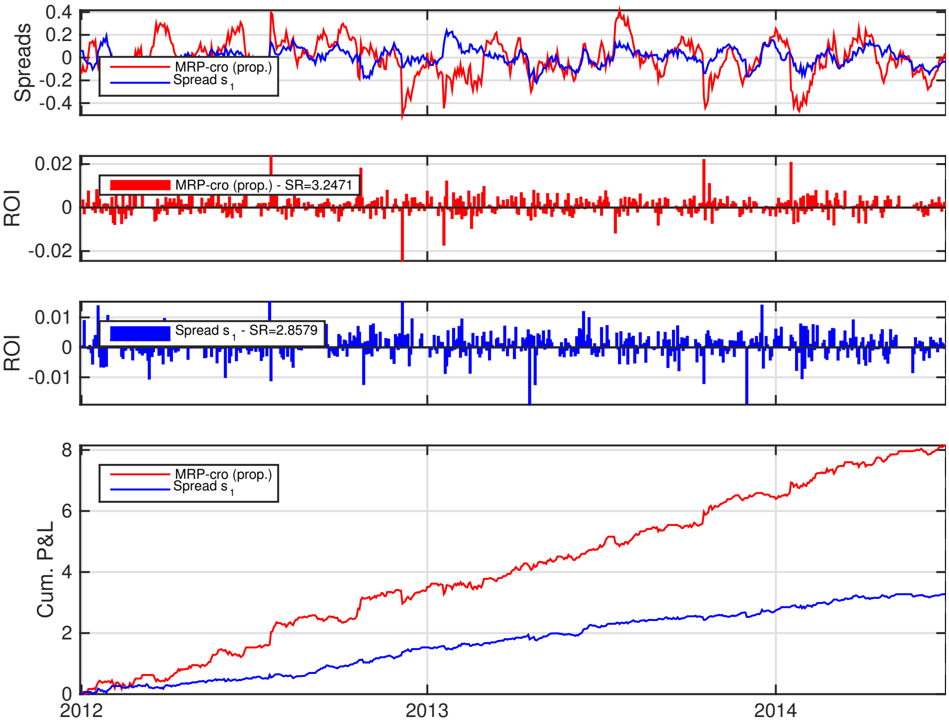


Figure 4.8: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP-cro (prop.) with one underlying spread denoted as Spread $s_1$.

Similarly, for a short position on the MRP, the $\mathrm{P\&L}_t(\tau)$ is computed as $z_{t-\tau} - z_t$.

## 4.10 Proof for Lemma 10

It is easy to see that, based on Lemma 4.5.2, only the second term of $f(\mathbf{w})$ in problem (4.5.7) is majorized. Then the overall majorizing function for $f(\mathbf{w})$ at $\mathbf{w}^{(k)}$ can be attained through replacing the second term by its majorizing function.

Replacing the the second term in the objective function of problem (4.5.9) by $u_1\left(\bar{\mathbf{W}}, \bar{\mathbf{W}}^{(k)}\right)$ in (4.5.11) and substituting $\bar{\mathbf{M}}$ in (4.5.10) back into the function, we get the following overall

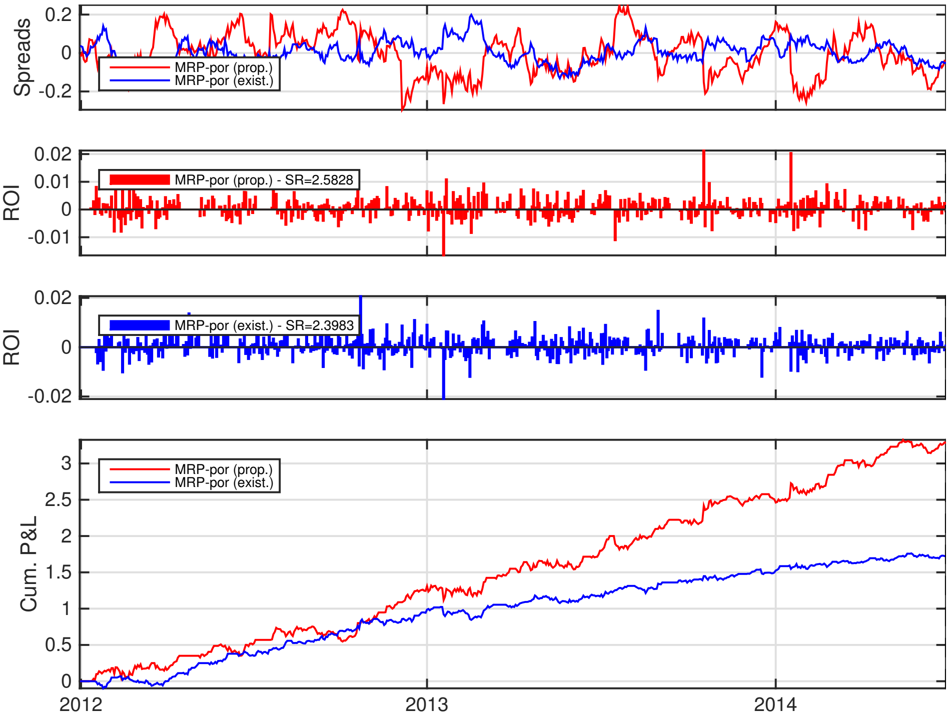


Figure 4.9: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP-por (prop.) and one existing benchmark method in [68] denoted as MRP-por (exist.).

majorizing function in variable $\bar{\mathbf{W}}$ as follows:

$$
\begin{aligned}
\bar{f}\left(\bar{\mathbf{W}}, \bar{\mathbf{W}}^{(k)}\right) &= \xi \mathrm{vec}\left(\bar{\mathbf{M}}_1\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) + u_1\left(\bar{\mathbf{W}}, \bar{\mathbf{W}}^{(k)}\right) \\
&= \xi \mathrm{vec}\left(\bar{\mathbf{M}}_1\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) + \psi(\bar{\mathbf{M}}) \mathrm{vec}\left(\bar{\mathbf{W}}\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) \\
&+ 2\zeta \left[\mathrm{vec}\left(\bar{\mathbf{M}}_1\right) \mathrm{vec}\left(\bar{\mathbf{M}}_1\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)\right]^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) \\
&+ 2\eta \left[\sum_{i=2}^{p} \mathrm{vec}\left(\bar{\mathbf{M}}_i\right) \mathrm{vec}\left(\bar{\mathbf{M}}_i\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)\right]^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) \\
&- 2\psi(\bar{\mathbf{M}}) \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) \\
&+ \psi(\bar{\mathbf{M}}) \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right) \\
&- \zeta \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)^T \mathrm{vec}\left(\bar{\mathbf{M}}_1\right) \mathrm{vec}\left(\bar{\mathbf{M}}_1\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right) \\
&- \eta \sum_{i=2}^{p} \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right)^T \mathrm{vec}\left(\bar{\mathbf{M}}_i\right) \mathrm{vec}\left(\bar{\mathbf{M}}_i\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}^{(k)}\right).
\end{aligned}
\tag{4.10.1}
$$

Then, by undoing the matrix lifting, i.e., changing variable $\bar{\mathbf{W}}$ back to $\bar{\mathbf{w}}$, and using $\mathrm{vec}\left(\bar{\mathbf{M}}_i\right)^T \mathrm{vec}\left(\bar{\mathbf{W}}\right) = \mathrm{Tr}\left(\bar{\mathbf{M}}_i \bar{\mathbf{W}}\right) = \bar{\mathbf{w}}^T \bar{\mathbf{M}}_i \bar{\mathbf{w}}$, we can get the majorizing function in $\bar{\mathbf{w}}$

given by

$$
\begin{aligned}
&\overline{f}_1\Big(\bar{\mathbf{w}},\bar{\mathbf{w}}^{(k)}\Big) = \xi\mathrm{Tr}\Big(\bar{\mathbf{M}}_1\bar{\mathbf{w}}\bar{\mathbf{w}}^T\Big) + \psi(\bar{\mathbf{M}})\mathrm{Tr}\Big(\bar{\mathbf{w}}\bar{\mathbf{w}}^T\bar{\mathbf{w}}\bar{\mathbf{w}}^T\Big)\\
&+ 2\zeta\mathrm{Tr}\Big(\bar{\mathbf{M}}_1\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\Big)\mathrm{Tr}\Big(\bar{\mathbf{M}}_1\bar{\mathbf{w}}\bar{\mathbf{w}}^T\Big)\\
&+ 2\eta\sum_{i=2}^{p}\Big[\mathrm{Tr}\Big(\bar{\mathbf{M}}_i\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\Big)\mathrm{Tr}\Big(\bar{\mathbf{M}}_i\bar{\mathbf{w}}\bar{\mathbf{w}}^T\Big)\Big]\\
&- 2\psi(\bar{\mathbf{M}})\mathrm{Tr}\Big(\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}\bar{\mathbf{w}}^T\Big)\\
&+ \psi(\bar{\mathbf{M}})\mathrm{Tr}\Big(\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\Big)\\
&- \zeta\mathrm{Tr}\Big(\bar{\mathbf{M}}_1\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\Big)^2 - \eta\sum_{i=2}^{p}\mathrm{Tr}\Big(\bar{\mathbf{M}}_i\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T\Big)^2\\
&= \bar{\mathbf{w}}^T\bar{\mathbf{H}}^{(k)}\bar{\mathbf{w}} + \psi(\bar{\mathbf{M}})\Big(\bar{\mathbf{w}}^T\bar{\mathbf{w}}\Big)^2 + \psi(\bar{\mathbf{M}})\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}^{(k)}\Big)^2\\
&- \zeta\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_1\bar{\mathbf{w}}^{(k)}\Big)^2 - \eta\sum_{i=2}^{p}\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_i\bar{\mathbf{w}}^{(k)}\Big)^2,
\end{aligned}
\tag{4.10.2}
$$

where in the objective function, $\bar{\mathbf{H}}^{(k)}$ is defined as follows:

$$
\begin{aligned}
\bar{\mathbf{H}}^{(k)} \triangleq\ & \xi\bar{\mathbf{M}}_1 + 2\zeta\left(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_1\bar{\mathbf{w}}^{(k)}\right)\bar{\mathbf{M}}_1\\
&+ 2\eta\sum_{i=2}^{p}\left(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_i\bar{\mathbf{w}}^{(k)}\right)\bar{\mathbf{M}}_i\\
&- 2\psi(\bar{\mathbf{M}})\bar{\mathbf{w}}^{(k)}\big(\bar{\mathbf{w}}^{(k)}\big)^T.
\end{aligned}
\tag{4.10.3}
$$

Dropping the constants in $\overline{f}_1\left(\bar{\mathbf{w}},\bar{\mathbf{w}}^{(k)}\right)$, problem (4.5.12) becomes

$$
\begin{aligned}
\underset{\bar{\mathbf{w}}}{\mathsf{minimize}}\quad & \bar{\mathbf{w}}^T\bar{\mathbf{H}}^{(k)}\bar{\mathbf{w}}\\
\mathsf{subject\ to}\quad & \bar{\mathbf{w}}^T\bar{\mathbf{w}} = \nu\\
& \bar{\mathbf{w}} \in \bar{\mathcal{W}}_i,\ (i = 0, 1).
\end{aligned}
\tag{4.10.4}
$$

From $\overline{f}_1\left(\bar{\mathbf{w}},\bar{\mathbf{w}}^{(k)}\right)$, by changing variable $\bar{\mathbf{w}}$ back to $\mathbf{w}$ based on $\bar{\mathbf{w}} = \mathbf{L}^T\mathbf{w}$ and considering the constraint $\mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu$, we have the majoring function in variable $\mathbf{w}$ given as follows:

$$\begin{aligned}\overline{f}_1\left(\mathbf{w},\mathbf{w}^{(k)}\right) &= \mathbf{w}^T\mathbf{H}^{(k)}\mathbf{w} + \psi(\bar{\mathbf{M}})\left(\mathbf{w}^T\mathbf{M}_0\mathbf{w}\right)^2\\ &+ \psi(\bar{\mathbf{M}})\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_0\mathbf{w}^{(k)}\right)^2 - \zeta\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_1\mathbf{w}^{(k)}\right)^2\\ &- \eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)^2\\ &= \mathbf{w}^T\mathbf{H}^{(k)}\mathbf{w} + 2\psi(\bar{\mathbf{M}})\nu^2 - \zeta\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_1\mathbf{w}^{(k)}\right)^2\\ &- \eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)^2,\end{aligned} \tag{4.10.5}$$

where $\mathbf{H}^{(k)}$ in the objective function is given by

$$\begin{aligned}\mathbf{H}^{(k)} \triangleq\ & \xi\mathbf{M}_1 + 2\zeta\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_1\mathbf{w}^{(k)}\right)\mathbf{M}_1\\ &+ 2\eta\sum_{i=2}^{p}\left(\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_i\mathbf{w}^{(k)}\right)\mathbf{M}_i\\ &- 2\psi(\bar{\mathbf{M}})\mathbf{M}_0\mathbf{w}^{(k)}\left(\mathbf{w}^{(k)}\right)^T\mathbf{M}_0.\end{aligned}$$

Finally, based on $\overline{f}_1\left(\mathbf{w},\mathbf{w}^{(k)}\right)$, the majorized problem is given as follows:

$$\begin{array}{ll}\underset{\mathbf{w}}{\mathsf{minimize}} & \mathbf{w}^T\mathbf{H}^{(k)}\mathbf{w}\\ \mathsf{subject\ to} & \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu\\ & \mathbf{w}\in\mathcal{W}_i,\ (i=0,1)\,.\end{array} \tag{4.10.6}$$

## 4.11 Proof for Lemma 11

The proof is similar to that for Lemma 10. Since the majorization step in Lemma 11 can be regarded as a second majorization for the majorizing function given in Lemma 10 (i.e., $\overline{f}_2\left(\mathbf{w},\mathbf{w}^{(k)}\right) \geq \overline{f}_1\left(\mathbf{w},\mathbf{w}^{(k)}\right)$), we can just start the proof from the first majorizing function in (4.10.2).

First, replacing the first term of $\overline{f}_1\left(\bar{\mathbf{w}},\bar{\mathbf{w}}^{(k)}\right)$ by its majorizing function $u_2\left(\bar{\mathbf{w}},\bar{\mathbf{w}}^{(k)}\right)$ in

(4.5.17), we have

$$\begin{aligned}
\overline{f}_2\Big(\bar{\mathbf{w}}, \bar{\mathbf{w}}^{(k)}\Big) &= \psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\bar{\mathbf{w}}^T\bar{\mathbf{w}}\Big) + 2\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{H}}^{(k)}\bar{\mathbf{w}} \\
&- 2\psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}\Big) + \psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}\Big) \\
&- \big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{H}}^{(k)}\bar{\mathbf{w}}^{(k)} + \psi\big(\bar{\mathbf{M}}\big)\Big(\bar{\mathbf{w}}^T\bar{\mathbf{w}}\Big)^2 \\
&+ \psi\big(\bar{\mathbf{M}}\big)\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{w}}^{(k)}\Big)^2 - \zeta\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_1\bar{\mathbf{w}}^{(k)}\Big)^2 \\
&- \eta\sum_{i=2}^{p}\Big(\big(\bar{\mathbf{w}}^{(k)}\big)^T\bar{\mathbf{M}}_i\bar{\mathbf{w}}^{(k)}\Big)^2.
\end{aligned} \tag{4.11.1}$$

Then, we change the variable $\bar{\mathbf{w}}$ back to $\mathbf{w}$, consider the constraint $\mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu$, and get the majoring function in variable $\mathbf{w}$ as follows:

$$\begin{aligned}
\overline{f}_2\left(\mathbf{w}, \mathbf{w}^{(k)}\right) &= \psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\mathbf{w}^T\mathbf{M}_0\mathbf{w}\Big) + 2\big(\mathbf{w}^{(k)}\big)^T\mathbf{H}^{(k)}\mathbf{w} \\
&- 2\psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_0\mathbf{w}\Big) + \psi\big(\bar{\mathbf{H}}^{(k)}\big)\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_0\mathbf{w}\Big) \\
&- \big(\mathbf{w}^{(k)}\big)^T\mathbf{H}^{(k)}\mathbf{w}^{(k)} + \psi\big(\bar{\mathbf{M}}\big)\Big(\mathbf{w}^T\mathbf{M}_0\mathbf{w}\Big)^2 \\
&+ \psi\big(\bar{\mathbf{M}}\big)\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_0\mathbf{w}^{(k)}\Big)^2 - \zeta\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_1\mathbf{w}^{(k)}\Big)^2 \\
&- \eta\sum_{i=2}^{p}\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_i\mathbf{w}^{(k)}\Big)^2 \\
&= 2\big(\mathbf{e}^{(k)}\big)^T\mathbf{w} - \big(\mathbf{w}^{(k)}\big)^T\mathbf{H}^{(k)}\mathbf{w}^{(k)} + 2\psi\big(\bar{\mathbf{H}}^{(k)}\big)\nu \\
&- \zeta\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_1\mathbf{w}^{(k)}\Big)^2 - \eta\sum_{i=2}^{p}\Big(\big(\mathbf{w}^{(k)}\big)^T\mathbf{M}_i\mathbf{w}^{(k)}\Big)^2 \\
&+ 2\psi\big(\bar{\mathbf{M}}\big)\nu^2
\end{aligned} \tag{4.11.2}$$

where

$$\mathbf{e}^{(k)} \triangleq \Big(\mathbf{H}^{(k)} - \psi\big(\bar{\mathbf{H}}^{(k)}\big)\mathbf{M}_0\Big)\mathbf{w}^{(k)}.$$

Finally, the majorized subproblem is accordingly given in the following way:

$$\begin{aligned}
\underset{\mathbf{w}}{\mathsf{minimize}} \quad & \big(\mathbf{e}^{(k)}\big)^T\mathbf{w} \\
\mathsf{subject\ to} \quad & \mathbf{w}^T\mathbf{M}_0\mathbf{w} = \nu \\
& \mathbf{w} \in \mathcal{W}_i,\ (i = 0, 1)\,.
\end{aligned} \tag{4.11.3}$$

## 4.12 Proof for Lemma 4.5.3

We show the proof for the case $\mathbf{w} \in \mathcal{W}_1$, and the other case follows accordingly. We first check the regularity conditions (or constraint qualifications). Problem (4.5.21) is equivalent to the following convex problem

$$\begin{aligned} \underset{\mathbf{w}}{\mathsf{minimize}} \quad & \left(\mathbf{e}^{(k)}\right)^T \mathbf{w} \\ \mathsf{subject\ to} \quad & \mathbf{w}^T \mathbf{M}_0 \mathbf{w} \leq \nu \\ & \mathbf{1}^T \mathbf{w} = 1, \end{aligned} \tag{4.12.1}$$

since the objective is linear and the optimal solution $\mathbf{w}^\star$ is always attained in the boundary of the quadratic constraint set. Slater's regularity condition holds for (4.12.1), i.e., it is strictly feasible. By variable changing $\mathbf{w} = \mathbf{F}\mathbf{x} + \mathbf{w}_0$, with $\mathbf{N}_0 = \mathbf{F}^T\mathbf{M}_0\mathbf{F}$, $\mathbf{p}_0 = \mathbf{F}^T\mathbf{M}_0\mathbf{w}_0$, and $b_0 = \mathbf{w}_0^T\mathbf{M}_0\mathbf{w}_0$, we have

$$\begin{aligned} \underset{\mathbf{w}}{\mathsf{minimize}} \quad & \left(\mathbf{e}^{(k)}\right)^T \mathbf{F}\mathbf{x} \\ \mathsf{subject\ to} \quad & \mathbf{x}^T \mathbf{N}_0 \mathbf{x} + 2\mathbf{p}_0^T \mathbf{x} + b_0 \leq \nu. \end{aligned} \tag{4.12.2}$$

The Karush-Kuhn-Tucker (KKT) conditions for primal and dual variable pair $(\mathbf{x}^\star, \lambda^\star)$ can be written as

$$\begin{cases} 2\lambda^\star \mathbf{N}_0 \mathbf{x}^\star + 2\lambda^\star \mathbf{p}_0 + \mathbf{F}^T \mathbf{e}^{(k)} = 0, \\ \mathbf{x}^{\star T} \mathbf{N}_0 \mathbf{x}^\star + 2\mathbf{p}_0^T \mathbf{x}^\star + b_0 \leq \nu, \\ \lambda^\star \geq 0, \\ \lambda^\star \left(\mathbf{x}^{\star T} \mathbf{N}_0 \mathbf{x}^\star + 2\mathbf{p}_0^T \mathbf{x}^\star + b_0 - \nu\right) = 0. \end{cases}$$

By solving the KKT conditions, we have

$$\begin{aligned} & \mathbf{x}^\star = \\ & -\left(\frac{\nu - \mathbf{w}_0^T\mathbf{M}_0\mathbf{w}_0 + \mathbf{w}_0^T\mathbf{M}_0\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{M}_0\mathbf{w}_0}{\left(\mathbf{e}^{(k)}\right)^T\mathbf{F}\left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)}}\right)^{\frac{1}{2}} \\ & \times \left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{e}^{(k)} - \left(\mathbf{F}^T\mathbf{M}_0\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{M}_0\mathbf{w}_0, \end{aligned}$$

and accordingly have $\mathbf{w}^\star = \mathbf{F}\mathbf{x}^\star + \mathbf{w}_0$.

## 4.13 Proof for Proposition 3

From the derivation of the MM-based algorithms (IRGEVP and IRGTRS in Algorithm 4.3 as well as E-IRGEVP and E-IRGTRS in Algorithm 4.4), we know that the objective function $f\left(\mathbf{w}\right)$ in (4.5.7) is majorized by functions $\overline{f}_1\left(\mathbf{w},\mathbf{w}^{(k)}\right)$ in Lemma 10 and $\overline{f}_2\left(\mathbf{w},\mathbf{w}^{(k)}\right)$ in Lemma 11 at $\mathbf{w}^{(k)}$ over the constraint $\mathcal{W}=\left\{\mathbf{w}^T\mathbf{M}_0\mathbf{w}=\nu\right\}\cap\mathcal{W}_i$, $(i=0,1)$. In the following, for the purpose of easy explanation, $\overline{f}_1\left(\mathbf{w},\mathbf{w}^{(k)}\right)$ and $\overline{f}_2\left(\mathbf{w},\mathbf{w}^{(k)}\right)$ will be jointly denoted as $\overline{f}\left(\mathbf{w},\mathbf{w}^{(k)}\right)$.

Based on (4.5.2) and (4.5.3) in Section 4.5.1, we can get the objective function value is monotonically nonincreasing at each iteration, i.e.,

$$\begin{aligned} f\left(\mathbf{w}^{(k+1)}\right) &\overset{(a)}{\leq} \overline{f}\left(\mathbf{w}^{(k+1)},\mathbf{w}^{(k)}\right) \\ &\overset{(b)}{\leq} \overline{f}\left(\mathbf{w}^{(k)},\mathbf{w}^{(k)}\right) \overset{(c)}{=} f\left(\mathbf{w}^{(k)}\right), \quad \forall k\in\mathbb{N}, \end{aligned}$$

where $(a)$ and $(c)$ follow from the A2) and A1) in (4.5.3), respectively, and $(b)$ follows from (4.5.2). It implies $\left\{f\left(\mathbf{w}^{(k)}\right)\right\}$ is a nonincreasing sequence, i.e.,

$$f\left(\mathbf{w}^{(0)}\right)\geq f\left(\mathbf{w}^{(1)}\right)\geq f\left(\mathbf{w}^{(2)}\right)\geq\cdots.$$

Assume that there exists a subsequence $\left\{\mathbf{w}^{(k_j)}\right\}$ converging to a limit point $\mathbf{w}^{(\infty)}$. We first have

$$\begin{aligned} \overline{f}\left(\mathbf{w}^{(k_{j+1})},\mathbf{w}^{(k_{j+1})}\right) &= f\left(\mathbf{w}^{(k_{j+1})}\right)\leq f\left(\mathbf{w}^{(k_j+1)}\right) \\ &\leq \bar{f}\left(\mathbf{w}^{(k_j+1)},\mathbf{w}^{(k_j)}\right)\leq \bar{f}\left(\mathbf{w},\mathbf{w}^{(k_j)}\right),\ \forall\mathbf{w}\in\mathcal{W}. \end{aligned}$$

Letting $j\to\infty$, we can further obtain

$$\bar{f}\left(\mathbf{w}^{(\infty)},\mathbf{w}^{(\infty)}\right)\leq\bar{f}\left(\mathbf{w},\mathbf{w}^{(\infty)}\right),\ \forall\mathbf{w}\in\mathcal{W},$$

i.e., $\mathbf{w}^{(\infty)}$ is the global minimizer of $\bar{f}\left(\mathbf{w},\mathbf{w}^{(\infty)}\right)$ over $\mathcal{W}$. Based on the B-stationarity defined in Section 4.5.1, we have

$$\bar{f}'\left(\mathbf{w}^{(\infty)},\mathbf{w}^{(\infty)};\mathbf{d}\right)\geq 0,\ \forall\mathbf{d}\in\mathcal{T}_{\mathcal{W}}\left(\mathbf{w}^{(\infty)}\right).$$

Then, according to the A3) in (4.5.3), we have

$$f'\left(\mathbf{w}^{(\infty)};\mathbf{d}\right)\geq 0,\,\forall\mathbf{d}\in\mathcal{T}_{\mathcal{W}}\left(\mathbf{w}^{(\infty)}\right),$$

which implies $\mathbf{w}^{(\infty)}$ is a B-stationary point of problem (4.5.7).

# Chapter 5

# Mean-Reverting Portfolio Design With A Leverage Constraint

The optimal mean-reverting portfolio (MRP) design problem is an important task for statistical arbitrage, a.k.a. pairs trading, in the financial markets. The target of the problem is to construct a portfolio of the underlying assets (possibly with an asset selection target) that can exhibit a satisfactory mean reversion property and a desirable variance property. In this chapter, the optimal MRP design problem is studied under an investment leverage constraint representing the total investment positions on the underlying assets. A general problem formulation is proposed by considering the design targets subject to a leverage constraint. To solve the problem, a unified optimization framework based on the successive convex approximation (SCA) method is developed. The superior performance of the proposed formulation and the algorithms are verified through numerical simulations on both synthetic data and real market data.

## 5.1 Introduction

Statistical arbitrage [32], also known as *Stat Arb*, is a quantitative investment and trading strategy widely used by many parties in the financial markets, e.g., institutional investors, hedge funds, mutual funds, proprietary trading firms, and individual investors [59]. In statistical arbitrage, the trading basket usually consists of many financial assets of possibly different categories such as equities, options, bonds, futures, commodities, etc. To arbitrage from the markets, investors need to buy the under-priced assets and short-sell or, more plainly, borrow and sell the over-priced ones. The profits will finally be locked in by unwinding the trading positions when the mispricings of the assets correct themselves in the future. Such an investment strategy is usually coined as a contrarian relative-value strategy [55]. In statistical arbitrage, the arbitrage opportunities exist as a consequence of the market inefficiency [56]. As revealed by the name, the design of trading baskets and trading actions largely relies on statistical analysis [92].

Statistical arbitrage dated back to the well-known trading strategy called pairs trading [46], [48], [54], which was firstly developed at Morgan Stanley by a quantitative trading group under the lead of Nunzio Tartaglia in the mid 1980s in the Wall Street [49]. Pairs trading, as a special scenario, falls into the umbrella of statistical arbitrage and, as indicated by the name, it is often used when there are only two assets in the trading basket. Since statistical arbitrage is able to hedge the overall market or systematic risk, and hence the profits are independent from the movements and the conditions of the prevailing markets (volatile, flat, or falling), it is also named as a market neutral strategy or an absolute return strategy [57], [58].

In statistical arbitrage, the trading basket is used to design a “spread” which characterizes the mis-pricings (also called the “relative pricing”) of the underlying assets. The designed spread is stationary, hence mean-reverting, and virtually represents the price for a synthetic mean-reverting asset [13]. In order to make profits, the trading process is carried out based on the mean reversion (MR) behavior of the spread around its statistical equilibrium, and hence named mean reversion trading or convergence trading [93]. For example, a simple mean reversion trading design could be buying the spread when it is below the equilibrium and selling it when it is above the equilibrium. Statistical arbitrage or pairs trading is accordingly also referred to as spread trading in the literature [50], [52], [53]. In practice, there are many existing methods to design a trading spread based on different philosophies, such as the distance method [47], the cointegration analysis method [46], the factor analysis method [63], the Copulas method [94], the stochastic modeling method [95], [96], and so on. In this chapter, we will only focus on the cointegration analysis method where the spread is constructed by a formal time series analysis [26]. The concept of “cointegration” was first come up with by Clive W. J. Granger in [27] and later in [28] to describe the linear stationary relationships within nonstationary time series which are named to be cointegrated. Later, the cointegrated vector autoregressive model was put forward into time series modeling [12],

[30] to efficiently estimate the cointegration relations. To honor the discovery of cointegration statistical property in time series, Granger was awarded the Nobel Prize in Economic Sciences in 2003. The cointegration relations have been verified by empirical analyses in many different financial markets to get statistical arbitrage opportunities [62], [64], [97].

Traditionally, cointegration analysis methods like the Engle-Granger ordinary least squares (OLS) method [28] and the Johansen method [30] are used to estimate the trading spreads from the underlying assets. An asset that naturally shows stationarity is a spread as well [98], e.g., the option implied volatility for stocks. Inherent correlations, however, may exist among different spreads. For example, when using the Johansen method, although many distinct spreads could be estimated from the same underlying assets, they essentially fall into the same "cointegration space". When having multiple spreads, a natural and interesting question is put forward: Can we design an optimal portfolio of these underlying assets? This question will be addressed in this chapter. A portfolio of the mean-reverting spreads is named a mean-reverting portfolio (MRP) or sometimes a long-short portfolio. To design an optimal MRP, two factors should be considered. Firstly, the designed MRP should exhibit a strong MR property so that it has frequent mean-crossing points and hence brings in more trading opportunities. Secondly, the designed MRP should exhibit large variance property so that each trade can provide sufficient profit. These two factors together naturally result in a multi-objective optimization problem, i.e., to find a desirable trade-off between MR and variance.

In [67], the author first proposed to design an MRP by optimizing a criterion characterizing the mean reversion strength. Later, authors in [68] and [69] realized that directly solving the MRP design problem in [67] could result in a portfolio with very low variance, then the variance control was taken into consideration and several new mean reversion criteria were also brought up. However, all the aforementioned MRP design problems were carried out by imposing an $\ell_2$-norm constraint on the portfolio weights. In [99], the authors argued that the $\ell_2$-norm has a physical meaning of power constraint in many signal processing problems (like beamforming in wireless communications), but its practical significance in the financial context is unclear. As a result, the investment budget constraint (a linear constraint) was firstly proposed in [99] and then in [13]. Compared to [68] and [69], the proposed methods in [13] make the designed portfolio more explainable and practical, in a sense that it explicitly represents the budget allocation on different underlying assets. However, one prominent issue incurred by the methods of using investment budget constraints as in [13] is that the designed portfolio could lead to a very large leverage (i.e., the total dollar position, both in longs and shorts), which makes the methods not always acceptable for practical use in real investments.

In this chapter, we are going to propose a new formulation for the optimal MRP design problem by jointly optimizing the two factors (i.e., MR and variance) subject to an investment leverage constraint [100]. To make it clear, the contributions of this chapter are summarized as follows.

- A general problem formulation for optimal MRP design is proposed that aims at finding a desirable trade-off between the MR and the variance of the portfolio, while subject to a practical leverage constraint instead of a budget constraint. Different MR criteria and variance criteria are considered in the formulation. The portfolio leverage constraint takes two cases into consideration, namely, the case of cointegration space and the case of naturally stationary assets.

- Besides the MR and variance criteria, the asset selection criterion is further considered in the optimal MRP design problem. Finally, the formulation becomes a constrained nonconvex problem.

- A unified algorithm framework based on the successive convex approximation (SCA) method named SCA-MRP is proposed to solve the MRP design problem, which tackles the original highly nonconvex problem by solving a sequence of easy convex subproblems.

- In order to efficiently solve the convex inner subproblems in SCA-MRP and to address different design cases in practice, several methods are proposed. The Armijo-like backtracking line search method is proposed to accelerate the SCA-MRP algorithm.

- The algorithm complexity and convergence to a stationary point are analyzed for the SCA-MRP algorithm.

- Numerical simulations on both synthetic and real market data are carried out to address the efficacy of the proposed MRP design problem formulation and the algorithms.

The remaining sections of this chapter are organized as follows. In Section 5.2, the optimal MRP design problem is briefly introduced. A general problem formulation for the optimal

MRP design is given in Section 5.3. Section 5.4 generally introduces the SCA method. The SCA-based algorithm called SCA-MRP is elaborated in Section 5.5 and three efficient algorithms to solve the convex subproblems are given in Section 5.6. The algorithm complexity analysis and convergence analysis are given in Section 5.7. Numerical performance is evaluated in Section 5.8 and, finally, concluding remarks are drawn in Section 5.9.

## 5.2 Optimal Mean-Reverting Portfolio Design

### 5.2.1 Mean-Reverting Portfolio (MRP)

For a financial asset, e.g., a common stock, a future contract, or a portfolio of them, its price at time $t$ is denoted by $p_t \in \mathbb{R}_+$, and then its corresponding logarithmic price or log-price $y_t \in \mathbb{R}$ is given by $y_t \triangleq \log(p_t)$[1]. Let $\mathbf{y}_t \triangleq [y_{1,t}, \ldots, y_{M,t}]^T$ denote the log-prices of $M$ assets. The (log-price) spread $s_t$ is given by $s_t \triangleq \boldsymbol{\beta}^T \mathbf{y}_t$, where $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_M]^T$ denotes the weights or hedge ratios. Suppose there exists a subspace, termed cointegration space, with $N$ (usually $N \leq M$) cointegration relations defined by $\mathbf{B} \triangleq [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N]$. Then these $N$ spreads are obtained as

$$\mathbf{s}_t \triangleq \mathbf{B}^T \mathbf{y}_t, \tag{5.2.1}$$

where every element of $\mathbf{s}_t$ is a spread. Specifically, if the asset log-prices are stationary in nature, every element of $\mathbf{y}_t$ can be defined as a spread, i.e., $\mathbf{s}_t = \mathbf{y}_t$ with $\mathbf{B} = \mathbf{I}$ $(N = M)$.

Different spreads may possess different mean reversion and variance properties in nature. The objective of the MRP design problem is to construct a portfolio of the underlying spreads to attain desirable trading properties. For the $N$ spreads in $\mathbf{s}_t$, the MRP is denoted by the portfolio weight $\mathbf{w} \triangleq [w_1, \ldots, w_N]^T$ with its resulting spread given by

$$z_t \triangleq \mathbf{w}^T \mathbf{s}_t = \sum_{n=1}^{N} w_n s_{n,t}. \tag{5.2.2}$$

Based on (5.2.1) and (5.2.2), we can further get the spread $z_t$ defined on the underlying assets as follows:

$$z_t \triangleq \mathbf{w}_p^T \mathbf{y}_t, \tag{5.2.3}$$

where $\mathbf{w}_p \triangleq \mathbf{B}\mathbf{w} \in \mathbb{R}^M$ denotes the MRP weights on the underlying assets and represents

[1] The $\log(\cdot)$ is the natural logarithm function.

the dollar value proportion invested on the underlying assets. For each asset $m = 1, \ldots, M$, the sign of $w_{p,m}$ indicates the type of positions, namely, $w_{p,m} > 0$ means a long position (i.e., it is bought), $w_{p,m} < 0$ means a short position (i.e., it is short-sold), and $w_{p,m} = 0$ means no position on the asset.

In the following, we continue to introduce some criteria for MR, variance, and asset selection.

## 5.2.2 Mean Reversion (MR) Criteria

Several MR criteria were used in [13], [69] and will be briefly introduced here. We start by defining the $i$th order (lag-$i$) autocovariance matrix for the spreads $\mathbf{s}_t$ as $\mathbf{M}_i \triangleq \mathsf{Cov}\left(\mathbf{s}_t, \mathbf{s}_{t+i}\right) = \mathsf{E}\left[\left(\mathbf{s}_t - \mathsf{E}\left[\mathbf{s}_t\right]\right)\left(\mathbf{s}_{t+i} - \mathsf{E}\left[\mathbf{s}_{t+i}\right]\right)^T\right]$ with $i \in \mathbb{N}$. Specifically, when $i = 0$, $\mathbf{M}_0$ stands for the (positive definite) covariance matrix of $\mathbf{s}_t$. Since we can always compute the centered form for $\mathbf{s}_t$, as $\tilde{\mathbf{s}}_t = \mathbf{s}_t - \mathsf{E}\left[\mathbf{s}_t\right]$, without loss of generality, $\mathbf{s}_t$ will be used to denote $\tilde{\mathbf{s}}_t$ in the following.

### 5.2.2.1 Predictability Statistics $\mathrm{pre}\left(\mathbf{w}\right)$

Consider a centered univariate stationary autoregressive process $z_t = \hat{z}_{t-1} + \epsilon_t$, where $\hat{z}_{t-1}$ is the prediction of $z_t$ at time $t-1$, and $\epsilon_t$ denotes a white noise. The predictability statistics [74] is proposed to measure how close a random process is to a white noise and defined by $\mathrm{pre} \triangleq \sigma_{\hat{z}}^2/\sigma_z^2$, where $\sigma_z^2 \triangleq \mathsf{E}\left[z_t^2\right]$ and $\sigma_{\hat{z}}^2 \triangleq \mathsf{E}\left[\hat{z}_{t-1}^2\right]$. Given the spread $z_t = \mathbf{w}^T\mathbf{s}_t$, the predictability statistics for spread $z_t = \mathbf{w}^T\mathbf{s}_t$ is computed as

$$\mathrm{pre}\left(\mathbf{w}\right) \triangleq \frac{\mathbf{w}^T\mathbf{T}\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}, \tag{5.2.4}$$

where $\mathbf{T} \triangleq \mathbf{M}_1^T\mathbf{M}_0^{-1}\mathbf{M}_1$. To design a spread $z_t$ as close as possible to white noise, we need to minimize $\mathrm{pre}\left(\mathbf{w}\right)$.

### 5.2.2.2 Portmanteau Statistics $\mathrm{por}\left(p, \mathbf{w}\right)$

The portmanteau statistics of order $p$ [75] for a centered univariate stationary process $z_t$ is defined as $\mathrm{por}\left(p\right) \triangleq \sum_{i=1}^{p} \rho_i^2$, where $\rho_i$ is the $i$th order autocorrelation of $z_t$ defined as $\rho_i \triangleq \mathsf{E}\left[z_t z_{t+i}\right]/\mathsf{E}\left[z_t^2\right]$. The measure $\mathrm{por}\left(p\right)$ is used to test whether a random process is close to a white noise. To design a spread $z_t$ close to white noise, we need to minimize $\mathrm{por}_z\left(p\right)$

with a prespecified order $p$. Given $z_t = \mathbf{w}^T\mathbf{s}_t$, we can get $\mathrm{por}\left(p, \mathbf{w}\right)$ as follows:

$$\mathrm{por}\left(p, \mathbf{w}\right) \triangleq \sum_{i=1}^{p} \left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2. \tag{5.2.5}$$

#### 5.2.2.3 Crossing Statistics $\mathrm{cro}\left(\mathbf{w}\right)$ and Penalized Crossing Statistics $\mathrm{pcro}\left(p, \mathbf{w}\right)$

The zero-crossing rate for a centered stationary process $z_t$ is defined as $\mathrm{zcr} \triangleq \left(T-1\right)^{-1}\sum_{t=2}^{T}\mathbf{1}\left\{z_t z_{t-1} \leq 0\right\}$, where

$$\mathbf{1}\left\{z_t z_{t-1} \leq 0\right\} \triangleq \begin{cases} 1, & z_t z_{t-1} \leq 0 \\ 0, & \mathrm{otherwise} \end{cases}$$

is the indicator function defined on $z_t$. It tests the probability that a process crosses its mean per unit of time. According to [77], for a centered stationary Gaussian process, zero-crossing rate is defined as $\mathrm{zcr} = \pi^{-1}\arccos\left(\rho_1\right)$. To design a spread exhibiting sufficient zero-crossings, we should minimize $\rho_1$. So given $z_t = \mathbf{w}^T\mathbf{s}_t$, we define the crossing statistics as

$$\mathrm{cro}\left(\mathbf{w}\right) \triangleq \frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}. \tag{5.2.6}$$

Besides minimizing the criterion $\mathrm{cro}\left(\mathbf{w}\right)$, it is also proposed to minimize the absolute high order autocorrelations $|\rho_i|$'s ($i = 2, \ldots, p$) [13]. Based on $\mathrm{cro}\left(\mathbf{w}\right)$, the penalized crossing statistics of order $p$ is defined as follows:

$$\mathrm{pcro}\left(p, \mathbf{w}\right) \triangleq \frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} + \eta\sum_{i=2}^{p}\left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2, \tag{5.2.7}$$

where $\eta$ is a positive prespecified factor.

### 5.2.3 Variance Criteria

Given a spread $z_t = \mathbf{w}^T\mathbf{s}_t$, its variance is naturally given by $\mathsf{Var}\left[z_t\right] = \mathsf{E}\left[z_t^2\right] = \mathbf{w}^T\mathbf{M}_0\mathbf{w}$. Another criterion we will consider is the standard deviation of $z_t$ which is given by $\mathsf{Std}\left[z_t\right] = \sqrt{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}$.

### 5.2.4 Asset Selection Criterion

In portfolio design problems, allocating capital to all the assets can increase significantly the transaction costs, which motivates to select a subset of assets [101]. To realize this asset selection goal in MRP design, it is desirable to pursue sparsity in the cointegration space $\mathbf{B}$. Based on the $\ell_0$-"norm"[2] [21], the asset selection criterion is accordingly given by $\|\mathbf{Bw}\|_0$.

### 5.2.5 Portfolio Leverage Constraint

The constraint on portfolio weights in portfolio design problems represent the investment policy and allocation [70]. As we mentioned in the introduction, the returns from statistical arbitrage are usually small. Hence, investors in practice may want to use leverage to multiply the returns.

In [67], [68], [69], the $\ell_2$-norm $\|\mathbf{w}\|_2$ was considered as a portfolio constraint. The $\ell_2$-norm is commonly used as a power constraint in electrical engineering like wireless communications and radar, but its practical significance in financial applications is unclear since imposing the $\ell_2$-norm on portfolio weights does not carry any physical meaning in a financial context. To address this issue, in our previous chapter [13], the budget constraint $\mathbf{1}^T\mathbf{w} = \mathbf{1}$ was proposed, but still fails to take the "portfolio leverage" into account which is the key practical consideration in portfolio design. In this chapter, we will use a general investment leverage constraint given as follows:

$$\mathcal{W} \triangleq \{\mathbf{w} \mid \|\mathbf{Bw}\|_1 \leq L\},$$

where $\mathbf{B}$ is the cointegration space and $L$ means the total investment leverage considering both long and short positions deployed on the underlying financial assets.

## 5.3 The Optimal MRP Design Problem

### 5.3.1 Problem Formulation

Considering the three design criteria previously described, i.e., MR criterion, variance criterion, and asset selection criterion, a general problem formulation for the optimal MRP design

[2] Strictly speaking, it is not a norm. For $\mathbf{x} \in \mathbb{R}^N$, the $\ell_0$-"norm" $\|\mathbf{x}\|_0 \triangleq \sum_{i=1}^{N} \operatorname{sgn}(|x_i|)$, where $\operatorname{sgn}(\cdot)$ is the sign function.

problem is given as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & F(\mathbf{w}) \triangleq U(\mathbf{w}) + \mu V(\mathbf{w}) + \gamma S(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{5.3.1}$$

which is a nonconvex constrained problem. The constant $\mu > 0$ defines the trade-off between the portfolio MR measure and variance preference. The regularizing parameter $\gamma \geq 0$ controls the sparsity level. The three terms $U(\mathbf{w})$, $V(\mathbf{w})$, and $S(\mathbf{w})$ composing the objective are described below in detail.

**T1)** The mean reversion term $U(\mathbf{w})$ considering different MR criteria can be jointly represented as

$$U(\mathbf{w}) \triangleq \xi \frac{\mathbf{w}^T\mathbf{H}\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} + \zeta \left(\frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2 + \eta \sum_{i=2}^{p} \left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}}\right)^2,$$

which contains as specific cases the predictability statistics $\mathrm{pre}(\mathbf{w})$ ($\xi = 1$, $\mathbf{H} = \mathbf{T}$, and $\zeta = \eta = 0$), the portmanteau statistics $\mathrm{por}(p, \mathbf{w})$ ($\xi = 0$ and $\zeta = \eta = 1$), the crossing statistics $\mathrm{cro}(\mathbf{w})$ ($\xi = 1$, $\mathbf{H} = \mathbf{M}_1$, and $\zeta = \eta = 0$), and the penalized crossing statistics $\mathrm{pcro}(p, \mathbf{w})$ ($\xi = 1$, $\mathbf{H} = \mathbf{M}_1$, $\zeta = 0$, and $\eta > 0$).

**T2)** The variance term $V(\mathbf{w})$ can be represented in the following four different forms:

$$V(\mathbf{w}) \triangleq \begin{cases} 1/\left(\mathbf{w}^T\mathbf{M}_0\mathbf{w}\right) & (\mathrm{VarInv}(\mathbf{w})) \\ 1/\sqrt{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} & (\mathrm{StdInv}(\mathbf{w})) \\ -\mathbf{w}^T\mathbf{M}_0\mathbf{w} & (\mathrm{VarNeg}(\mathbf{w})) \\ -\sqrt{\mathbf{w}^T\mathbf{M}_0\mathbf{w}} & (\mathrm{StdNeg}(\mathbf{w})). \end{cases}$$

**T3)** The asset selection term $S(\mathbf{w})$ is given by

$$S(\mathbf{w}) \triangleq \|\mathbf{B}\mathbf{w}\|_0 = \textstyle\sum_{i=1}^{M} \mathrm{sgn}\left(\left|[\mathbf{B}\mathbf{w}]_i\right|\right),$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $[\mathbf{a}]_i$ denotes the $i$th element in $\mathbf{a}$.

### 5.3.2 Observations and Insight

In this section, we will focus on the analysis of the optimal MRP design problem formulation in (5.3.1). Some observations and insight are given in the following.

**Lemma 12.** *Given any two colinear MRPs:* $\mathbf{w}_1$ *and* $\mathbf{w}_2 \triangleq \alpha \mathbf{w}_1$ *(*$\mathbf{w}_1 \neq \mathbf{0}$ *and* $\alpha \neq 0$*), we have: i)* $U(\mathbf{w}_1) = U(\mathbf{w}_2)$*; ii)* $V(\mathbf{w}_1) > (<, =)\, V(\mathbf{w}_2)$*, when* $|\alpha| < (>, =)\, 1$*; iii)* $S(\mathbf{w}_1) = S(\mathbf{w}_2)$*; iv)* $F(\mathbf{w}_1) = F(\mathbf{w}_2)$*, when* $|\alpha| = 1$*; and v)* $|\alpha| \|\mathbf{B}\mathbf{w}_1\|_1 = \|\mathbf{B}\mathbf{w}_2\|_1$.

**Proof 5.** *The proof is trivial and hence omitted.*

In Lemma 12, the points *i)-iii)* reveal that increasing the leverage level $L$ on an MRP can only increase its variance, but not change its MR and asset selection properties. The point *iv)* reveals that two MRPs with the opposite sign of weights $\mathbf{w}$ are essentially the same; or in other words, two trading spreads defined by $\mathbf{w}^T\mathbf{s}_t$ and $-\mathbf{w}^T\mathbf{s}_t$ are the same. This is really to the nature of MRP design, because in statistical arbitrage the actual investment not only depends on $\mathbf{w}$, which defines a spread, but also on whether a long or short position is taken on this spread later in the trading stage.

Based on Lemma 12, we further have the following result. Denote the set of optimal solutions of Problem (5.3.1) as $\mathcal{W}^\star \triangleq \{\mathbf{w}^\star | F(\mathbf{w}^\star) \leq F(\mathbf{w}), \forall \mathbf{w} \neq \mathbf{0}, \mathbf{w} \in \mathcal{W}\}$, then $\mathcal{W}^\star \subseteq \mathrm{bd}(\mathcal{W})$ (the boundary of set $\mathcal{W}$). The proof is trivial and hence omitted. Lemma 5.3.2 essentially reveals the inequality leverage constraint is always active, i.e., the designed MRPs always attain the total leverage $L$.

As mentioned before, the cointegration matrix $\mathbf{B}$, in practice, is commonly estimated based on time series modeling. However, the matrix $\mathbf{B}$ is not unique [26]. (Assuming the singular value decomposition $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, the cointegration space (i.e., the column space of $\mathbf{B}$) is given by $\mathcal{R}(\mathbf{U})$). Then based on Lemma 5.3.2, another intriguing observation for the MRP design problem (5.3.1) is given in the following.

**Proposition 4.** *Suppose there exist* $\mathbf{B}, \mathbf{B}' \in \mathcal{R}(\mathbf{U})$ *with the corresponding designed optimal MRPs from Problem* (5.3.1) *denoted as* $\mathcal{W}_p^\star \triangleq \left\{\mathbf{w}_p^\star | \mathbf{w}_p^\star = \mathbf{B}\mathbf{w}^\star, \forall \mathbf{w}^\star \in \mathcal{W}^\star\right\}$ *and* $\mathcal{W}_p'^\star \triangleq \left\{\mathbf{w}_p'^\star | \mathbf{w}_p'^\star = \mathbf{B}'\mathbf{w}'^\star, \forall \mathbf{w}'^\star \in \mathcal{W}'^\star\right\}$ *respectively, we have* $\mathcal{W}_p^\star = \mathcal{W}_p'^\star$.

**Proof 6.** *See Appendix 5.10.*

This result reveals that the optimal MRP $\mathbf{w}_p^\star$ designed from Problem (5.3.1) does not depend on the explicit form of $\mathbf{B}$, but instead only on the subspace $\mathcal{R}(\mathbf{U})$.

### 5.3.3 Mild Problem Modifications

The objective function $F(\mathbf{w})$ in (5.3.1) is not well-defined at $\mathbf{0}$ making it discontinuous over $\mathcal{W}$. Some mild modifications to $F(\mathbf{w})$ will be introduced in this section. Firstly, since $U(\mathbf{w})$ and $V(\mathbf{w})$ (refer to VarInv $(\mathbf{w})$ and StdInv $(\mathbf{w})$) are singular at $\mathbf{0}$, we propose to reduce this "singularity" by defining two modified criteria $U^{\epsilon}(\mathbf{w})$ and $V^{\epsilon}(\mathbf{w})$ as follows:

$$\begin{aligned} U^{\epsilon}(\mathbf{w}) \triangleq\, & \xi \frac{\mathbf{w}^T\mathbf{H}\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon} + \zeta\left(\frac{\mathbf{w}^T\mathbf{M}_1\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon}\right)^2 \\ & + \eta\sum_{i=2}^{p}\left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon}\right)^2, \end{aligned} \tag{5.3.2}$$

and

$$V^{\epsilon}(\mathbf{w}) \triangleq \begin{cases} 1/\left(\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon\right) & (\text{VarInv}(\mathbf{w})) \\ 1/\sqrt{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon} & (\text{StdInv}(\mathbf{w})) \\ -\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon & (\text{VarNeg}(\mathbf{w})) \\ -\sqrt{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon} & (\text{StdNeg}(\mathbf{w})), \end{cases} \tag{5.3.3}$$

where $\epsilon > 0$ is a small constant. Secondly, since the sparsity criterion $S(\mathbf{w})$ is nonconvex and discontinuous, the following smooth nonconvex sparsity function will be considered

$$S^{\epsilon}(\mathbf{w}) \triangleq \textstyle\sum_{m=1}^{M}\left[1-\exp\left(-\epsilon^{-1}\left|[\mathbf{B}\mathbf{w}]_m\right|^2\right)\right], \tag{5.3.4}$$

where compared to $S(\mathbf{w})$ the function $1-\exp(-\epsilon^{-1}(\cdot)^2)$ is used to approximate $\mathrm{sgn}(\cdot)$ with $\epsilon > 0$ controlling the approximation tightness [102]. Finally, the the modified objective is given as $F^{\epsilon}(\mathbf{w}) \triangleq U^{\epsilon}(\mathbf{w}) + \mu V^{\epsilon}(\mathbf{w}) + \gamma S^{\epsilon}(\mathbf{w})$ which is almost "equivalent" to $F(\mathbf{w})$.

Now, we are ready to discuss the solving procedure for the optimal MRP design problem in (5.3.1). We will firstly introduce a general algorithmic framework based on the idea of successively approximating the original nonconvex problem with a series of convex subproblems, and the derived algorithms are expected to be simple and efficient with provable convergence to a stationary point.

## 5.4 The Successive Convex Approximation Method

The successive convex approximation (SCA) method [103] is a general optimization framework especially for solving nonconvex optimization problems. In this chapter, we will use the SCA method proposed in [104] which is based on solving a sequence of simpler strongly convex problems, preserves feasibility of the iterates for the original nonconvex problem, and also has guaranteed convergence.

Specifically, an optimization problem is given as follows:

$$\begin{array}{ll} \underset{\mathbf{x}}{\mathsf{minimize}} & f\left(\mathbf{x}\right) \\ \mathsf{subject\ to} & \mathbf{x} \in \mathcal{X}, \end{array} \tag{5.4.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^N$ is convex and $f\left(\mathbf{x}\right)$ is nonconvex and (possibly) nonsmooth. In order to solve Problem (5.4.1) which could be difficult to tackle directly, starting from an initial feasible point $\mathbf{x}^{(0)}$, the SCA method solves a series of subproblems with surrogate functions $\tilde{f}\left(\mathbf{x};\mathbf{x}^{(k)}\right)$ (or simply denoted as $\tilde{f}^{(k)}\left(\mathbf{x}\right)$) approximating the original objective $f\left(\mathbf{x}\right)$ over the set $\mathcal{X}$. A sequence $\left\{\mathbf{x}^{(k)}\right\}$ is generated by the following rules:

$$\begin{cases} \hat{\mathbf{x}}^{(k+1)} = \arg\min\limits_{\mathbf{x}\in\mathcal{X}} \tilde{f}(\mathbf{x};\mathbf{x}^{(k)}) & (a) \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \gamma^{(k)}(\hat{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}). & (b) \end{cases} \tag{5.4.2}$$

The first step is to generate the descent direction (i.e., $\hat{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}$) by solving a best-response problem $(a)$, and the second step is the variable update rule with $\gamma^{(k)}$ to be the step-size.

Convergence to a stationary solution of the original nonconvex optimization problem in (5.4.1) can be established under the following mild assumptions on the problem:

**A1)** $\mathcal{X}$ is nonempty, closed, and convex;

**A2)** $\nabla_{\mathbf{x}} f\left(\mathbf{x}\right)$ is $L_{\nabla f}$-Lipschitz continuous on $\mathcal{X}$;

**A3)** $f\left(\mathbf{x}\right)$ is coercive on $\mathcal{X}$.

And as to $\tilde{f}\left(\mathbf{x};\mathbf{x}^{(k)}\right)$, the following conditions are needed:

**B1)** given $\mathbf{x}^{(k)}$, $\tilde{f}\left(\mathbf{x};\mathbf{x}^{(k)}\right)$ is $c$-strongly convex on $\mathcal{X}$ for some $c > 0$, i.e., $\nabla^2_{\mathbf{x}}\tilde{f}\left(\mathbf{x};\mathbf{x}^{(k)}\right) \succeq c\mathbf{I}$;

**B2)** $\nabla_{\mathbf{x}}\tilde{f}\left(\mathbf{x}^{(k)};\mathbf{x}^{(k)}\right)=\nabla_{\mathbf{x}}f\left(\mathbf{x}^{(k)}\right)$ for all $\mathbf{x}^{(k)}\in\mathcal{X}$;

**B3)** $\nabla_{\mathbf{x}}\tilde{f}\left(\mathbf{x};\mathbf{x}\right)$ is continuous for all $\mathbf{x}\in\mathcal{X}$.

It is easy to see that the key point to use SCA is to find a good approximation function $\tilde{f}\left(\mathbf{x};\mathbf{x}^{(k)}\right)$, which could make the best response problem easy to solve and result in a fast convergence. In the following, we are going to apply the SCA method for the optimal MRP design problem in (5.3.1).

## 5.5 Problem Solving Based on The SCA Method

### 5.5.1 Using SCA For MRP Design

Applying the SCA method to Problem (5.3.1), we have the convex approximation function $\tilde{F}^{(k)}(\mathbf{w})$ at the $(k+1)$th iteration for the objective $F^{\epsilon}\left(\mathbf{w}\right)$ given as follows:

$$\tilde{F}^{(k)}(\mathbf{w})\triangleq\tilde{U}^{(k)}(\mathbf{w})+\mu\tilde{V}^{(k)}(\mathbf{w})+\gamma\tilde{S}^{(k)}(\mathbf{w})+\tau\|\mathbf{w}-\mathbf{w}^{(k)}\|_2^2,\tag{5.5.1}$$

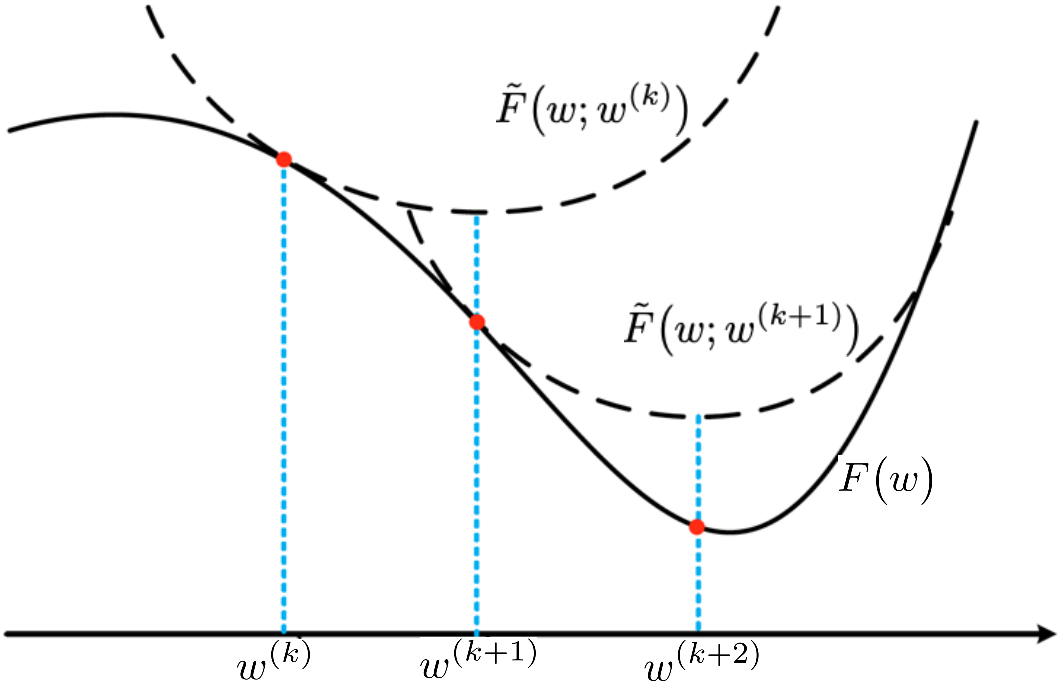
with $\tau\geq0$ denoting a parameter on the proximal term added for convergence [104]. An illustrative figure for the relation between $F^{\epsilon}\left(\mathbf{w}\right)$ and $\tilde{F}^{(k)}(\mathbf{w})$ is given in Figure 5.1.

**On The Approximation Term $\tilde{U}^{(k)}\left(\mathbf{w}\right)$** The term $\tilde{U}^{(k)}\left(\mathbf{w}\right)$ is a convex approximation for the MR term $U^{\epsilon}\left(\mathbf{w}\right)$. Based on the general idea of SCA, there could be many choices on deriving such an approximation. When $U^{\epsilon}\left(\mathbf{w}\right)$ is chosen as $\mathrm{pre}\left(\mathbf{w}\right)$ or $\mathrm{cro}\left(\mathbf{w}\right)$ (i.e., ratio of quadratic functions), the convex approximation is chosen to be a linearization of the criterion, which is simply given as

$$\tilde{U}^{(k)}\left(\mathbf{w}\right)\triangleq\mathbf{b}_U^{(k)T}\mathbf{w},\tag{5.5.2}$$

where

$$\mathbf{b}_U^{(k)}\triangleq2\xi\big(\mathbf{d}_{0,h}^{(k)}-\mathbf{d}_{h,0}^{(k)}\big)+2\zeta r_1^{(k)}\big(\mathbf{d}_{0,1}^{(k)}-\mathbf{d}_{1,0}^{(k)}\big)+2\eta\sum_{i=2}^{p}r_i^{(k)}\big(\mathbf{d}_{0,i}^{(k)}-\mathbf{d}_{i,0}^{(k)}\big),\tag{5.5.3}$$



$\tilde{F}\left(w; w^{(k)}\right)$
$\tilde{F}\left(w; w^{(k+1)}\right)$
$F(w)$
$w^{(k)}$
$w^{(k+1)}$
$w^{(k+2)}$

Figure 5.1: Solving Problem (5.3.1) with objective $F(\mathbf{w})$ by solving a sequence of strongly convex subproblems with quadratic objective functions $\tilde{F}\left(\mathbf{w}; \mathbf{w}^{(k)}\right)$ (or $\tilde{F}^{(k)}(\mathbf{w})$) in (5.5.1). (Illustration is shown in one dimension.)

with

$$\begin{aligned}
r_i^{(k)} &\triangleq \left(\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right),\ i=1,\ldots,p,\\
r_h^{(k)} &\triangleq \left(\mathbf{w}^{(k)T}\mathbf{H}\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right),\\
\mathbf{d}_{0,i}^{(k)} &\triangleq \left(\mathbf{M}_i\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right),\ i=1,\ldots,p,\\
\mathbf{d}_{0,h}^{(k)} &\triangleq \left(\mathbf{H}\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right),\\
\mathbf{d}_{i,0}^{(k)} &\triangleq \left(r_i^{(k)}\mathbf{M}_0\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right),\ i=1,\ldots,p,\\
\mathbf{d}_{h,0}^{(k)} &\triangleq \left(r_h^{(k)}\mathbf{M}_0\mathbf{w}^{(k)}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right).
\end{aligned} \tag{5.5.4}$$

When $U^\epsilon(\mathbf{w})$ is chosen as $\mathrm{por}(p, \mathbf{w})$ or $\mathrm{pcro}(p, \mathbf{w})$, in which case a "square of ratio of quadratic functions" term is involved, a nice approximation technique by exploring the convex curvature of $U^\epsilon(\mathbf{w})$ given in Example 1 will be employed.

**Remark 1.** *We first define a function* $u(\mathbf{w})$ *as*

$$u(\mathbf{w}) \triangleq \left(\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^T\mathbf{M}_0\mathbf{w}+\epsilon}\right)^2. \tag{5.5.5}$$

*At a given point* $\mathbf{w}^{(k)}$*, we describe three possible choices of the convex approximation function* $\tilde{u}^{(k)}(\mathbf{w})$ *in Equation* (5.5.6) *together with their visualizations in Figure 5.2.*

$$
\begin{aligned}
\tilde{u}_1^{(k)}\left(\mathbf{w}\right) &\triangleq \left(r_i^{(k)}\right)^2 + 4r_i^{(k)}\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)^T\left(\mathbf{w} - \mathbf{w}^{(k)}\right) \\
\tilde{u}_2^{(k)}\left(\mathbf{w}\right) &\triangleq 4\left(r_i^{(k)}\right)^3 - \left(r_i^{(k)}\right)^2 - 4r_i^{(k)}\left(\mathbf{d}_{i,0}^{(k)}\right)^T\mathbf{w} + 2r_i^{(k)}\left(\mathbf{w}^T\mathbf{M}_i\mathbf{w}\right)/\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)} + \epsilon\right) \\
\tilde{u}_3^{(k)}\left(\mathbf{w}\right) &\triangleq \left[r_i^{(k)} + 2\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)^T\left(\mathbf{w} - \mathbf{w}^{(k)}\right)\right]^2 \\
&= \left(r_i^{(k)}\right)^2 + 4r_i^{(k)}\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)^T\left(\mathbf{w} - \mathbf{w}^{(k)}\right) + 4\left(\mathbf{w} - \mathbf{w}^{(k)}\right)^T\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)^T\left(\mathbf{w} - \mathbf{w}^{(k)}\right)
\end{aligned}
\tag{5.5.6}
$$

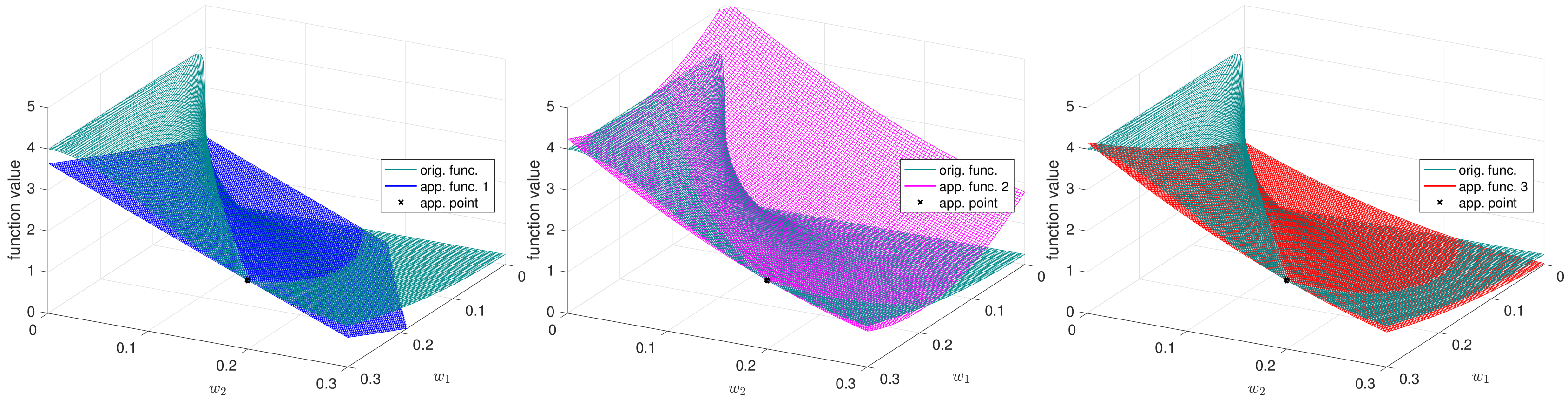


Figure 5.2: Approximation functions $\tilde{u}_1\left(\mathbf{w};\mathbf{w}^{(k)}\right)$, $\tilde{u}_2\left(\mathbf{w};\mathbf{w}^{(k)}\right)$, and $\tilde{u}_3\left(\mathbf{w};\mathbf{w}^{(k)}\right)$ (Three different convex approximation functions for $u\left(\mathbf{w}\right)$ in Equation (5.5.5) at approximation point $\mathbf{w}^{(k)} = (0.3, 0.2)^T$.).

The function $\tilde{u}_1^{(k)}(\mathbf{w})$ is based on the direct linearization of $u(\mathbf{w})$ and $\tilde{u}_2^{(k)}(\mathbf{w})$ is designed by linearization w.r.t. partial variables[3]. To obtain the third approximation function $\tilde{u}_3^{(k)}(\mathbf{w})$, $u(\mathbf{w})$ is convexified by linearizing the fractional term inside the square operation $(\cdot)^2$.

By comparing the subfigures (a), (b), and (c) in Figure **??**, we can find that $\tilde{u}_3^{(k)}(\mathbf{w})$ gives a much tighter approximation than the other two. For $\tilde{u}_3^{(k)}(\mathbf{w})$, it is easy to verify that the approximation technique ensures it has the same gradient as $u(\mathbf{w})$ at $\mathbf{w}^{(k)}$. We can also observe that the matrix $\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)\left(\mathbf{d}_{0,i}^{(k)} - \mathbf{d}_{i,0}^{(k)}\right)^T$ in $\tilde{u}_3^{(k)}(\mathbf{w})$ is in fact an approximation to the true Hessian matrix, which is known as an outer product approximation or Levenberg-Marquardt approximation [105]. Thus, it is reasonable to assume that based on $\tilde{u}_3^{(k)}(\mathbf{w})$ the overall resulting algorithm is able to largely maintain the information on the curvature of the cost function, even if higher order derivatives with respect to the gradient are never explicitly computed. In fact, this is an interesting line of reasoning, which could eventually lead to improved approximations for the cost functions. Finally, for $\mathrm{por}(p, \mathbf{w})$ or $\mathrm{pcro}(p, \mathbf{w})$, using the approximation technique for $\tilde{u}_3^{(k)}(\mathbf{w})$ in Example 1, we have $\tilde{U}^{(k)}(\mathbf{w})$ as follows:

$$\tilde{U}^{(k)}(\mathbf{w}) \triangleq \mathbf{w}^T \mathbf{A}_U^{(k)} \mathbf{w} + \mathbf{b}_U^{(k)T} \mathbf{w}, \tag{5.5.7}$$

where $\mathbf{A}_U^{(k)}$ is given by

$$\begin{aligned} \mathbf{A}_U^{(k)} \triangleq 4\zeta\big(&\mathbf{d}_{0,1}^{(k)}\mathbf{d}_{0,1}^{(k)T} + \mathbf{d}_{1,0}^{(k)}\mathbf{d}_{1,0}^{(k)T} - \mathbf{d}_{0,1}^{(k)}\mathbf{d}_{1,0}^{(k)T} - \mathbf{d}_{1,0}^{(k)}\mathbf{d}_{0,1}^{(k)T}\big) \\ &+ 4\eta \sum_{i=2}^{p} \big(\mathbf{d}_{0,i}^{(k)}\mathbf{d}_{0,i}^{(k)T} + \mathbf{d}_{i,0}^{(k)}\mathbf{d}_{i,0}^{(k)T} - \mathbf{d}_{0,i}^{(k)}\mathbf{d}_{i,0}^{(k)T} - \mathbf{d}_{i,0}^{(k)}\mathbf{d}_{0,i}^{(k)T}\big) \end{aligned} \tag{5.5.8}$$

with $\mathbf{d}_{0,i}^{(k)}$'s and $\mathbf{d}_{i,0}^{(k)}$ defined in Eq. (5.5.4) and $\mathbf{b}_U^{(k)}$ is given in (5.5.3).

**On The Approximation Term $\tilde{V}^{(k)}(\mathbf{w})$** The $\tilde{V}^{(k)}(\mathbf{w})$ denotes the convex (i.e., linearization) approximation for the variance term $V^{\epsilon}(\mathbf{w})$ given by

$$\tilde{V}^{(k)}(\mathbf{w}) \triangleq \mathbf{b}_V^{(k)T} \mathbf{w}, \tag{5.5.9}$$

[3] The details are given in Appendix 5.11.

where

$$\mathbf{b}_V^{(k)} \triangleq$$
$$\begin{cases} -2(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon)^{-2}\mathbf{M}_0\mathbf{w}^{(k)} & (\mathrm{VarInv}\,(\mathbf{w})) \\ -(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon)^{-\frac{3}{2}}\mathbf{M}_0\mathbf{w}^{(k)} & (\mathrm{StdInv}\,(\mathbf{w})) \\ -2\mathbf{M}_0\mathbf{w}^{(k)} & (\mathrm{VarNeg}\,(\mathbf{w})) \\ -(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon)^{-\frac{1}{2}}\mathbf{M}_0\mathbf{w}^{(k)} & (\mathrm{StdNeg}\,(\mathbf{w}))\,. \end{cases}$$

**On The Approximation Term $\tilde{S}^{(k)}(\mathbf{w})$** The $\tilde{S}^{(k)}(\mathbf{w})$ is the convex approximation for the sparsity term $S^{\epsilon}(\mathbf{w})$. To derive it, we need the following lemma. At any point $x^{(k)} \in \mathbb{R}$, a tight upperbound function for $s(x) \triangleq 1-\exp(-\epsilon^{-1}x^2)$ is obtained as follows:

$$\begin{aligned} s(x) \leq \tilde{s}^{(k)}(x) \triangleq & \epsilon^{-1}\exp\big(-\epsilon^{-1}(x^{(k)})^2\big)x^2 \\ & + 1-\exp\big(-\epsilon^{-1}(x^{(k)})^2\big)\big(1+\epsilon^{-1}(x^{(k)})^2\big), \end{aligned}$$

for $\forall x \in \mathbb{R}$. The proof is trivial and hence omitted. From Lemma 5.5.1, we have $\nabla_x \tilde{s}^{(k)}\left(x^{(k)}\right) = \nabla_x s\left(x^{(k)}\right)$. Then, based on the function $\tilde{s}^{(k)}(x)$, the approximation for $S(\mathbf{w})$ is accordingly given as follows:

$$\tilde{S}^{(k)}(\mathbf{w}) \triangleq \mathbf{w}^T\mathbf{A}_S^{(k)}\mathbf{w}, \tag{5.5.10}$$

with $\mathbf{A}_S^{(k)} \triangleq \epsilon^{-1}\mathbf{B}^T\mathrm{Diag}\left[\exp\left(-\epsilon^{-1}\left(\mathbf{B}\mathbf{w}^{(k)}\odot\mathbf{B}\mathbf{w}^{(k)}\right)\right)\right]\mathbf{B}$, where $\exp(\cdot)$ is taken elementwise and $\mathrm{Diag}[\mathbf{d}]$ is a matrix with diagonal elements formed by $\mathbf{d}$.

Finally, by combining $\tilde{U}^{(k)}(\mathbf{w})$, $\tilde{V}^{(k)}(\mathbf{w})$, and $\tilde{S}^{(k)}(\mathbf{w})$, $\tilde{F}^{(k)}(\mathbf{w})$ is accordingly written as

$$\tilde{F}^{(k)}(\mathbf{w}) \triangleq \mathbf{w}^T\mathbf{A}^{(k)}\mathbf{w}+\mathbf{b}^{(k)T}\mathbf{w}, \tag{5.5.11}$$

where $\mathbf{A}^{(k)} \triangleq \mathbf{A}_U^{(k)}+\gamma\mathbf{A}_S^{(k)}+\tau\mathbf{I}$, and $\mathbf{b}^{(k)} \triangleq \mathbf{b}_U^{(k)}+\mu\mathbf{b}_V^{(k)}-2\tau\mathbf{w}^{(k)}$. Based on the approximation $\tilde{F}^{(k)}(\mathbf{w})$ for $F(\mathbf{w})$, the subproblem to solve in the $(k+1)$th iteration is

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T\mathbf{A}^{(k)}\mathbf{w}+\mathbf{b}^{(k)T}\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{B}\mathbf{w}\|_1 \leq L, \end{aligned} \tag{5.5.12}$$

which is a convex problem. We can observe that the objective function in Problem (5.5.12) is quadratic in variable $\mathbf{w}$ instead of nonconvex in $\mathbf{w}$ as in Problem (5.3.1). Since it is convex, this problem can be efficiently solved, and some efficient methods for different cases will be discussed in detail in Section 5.6.

### 5.5.2 SCA-MRP: The Overall Algorithm

Based on SCA, in order to solve the original nonconvex problem in (5.3.1), we just need to iteratively solve a convex subproblem in (5.5.12). We name this SCA-based algorithm SCA-MRP and summarize it in Algorithm 5.1. The algorithm can be guaranteed to converge globally when the step-size $\gamma^{(k)}$ is chosen properly. A practical approach to choosing $\gamma^{(k)}$ is the Armijo-like backtracking line search rule [106], which is given as follows:

$$\begin{aligned} \text{Given} \quad & \alpha, \beta \in (0,1)\,,\ l = 0 \\ \text{While} \quad & \Delta F^{\epsilon}(\mathbf{w}^{(k)}) > -\alpha\beta^{l}\|\Delta\mathbf{w}^{(k)}\|_2^2 \\ & \quad l = l + 1 \\ \text{Let} \quad & \gamma^{(k)} = \beta^{l} \text{ for } k = 0, 1, 2, \ldots, \end{aligned}$$

where $\Delta F^{\epsilon}\left(\mathbf{w}^{(k)}\right) \triangleq F^{\epsilon}\left(\mathbf{w}^{(k)} + \beta^{l}\Delta\mathbf{w}^{(k)}\right) - F^{\epsilon}\left(\mathbf{w}^{(k)}\right)$ with $\Delta\mathbf{w}^{(k)} \triangleq \hat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k)}$.

**Algorithm 5.1** SCA-MRP: An SCA-Based Algorithm for The Optimal MRP Design Problem (5.3.1)

```
Require: $\mathbf{H}$, $\mathbf{M}_i$ $(i = 0, \ldots, p)$, $\mu$, $\gamma$, $\mathbf{B}$, $L$ and $\tau$
Set $k = 0$, $\gamma^{(0)}$ and $\mathbf{w}^{(0)}$.
repeat
    Compute $\mathbf{A}^{(k)}$ and $\mathbf{b}^{(k)}$ in (5.5.11)
    $\hat{\mathbf{w}}^{(k+1)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^T\mathbf{A}^{(k)}\mathbf{w} + \mathbf{b}^{(k)T}\mathbf{w}$
    $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \gamma^{(k)}(\hat{\mathbf{w}}^{(k+1)} - \mathbf{w}^{(k)})$
    $k \leftarrow k + 1$
until some convergence criterion is met
```

## 5.6 Solving Methods for The Inner Subproblem

In SCA-MRP, we need to solve a sequence of convex subproblems in each iteration (see Step 4 in Algorithm 5.1). This inner subproblem has no closed-form solution, but we can resort to the off-the-shelf public or commercial solvers like SeDuMi [107], SDPT3 [108], and MOSEK [109] or some popular convex optimization toolboxes (scripting languages) like YALMIP [110] and CVX [111]. However, as an alternative to the general-purpose solvers and toolboxes, we can also develop problem-specific algorithms to solve this problem more efficiently.

### 5.6.1 Algorithm Based on the ADMM Method

For the sake of notational simplicity, we omit the superscript $(k)$ in the SCA subproblem (5.5.12) and recast it as follows:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{B}\mathbf{w}\|_1 \leq L, \end{aligned} \tag{5.6.1}$$

where $\mathbf{A} \succ \mathbf{0}$.

The alternating direction method of multipliers (ADMM) is a method that can solve a convex optimization problem by breaking it into smaller parts, each of which are then easier to handle [112]. It has recently been applied on applications in a number of areas. To solve Problem (5.6.1) based on ADMM, we first rewrite it by introducing an auxiliary variable $\mathbf{z} = \mathbf{B}\mathbf{w}$, then the problem becomes

$$\begin{aligned} \underset{\mathbf{w},\mathbf{z}}{\text{minimize}} \quad & \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{z}\|_1 \leq L,\ \mathbf{B}\mathbf{w} - \mathbf{z} = \mathbf{0}, \end{aligned} \tag{5.6.2}$$

We further define an indicator function for the $\ell_1$-norm ball set as

$$I_{\mathcal{C}}(\mathbf{z}) \triangleq \begin{cases} 0, & \mathbf{z} \in \mathcal{C} \triangleq \left\{\mathbf{z} \middle| \left\|\mathbf{z}\right\|_1 \leq L\right\} \\ +\infty, & \text{otherwise}, \end{cases}.$$

Problem (5.6.2) can be written in the following standard ADMM form:

$$\begin{aligned} \underset{\mathbf{w},\mathbf{z}}{\text{minimize}} \quad & \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} + I_{\mathcal{C}}(\mathbf{z}) \\ \text{subject to} \quad & \mathbf{B}\mathbf{w} - \mathbf{z} = \mathbf{0}. \end{aligned} \tag{5.6.3}$$

And the augmented Lagrangian is given as follows:

$$\begin{aligned} & \mathcal{L}_\rho\left(\mathbf{w}, \mathbf{z}, \mathbf{u}\left(\mathbf{y}\right)\right) \\ =& \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} + I_{\mathcal{C}}(\mathbf{z}) + \mathbf{y}^T\left(\mathbf{B}\mathbf{w} - \mathbf{z}\right) + \frac{\rho}{2}\left\|\mathbf{B}\mathbf{w} - \mathbf{z}\right\|_2^2 \\ =& \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} + I_{\mathcal{C}}(\mathbf{z}) + \frac{\rho}{2}\left\|\mathbf{B}\mathbf{w} - \mathbf{z} + \mathbf{u}\right\|_2^2 + const., \end{aligned}$$

where $\rho > 0$ is the penalty parameter which serves as the dual update step-size and the scaled dual variable $\mathbf{u} \triangleq \frac{1}{\rho}\mathbf{y}$. Then, the ADMM updates are given in three variable blocks $(\mathbf{w}, \mathbf{z}, \mathbf{u})$ by

$$\begin{cases} \mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \Big\{ \mathbf{w}^T \mathbf{A} \mathbf{w} + \mathbf{b}^T \mathbf{w} \\ \qquad\qquad\qquad\qquad + \frac{\rho}{2} \left\| \mathbf{B}\mathbf{w} - \mathbf{z}^{(k)} + \mathbf{u}^{(k)} \right\|_2^2 \Big\} \\ \mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} \left\{ I_{\mathcal{C}}(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{z} - \mathbf{B}\mathbf{w}^{(k+1)} - \mathbf{u}^{(k)} \right\|_2^2 \right\} \\ \qquad\quad = \Pi_{\mathcal{C}}(\mathbf{B}\mathbf{w}^{(k+1)} + \mathbf{u}^{(k)}) \\ \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{B}\mathbf{w}^{(k+1)} - \mathbf{z}^{(k+1)}, \end{cases}$$

where the $\mathbf{z}$ update step is essentially a projection onto set $\mathcal{C}$ with $\Pi_{\mathcal{C}}(\cdot)$ denoting the projection operator.

Specifically, the update of variable $\mathbf{w}$ amounts to solving a convex quadratic programming (QP) which has the closed-form solution:

$$\mathbf{w}^{(k+1)} = -(2\mathbf{A} + \rho \mathbf{B}^T \mathbf{B})^{-1} (\mathbf{b} + \rho \mathbf{B}^T (\mathbf{u}^{(k)} - \mathbf{z}^{(k)})).$$

By defining $\mathbf{h}^{(k)} \triangleq \mathbf{B}\mathbf{w}^{(k+1)} + \mathbf{u}^{(k)}$, the variable $\mathbf{z}$ update is equivalent to solving

$$\begin{array}{ll} \underset{\mathbf{z}}{\text{minimize}} & \left\| \mathbf{z} - \mathbf{h}^{(k)} \right\|_2^2 \\ \text{subject to} & \|\mathbf{z}\|_1 \leq L, \end{array} \tag{5.6.4}$$

which is the classical projection onto the $\ell_1$-norm ball problem with efficient algorithms for problem solving [113, Lemma 1][114]. An efficient water-filling-like algorithm based on sorting is given in Algorithm 5.2.

In Algorithm 5.2, $\operatorname{sgn}(\cdot)$ is the sign function which extracts the sign of a real number; $\operatorname{abs}(\cdot)$ is the absolute value function; and $b_{(j)}$ $(1 \leq j \leq N)$ denotes the $j$-th largest element in $\mathbf{b}$. This algorithm gives a water-filling-like closed-form solution to Problem (5.6.4). In this ADMM-based algorithm, we have three blocks of variables to minimize, which could possibly be slow for convergence. For the primal variable $\mathbf{w}$ update, we also need to solve a convex QP involving the matrix inversion. In the following, we will develop an alternative algorithm.

**Algorithm 5.2** Euclidean Projection Onto An $\ell_1$-Norm Ball (5.6.4)

```
Require: $\mathbf{h}$ and $L$
if $||\mathbf{h}||_1 \leq L$ then
    $\mathbf{z} = \mathbf{h}$
    return $\mathbf{z}$
else
    $\mathbf{a} = \mathrm{sign}(\mathbf{h})$ and $\mathbf{b} = \mathrm{abs}(\mathbf{h})$
    Sort the elements in $\mathbf{b}$ as $b_{(1)} \geq b_{(2)} \geq \cdots \geq b_{(N)}$
    $\rho = \arg \max_{1 \leq j \leq N} \left\{ j \mid b_{(j)} > \frac{1}{j}\left( \sum_{i=1}^{j} b_{(i)} - L \right) \right\}$
    $\theta = \frac{1}{\rho}\left( \sum_{i=1}^{\rho} b_{(i)} - L \right)$
    $z_j = a_j \max\{b_j - \theta, 0\}, 1 \leq j \leq N$
    return $\mathbf{z}$
end if
```

### 5.6.2 Algorithm Based on the M-ADMM Method

The following method to solve the convex inner problem in (5.6.1) is based on majorized ADMM (M-ADMM) [115]. Compared to the vanilla ADMM, M-ADMM introduces the majorization-minimization (MM) [19] idea to find an upperbound function for the variable update. By minimizing instead an upperbound function, a cheap closed-form variable update can be achieved in many cases.

To use the M-ADMM method, based on Problem (5.6.1), we first define a new variable $\tilde{\mathbf{w}} \triangleq \mathbf{B}\mathbf{w}$. Then, we can equivalently have $\mathbf{w} = \mathbf{B}^{\dagger}\tilde{\mathbf{w}}$ and $\left(\mathbf{B}^{\perp}\right)^T \tilde{\mathbf{w}} = \mathbf{0}$,[4] where $\mathbf{B}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{B}$, and the columns of $\mathbf{B}^{\perp}$ span the orthogonal complementary subspace of $\mathbf{B}$, respectively. Then by defining $\tilde{\mathbf{A}} \triangleq \left(\mathbf{B}^{\dagger}\right)^T \mathbf{A}\mathbf{B}^{\dagger}$ and $\tilde{\mathbf{b}} \triangleq \left(\mathbf{B}^{\dagger}\right)^T \mathbf{b}$, Problem (5.6.1) can be equivalently rewritten in terms of $\tilde{\mathbf{w}}$ as

$$\begin{aligned} \underset{\tilde{\mathbf{w}}}{\text{minimize}} \quad & \tilde{\mathbf{w}}^T \tilde{\mathbf{A}} \tilde{\mathbf{w}} + \tilde{\mathbf{b}}^T \tilde{\mathbf{w}} \\ \text{subject to} \quad & \|\tilde{\mathbf{w}}\|_1 \leq L,\ \left(\mathbf{B}^{\perp}\right)^T \tilde{\mathbf{w}} = \mathbf{0}. \end{aligned} \tag{5.6.5}$$

Based on the indicator function $I_{\mathcal{C}}\left(\tilde{\mathbf{w}}\right)$ defined before, the above problem can be rewritten in the following form:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \tilde{\mathbf{w}}^T \tilde{\mathbf{A}} \tilde{\mathbf{w}} + \tilde{\mathbf{b}}^T \tilde{\mathbf{w}} + I_{\mathcal{C}}(\tilde{\mathbf{w}}) \\ \text{subject to} \quad & \left(\mathbf{B}^{\perp}\right)^T \tilde{\mathbf{w}} = \mathbf{0}. \end{aligned} \tag{5.6.6}$$

[4] A simple proof for this is given in Appendix 5.12.

And the augmented Lagrangian for (5.6.6) can be written as

$$\begin{aligned}&\mathcal{L}_\rho\left(\tilde{\mathbf{w}},\mathbf{u}\left(\mathbf{y}\right)\right)\\=&\tilde{\mathbf{w}}^T\tilde{\mathbf{A}}\tilde{\mathbf{w}}+\tilde{\mathbf{b}}^T\tilde{\mathbf{w}}+I_{\mathcal{C}}(\tilde{\mathbf{w}})+\mathbf{y}^T\left(\mathbf{B}^\perp\right)^T\tilde{\mathbf{w}}+\frac{\rho}{2}\left\|\left(\mathbf{B}^\perp\right)^T\tilde{\mathbf{w}}\right\|_2^2\\=&\tilde{\mathbf{w}}^T\tilde{\mathbf{A}}\tilde{\mathbf{w}}+\tilde{\mathbf{b}}^T\tilde{\mathbf{w}}+I_{\mathcal{C}}(\tilde{\mathbf{w}})+\frac{\rho}{2}\left\|\left(\mathbf{B}^\perp\right)^T\tilde{\mathbf{w}}+\mathbf{u}\right\|_2^2+const.,\end{aligned}$$

where $\rho>0$ is the penalty parameter and the scaled dual variable $\mathbf{u}=\frac{1}{\rho}\mathbf{y}$. Based on the augmented Lagrangian, it is easy to see that we only have two variable blocks $(\tilde{\mathbf{w}},\mathbf{u})$ for alternating minimization. Before we drive the variable update rule, we give the following useful result.

**Lemma 13.** *Let* $\mathbf{A}\in\mathbb{S}^K$ *and* $\mathbf{B}\in\mathbb{S}^K$ *such that* $\mathbf{B}\succeq\mathbf{A}$. *At any point* $\mathbf{x}^{(k)}\in\mathbb{R}^K$*, we have* $\mathbf{x}^T\mathbf{A}\mathbf{x}\geq\mathbf{x}^T\mathbf{B}\mathbf{x}+2\mathbf{x}^{(k)T}\left(\mathbf{A}-\mathbf{B}\right)\mathbf{x}+\mathbf{x}^{(k)T}\left(\mathbf{B}-\mathbf{A}\right)\mathbf{x}^{(k)}$.

Then for the $\tilde{\mathbf{w}}$ update, at the $(k+1)$th iteration with iterates $\left(\tilde{\mathbf{w}}^{(k)},\mathbf{u}^{(k)}(\mathbf{y}^{(k)})\right)$, by taking $\mathbf{M}^{\mathrm{M-ADMM}}\triangleq\tilde{\mathbf{A}}+\frac{\rho}{2}\mathbf{B}^\perp\left(\mathbf{B}^\perp\right)^T$ as $\mathbf{A}$ and choosing $\mathbf{B}=\lambda_{\max}^{\mathrm{M-ADMM}}\mathbf{I}$ where $\lambda_{\max}^{\mathrm{M-ADMM}}\triangleq\lambda_{\max}\left(\mathbf{M}^{\mathrm{M-ADMM}}\right)$ in Lemma 13, we get

$$\begin{aligned}&\mathcal{L}_\rho\left(\tilde{\mathbf{w}};\tilde{\mathbf{w}}^{(k)},\mathbf{u}^{(k)}\left(\mathbf{y}^{(k)}\right)\right)\\&=\tilde{\mathbf{w}}^T\mathbf{M}^{\mathrm{M-ADMM}}\tilde{\mathbf{w}}+\left(\tilde{\mathbf{b}}+\mathbf{B}^\perp\mathbf{y}^{(k)}\right)^T\tilde{\mathbf{w}}+I_{\mathcal{C}}(\tilde{\mathbf{w}})+const.\\&\leq\lambda_{\max}^{\mathrm{M-ADMM}}\tilde{\mathbf{w}}^T\tilde{\mathbf{w}}+2\tilde{\mathbf{w}}^{(k)T}\left(\mathbf{M}^{\mathrm{M-ADMM}}-\lambda_{\max}^{\mathrm{M-ADMM}}\mathbf{I}\right)\tilde{\mathbf{w}}\\&\quad+\left(\tilde{\mathbf{b}}+\mathbf{B}^\perp\mathbf{y}^{(k)}\right)^T\tilde{\mathbf{w}}+I_{\mathcal{C}}(\tilde{\mathbf{w}})+const.\\&=\lambda_{\max}^{\mathrm{M-ADMM}}\left\|\tilde{\mathbf{w}}-\mathbf{h}^{(k)}\right\|_2^2+I_{\mathcal{C}}(\tilde{\mathbf{w}})+const.,\end{aligned}$$

where $\mathbf{h}^{(k)}\triangleq-\left(\left(\lambda_{\max}^{\mathrm{M-ADMM}}\right)^{-1}\mathbf{M}^{\mathrm{M-ADMM}}-\mathbf{I}\right)\tilde{\mathbf{w}}^{(k)}-\frac{1}{2}\left(\lambda_{\max}^{\mathrm{M-ADMM}}\right)^{-1}\left(\tilde{\mathbf{b}}+\mathbf{B}^\perp\mathbf{y}^{(k)}\right)$. Then, the variable updates in M-ADMM are given as follows:

$$\begin{cases}\tilde{\mathbf{w}}^{(k+1)}=\arg\min\limits_{\tilde{\mathbf{w}}}\left\{\left\|\tilde{\mathbf{w}}-\mathbf{h}^{(k)}\right\|_2^2+I_{\mathcal{C}}\left(\tilde{\mathbf{w}}\right)\right\}=\Pi_{\mathcal{C}}(\mathbf{h}^{(k)})\\\mathbf{u}^{(k+1)}=\mathbf{u}^{(k)}+\left(\mathbf{B}^\perp\right)^T\tilde{\mathbf{w}}^{(k+1)}.\end{cases}$$

Specifically, for the variable $\tilde{\mathbf{w}}$ update, it is the projection onto the $\ell_1$-norm ball problem as in (5.6.4). In the M-ADMM algorithm, the number of variable blocks is reduced to 2 compared to the 3 variable blocks in the ADMM algorithm. In fact, when $\mathbf{B}=\mathbf{I}$, by leveraging on this specific structure, more efficient algorithm can be derived.

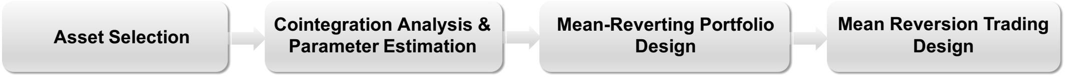

Asset Selection
Cointegration Analysis & Parameter Estimation
Mean-Reverting Portfolio Design
Mean Reversion Trading Design

Figure 5.3: A system view of the statistical arbitrage trading strategy in finance.

### 5.6.3 Specialized Algorithm Based on the MM Method

When $\mathbf{B} = \mathbf{I}$, the convex subproblem in (5.6.1) is written as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} \\ \text{subject to} & \|\mathbf{w}\|_1 \leq L. \end{array} \tag{5.6.7}$$

Besides using ADMM and M-ADMM, this problem can be more efficiently solved by the majorization-minimization (MM) method [19]. Using this primal-only method, we can get rid of the dual variable update in ADMM and M-ADMM.

From (5.6.7), based on Lemma 13, at the $(k+1)$th iteration with iterate $\mathbf{w}^{(k)}$, the objective function in (5.6.7) can be majorized as follows:

$$\begin{aligned} & \mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} \\ \leq & \lambda_{\max}(\mathbf{A})\,\mathbf{w}^T\mathbf{w} + 2\mathbf{w}^{(k)T}(\mathbf{A} - \lambda_{\max}(\mathbf{A})\,\mathbf{I})\,\mathbf{w} \\ & + \mathbf{b}^T\mathbf{w} + \mathbf{w}^{(k)T}(\lambda_{\max}(\mathbf{A})\,\mathbf{I} - \mathbf{A})\,\mathbf{w}^{(k)} \\ = & \lambda_{\max}(\mathbf{A})\left\|\mathbf{w} - \mathbf{h}^{(k)}\right\|_2^2 + const., \end{aligned}$$

where $\mathbf{h}^{(k)} \triangleq -\left(\lambda_{\max}^{-1}(\mathbf{A})\,\mathbf{A} - \mathbf{I}\right)\mathbf{w}^{(k)} - \frac{1}{2}\lambda_{\max}^{-1}(\mathbf{A})\,\mathbf{b}$. Then, the subproblem to solve in MM is given by

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \left\|\mathbf{w} - \mathbf{h}^{(k)}\right\|_2^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \leq L, \end{array}$$

which is still a projection onto the $\ell_1$-norm ball problem and can be solved based on Algorithm 5.2.

## 5.7 Complexity and Convergence Analysis

### 5.7.1 Complexity Analysis

In this section, we give a detailed discussion on the computational complexity of our proposed algorithms in Section 5.6. We analyze the per-iteration computational cost of the algorithms proposed to solve the inner convex subproblems, i.e., the ADMM-based algorithm, the M-ADMM-based algorithm, and the MM-based algorithm.

For the ADMM-based algorithm, the computational cost for updating three variable blocks $\mathbf{w}$, $\mathbf{z}$, and $\mathbf{u}$ are analyzed separately. The computational cost for updating $\mathbf{w}$ is $\mathcal{O}(N^3+MN+3N^2+M+2N)$. For updating $\mathbf{z}$, the cost is $\mathcal{O}(MN+M)$ (to calculate $\mathbf{h}^{(k)}$) plus $\mathcal{O}(M)$ (to do projection). The cost for updating $\mathbf{u}$ is $\mathcal{O}(MN+2M)$. So, the total cost per iteration ($M \geq N$) is $\mathcal{O}(N^3+MN+3N^2+M+2N)+\mathcal{O}(MN+M)+\mathcal{O}(M)+\mathcal{O}(MN+2M) \approx \mathcal{O}(N^3)$.

In the M-ADMM-based algorithm, for pre-processing, computing the Moore-Penrose pseudo-inverse the $\mathbf{B}^{\dagger}$ requires complexity of $\mathcal{O}(MN^2)$ and computing the orthogonal compliment $\mathbf{B}^{\perp}$ needs complexity of $\mathcal{O}(MN^2)$. So computing $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ requires complexity of $\mathcal{O}(NM^2+MN^2)$ and $\mathcal{O}(MN)$, separately. To recover the variable $\mathbf{w}$, i.e., post-processing, still need $\mathcal{O}(MN)$ time. Therefore the cost for computation outside the iterations is $\mathcal{O}(NM^2+MN^2)+\mathcal{O}(MN)+\mathcal{O}(MN) \approx \mathcal{O}(NM^2)$. To compute $\mathbf{h}^{(k)}$ in each iteration, it contains costs of $\mathcal{O}(2M^2)$ and $\mathcal{O}(M^3)$ to calculate a $M \times M$ matrix (i.e., $\tilde{\mathbf{A}}+\frac{\rho}{2}\left(\mathbf{B}^{\perp}\right)^T\mathbf{B}^{\perp}$) and its largest eigenvalue. However, the $\mathcal{O}(M^3)$ complexity can be reduced by simply replacing the largest eigenvalue with some easily computed quantity (like the Frobenius norm since $\|\mathbf{A}\|_F \geq \lambda\left(\mathbf{A}\right)_{\max}$) of that matrix, which only requires cost $\mathcal{O}(2M^2)$. Therefore, the overall cost for calculating $\mathbf{h}^{(k)}$ is $\mathcal{O}(6M^2+M^2+4M-MN)$ and the cost for updating $\tilde{\mathbf{w}}$ is $\mathcal{O}(M)$. Besides, it requires $\mathcal{O}(M^2+M-MN-N)$ to update $\mathbf{u}$. Then the overall cost for each iteration is $\mathcal{O}(6M^2+M^2+4M-MN)+\mathcal{O}(M)+\mathcal{O}(M^2+M-MN-N) \approx \mathcal{O}(M^2)$.

The MM-based algorithm is proposed for the $\mathbf{B}=\mathbf{I}$ case. It also needs pretreatment, which costs $\mathcal{O}(N^3)$ to calculate the maximum eigenvalue of a $N \times N$ matrix and $\mathcal{O}(N^2+N)$ to calculate the constant part of $\mathbf{h}^{(k)}$. The overall computation cost is of order $\mathcal{O}(N^3)$. In each iteration, to update $\mathbf{w}$, $\mathcal{O}(N^2+N)$ is needed for constructing the majorization function and $\mathcal{O}(N)$ for projection onto $l_1$-norm ball. The total cost per iteration is of order $\mathcal{O}(N^2)$.

The three algorithms for solving the subproblem should be properly chosen in order to achieve a better computational performance. The per-iteration computational cost for the ADMM-based and M-ADMM-based algorithms are $\mathcal{O}(N^3)$ and $\mathcal{O}(M^2)$, respectively. So, under the condition $M \geq N^{1.5}$, ADMM is recommended; otherwise, M-ADMM should be more appropriate. Compared to the $\mathcal{O}(N^3)$ complexity in ADMM and the $\mathcal{O}(M^2)$ complexity in M-ADMM for each iteration, MM-based algorithm just need $\mathcal{O}(N^2)$ computation per iteration. The time complexity of the MM-based algorithm is also lower in the pre-processing stage compared with M-ADMM. So the MM-based algorithm is highly recommended when $\mathbf{B} = \mathbf{I}$.

### 5.7.2 Convergence Analysis

The convergence property for the SCA-MRP algorithm is given in the following. Under assumptions A1)-A3) and B1)-B3), suppose $\tau \geq 0$, $\gamma^{(k)} \in (0, 1]$, $\gamma^{(k)} \to 0$ and $\sum_k \gamma^{(k)} = +\infty$, and let $\left\{\mathbf{w}^{(k)}\right\}$ be the sequence generated by SCA-MRP. Then either SCA-MRP converges in a finite number of iterations to a stationary solution of Problem (5.3.1) or every limit of sequence $\left\{\mathbf{w}^{(k)}\right\}$ (at least one such point exists) is a stationary solution of Problem (5.3.1). We can first check that the proposed problem satisfies Assumptions A1)-A3) in Section 5.4. Given $\tau \geq 0$ and $\gamma^{(k)}$ as above, it is easy to check that the approximation function (5.5.11) is a strongly convex quadratic function and satisfies Assumptions B1)-B3) in Section 5.4. Then this result directly follows from the proof in [106, Theorem 2].

## 5.8 Numerical Simulations

In this section, we first give a system view of the statistical arbitrage strategy. Then several performance evaluation metrics on portfolio investment will be introduced. The performance of our proposed MRP design problem and the algorithms will finally be given based on both synthetic data and real market data.

### 5.8.1 A Flow Diagram of The Statistical Arbitrage Strategy

We summarize the whole statistical arbitrage strategy as shown in Figure 5.3.

**Asset Selection** In this stage, a collection of (possibly cointegrated) asset candidates are selected to construct an asset pool. Conducting this process may require prior knowledge on the underlying financial assets.

**Parameter Estimation and Cointegration Analysis** The cointegration analysis (say, Engle-Granger two-step test, Johansen test, Phillips-Ouliaris test, etc.) will be conducted to test the hypothesis that there is a statistically significant stationarity connection within the underlying asset prices. Accordingly, a cointegration space will be identified in this stage.

**Mean-Reverting Portfolio Design** This stage is the focus of this chapter. An optimal MRP is designed considering different criteria based on the assets within the identified cointegration space. Unit root test may be applied to test the stationarity of the finally designed spread.

**Mean Reversion Trading Design** The designed spread will be firstly traded for an in-sample testing period for parameter estimation and trading actions optimization, such as the mean reversion equilibrium, trading threshold, timing of entering a position, lightening up a position, adding to a position, or exiting a position, and so on. After these trading parameters are obtained, the designed MRP can finally be invested for the out-of-sample trading.

### 5.8.2 Performance Evaluation Metrics

Some performance metrics for mean reversion trading used in [13] are briefly introduced in the following.

**Profit and Loss** We define the profit and loss (P&L) for the MRP at time $t$ as $\mathrm{P\&L}_t \triangleq \mathbf{w}_p^T \mathbf{r}_t$ where the asset returns $\mathbf{r}_t \triangleq \mathbf{y}_t - \mathbf{y}_{t-1}$ (please refer to [13] for further details). P&L measures the amount of profits or losses (in units of dollars) of an investment on the portfolio for one holding period. In order to measure the cumulative return performance, we define the cumulative P&L (not compounding) in one trading from time $t_1$ to $t_2$ as Cum. $\mathrm{P\&L}\,(t_1, t_2) \triangleq \sum_{t=t_1}^{t_2} \mathrm{P\&L}_t$.

**Return On Investment** Different MRPs may have different leverage properties, the return on investment (ROI) is introduced as another measure as the rate of return. Within one trading period, the ROI at time $t$ is defined as $\mathrm{ROI}_t \triangleq \frac{\mathrm{P\&L}_t}{\|\mathbf{w}_p\|_1}$.

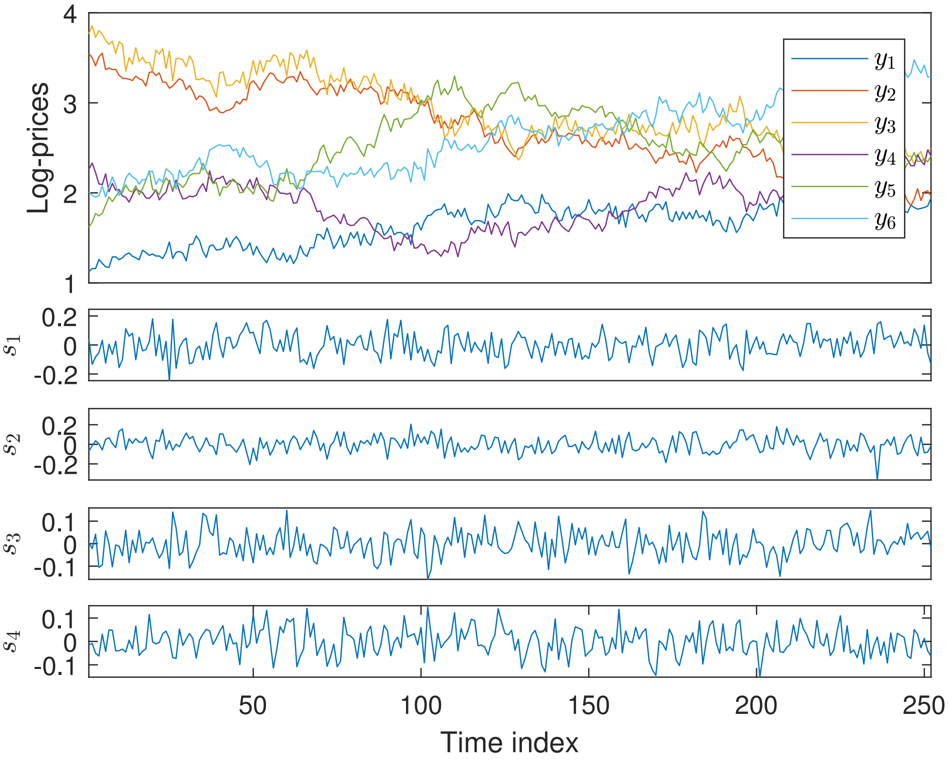


Figure 5.4: Synthetic log-prices ($M = 6$) and the spreads ($N = 4$) generated from a VECM model of order $1$.

**Sharpe Ratio** The Sharpe ratio (SR) describes how much excess return one can receive for the extra volatility (square root of variance). The annualized SR for a trading stage from time $t_1$ to $t_2$ is defined as $\mathrm{SR}_{\mathrm{ROI}}(t_1, t_2) \triangleq \sqrt{252}\frac{\mu_{\mathrm{ROI}}}{\sigma_{\mathrm{ROI}}}$, where $\mu_{\mathrm{ROI}} \triangleq \frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2} \mathrm{ROI}_t$ is the sample return and $\sigma_{\mathrm{ROI}} \triangleq \left[\frac{1}{t_2-t_1}\sum_{t=t_1}^{t_2} \left(\mathrm{ROI}_t - \mu_{\mathrm{ROI}}\right)^2\right]^{\frac{1}{2}}$ is the sample standard deviation, and the factor $\sqrt{252}$ relates the daily $\mathrm{SR}$ to the annualized $\mathrm{SR}$ (assuming $252$ trading days per year).

### 5.8.3 Synthetic Data Simulations

In this section, we will first show the superiority of the proposed algorithm SCA-MRP over some off-the-shelf solvers based on synthetic data. Following that, we will show the MRP design problem proposed in this chapter is able to design a portfolio attaining a trade-off between MR and variance, which is a practical and desirable property for MRP design, but has never been considered in the literature. The synthetic data is generated using a vector error correction model (VECM) [26], which models the stock log-prices with underlying cointegration relations as shown in Figure 5.4.

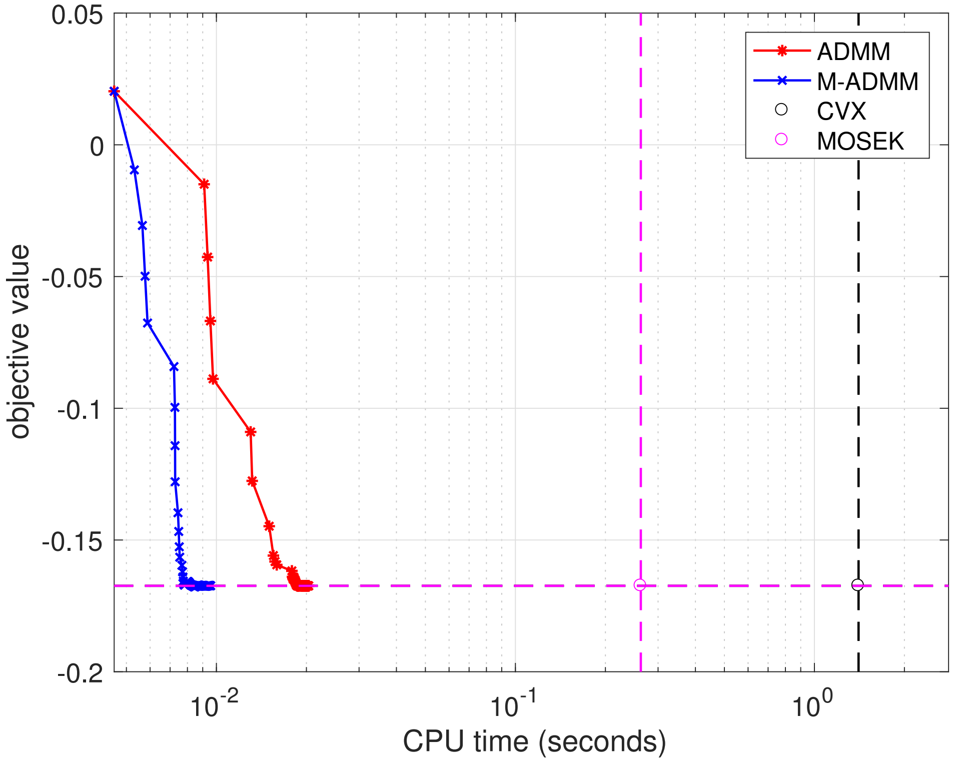


Figure 5.5: Convergence of the objective function value of different solving methods for the inner convex problem: Case 1 ($M = 6$ and $N = 4$).

#### 5.8.3.1 Algorithm Performance

We first compare our proposed algorithms for the inner convex problems in SCA-MRP, i.e., the ADMM method, the M-ADMM method, and the MM method. The proposed methods are first compared with the standard off-the-shelf packages CVX and MOSEK in Figures 5.5 and 5.6. Based on our simulations, the M-ADMM and ADMM algorithms can converge to the optimal solution orders of magnitude faster compared to CVX and MOSEK. In Case 1, M-ADMM method outperforms ADMM method. And in Case 2, where $\mathbf{B} = \mathbf{I}$, the MM method achieves the best convergence performance in terms of runtime in all the tested algorithms as expected. These convergence results match the complexity analyses given in Section 5.7.

We now compare the solution of the original problem based on SCA-MRP algorithm with the standard solver fmincon in MATLAB Optimization Toolbox for the MRP design problem where the MR criterion and the variance criterion are chosen as the portmanteau statistics of order 3, i.e., $\mathrm{pro}\left(\mathbf{w}, 3\right)$ and $\mathrm{VarInv}\left(\mathbf{w}\right)$, respectively. For the SCA-MRP algorithm, the inner convex problem is solved by different proposed algorithms. In the simulations, we use $\alpha = 10^{-5}$ and $\beta = 0.8$ in choosing the stepsize. From Figure 5.7, it is easy to see that the SCA-MRP algorithms can converge to better local optimal solutions with faster convergence

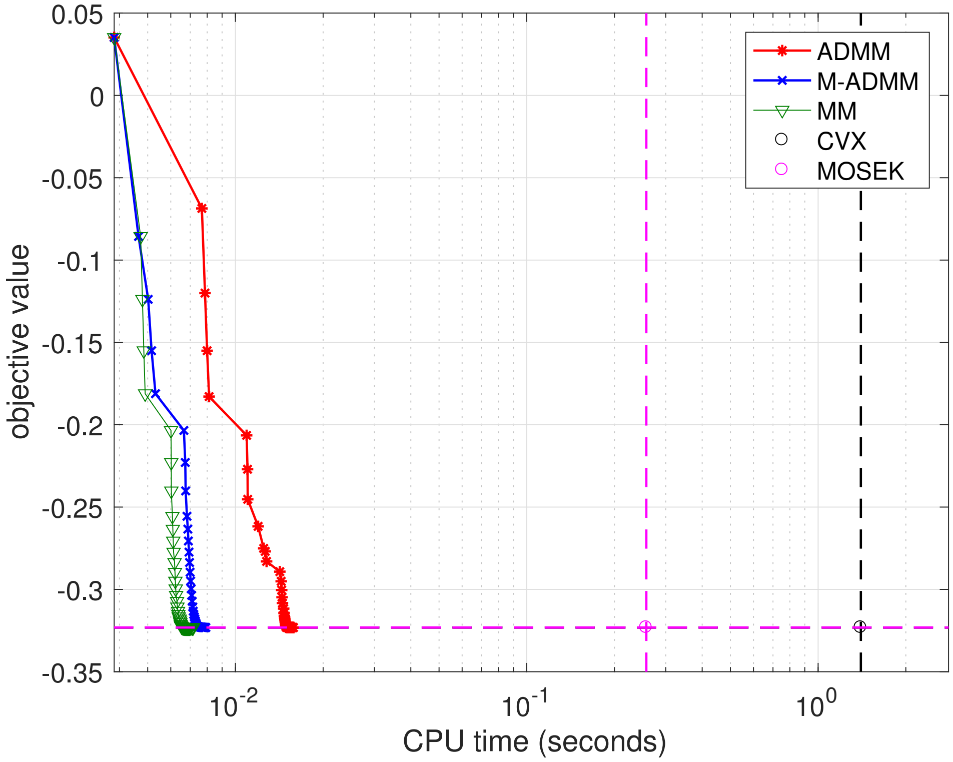


Figure 5.6: Convergence of the objective function value of different solving methods for the inner convex problem: Case 2 ($M = 10$ and $N = 10$).

speed compared to fmincon which is a general-purpose solver. Within all the SCA-MRP algorithms, the algorithms with inner problem solved by M-ADMM and ADMM show better convergence performance over those using CVX and MOSEK.

#### 5.8.3.2 Formulation Property

In this section, we will show that our proposed MRP design problem in (5.3.1) is more practical and flexible. We compare the design problem model in this chapter with the existing problem formulations in [13], [68], [69], [99]. Given a fixed portfolio variance and a fixed $\ell_1$-norm as in [68], [69] or a fixed portfolio budget $B$ as in [13], [99], we compute the MRP $\mathbf{w}$ (denoted as "MRP with $\ell_2$-norm" in Figure 5.8 and "MRP with budget" in Figure 5.9). Since, in real markets, the investment is always guided by the leverage which essentially tells the total amount of money people can employ, we accordingly compute the investment leverage $L = \|\mathbf{Bw}\|_1$ in these cases. Based on this leverage $L$, we use the newly proposed MRP design problem in (5.3.1) to design a series of MRPs (denoted as "MRP with leverage" in Figures 5.8 and 5.9) where the MR criterion and the variance criterion are chosen as $\mathrm{por}(\mathbf{w}, 3)$ and $\mathrm{VarInv}(\mathbf{w})$, respectively. We first design the portfolio with the minimal

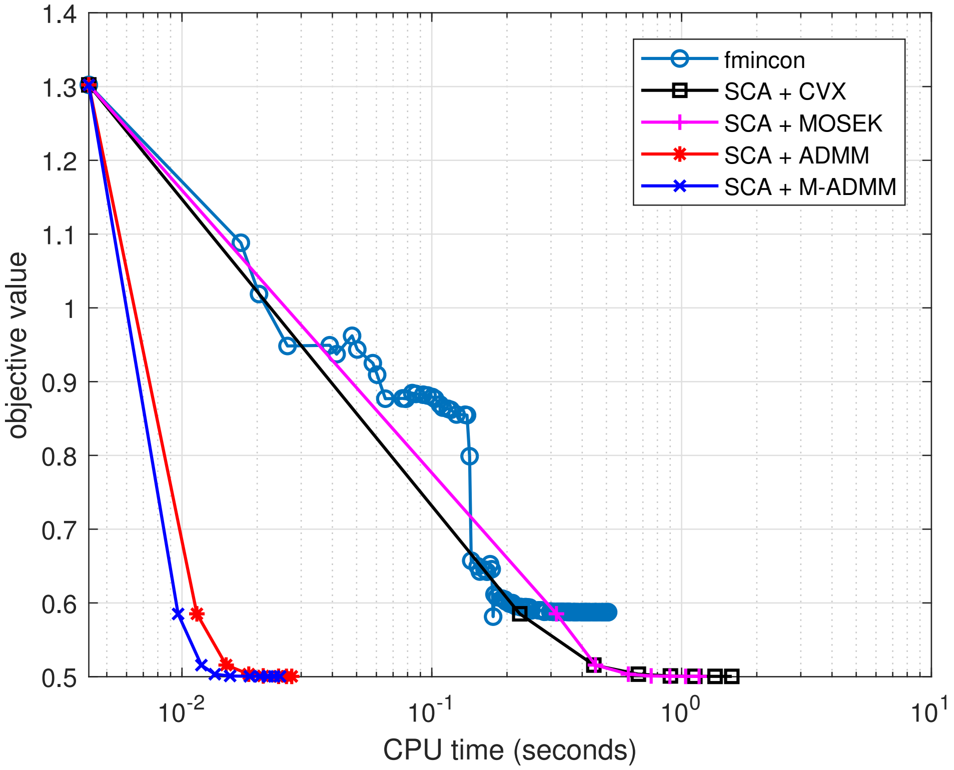


Figure 5.7: Convergence of the objective function value for different solving methods for $\mathrm{pro}\left(\mathbf{w},3\right)$ and $L=1.3$.

MR denoted as $\mathbf{w}^{\star}_{\mathrm{min\_MR}}$ (corresponding to the case when $\mu\to 0$ in Problem (5.3.1)) and the portfolio with the maximal variance denoted as $\mathbf{w}^{\star}_{\mathrm{max\_Var}}$ (corresponding to the case when $\mu\to\infty$ in Problem (5.3.1)). We also plot the path of the designed MRPs by tuning the parameter $\mu$. In both Figure 5.8 and Figure 5.9, it can be found that by tuning parameter $\mu$, for a fixed leverage the newly proposed design problem can easily get a trade-off between MR and variance of the MRP. However, although under the same investment leverage $L$, the MRP designed from [13], [68], [69], [99] is suboptimal no mater considering its MR property or variance property.

#### 5.8.3.3 Trading Performance

We test the performance our designed MRP through mean reversion trading. The performance of the designed portfolio based on the proposed problem formulation is compared with spread $s_1$. Some performance metrics are reported in Figures 5.10. From the simulations, we can conclude that the MRPs designed based on the proposed formulation is able to generate consistent positive profits and can outperform the underlying spreads with higher Sharpe ratios of ROIs and higher cumulative P&Ls.

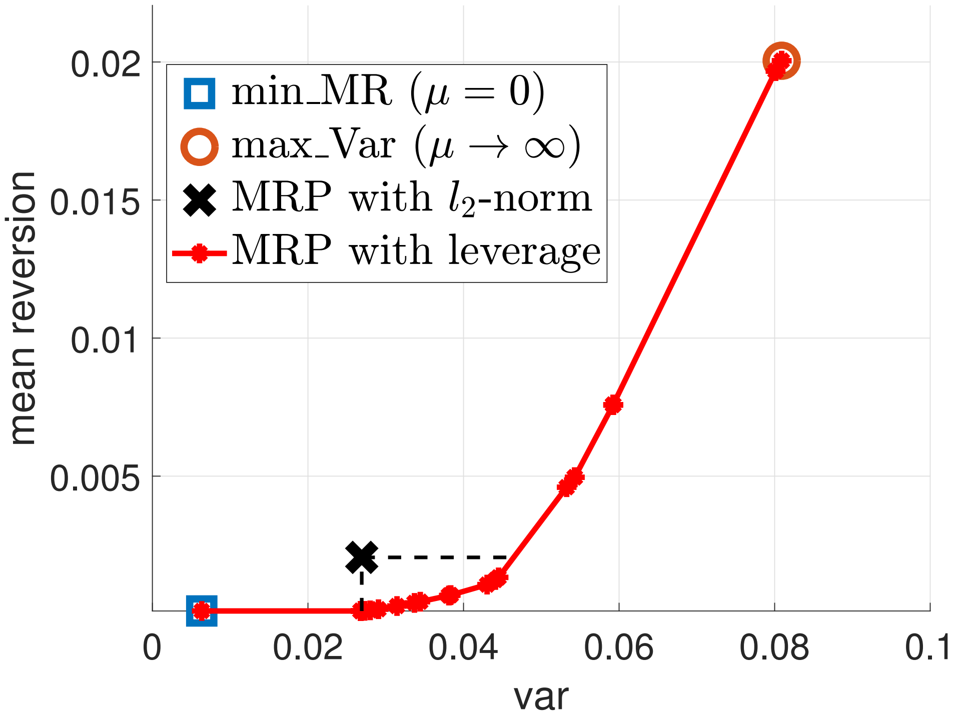


Figure 5.8: Comparison between the proposed methods and the methods in [68], [69] (Each point is averaged based on 100 Monte Carlo simulations with random initializations.).

### 5.8.4 Real Data Simulations

In this section, we test the proposed problem formulation and algorithms based on real market data. We first select stocks from the Standard & Poor's 500 (S&P 500) Index to construct an asset pool, which are denoted by their ticker labels as {APA, AXP, CAT, COF, FCX, IBM, MMM}. The data is retrieved from Google Finance (https://www.google.com/finance). Then, a VECM model is fitted to identify the cointegration space $\mathcal{R}(\mathbf{B})$. After that, the MRP design problem proposed in this chapter is used to design the optimal MRP where the MR is chosen as $\mathrm{pre}(\mathbf{w})$ and the variance criterion is chosen as $\mathrm{VarInv}(\mathbf{w})$. In Figure 5.11, we show the stock log-prices and spreads constructed from our asset pool. In Figures 5.12, 5.13, and 5.14, we show the performance comparisons between our designed MRP and one underlying spread $s_2$ and the MRPs from the literature [13], [68], [69], [99]. The log-prices for the designed spread, and the out-of-sample performance like ROI, Sharpe ratios of ROI, and cumulative P&Ls are reported. The in-sample training (learning) period is chosen from February 1st, 2010 to March 4th, 2013, and the out-of-sample trading (investing) period is from March 5th, 2013 to June 27th, 2014. It is easy to see the designed optimal MRP can achieve a higher Sharpe ratio and a better final cumulative return performance.

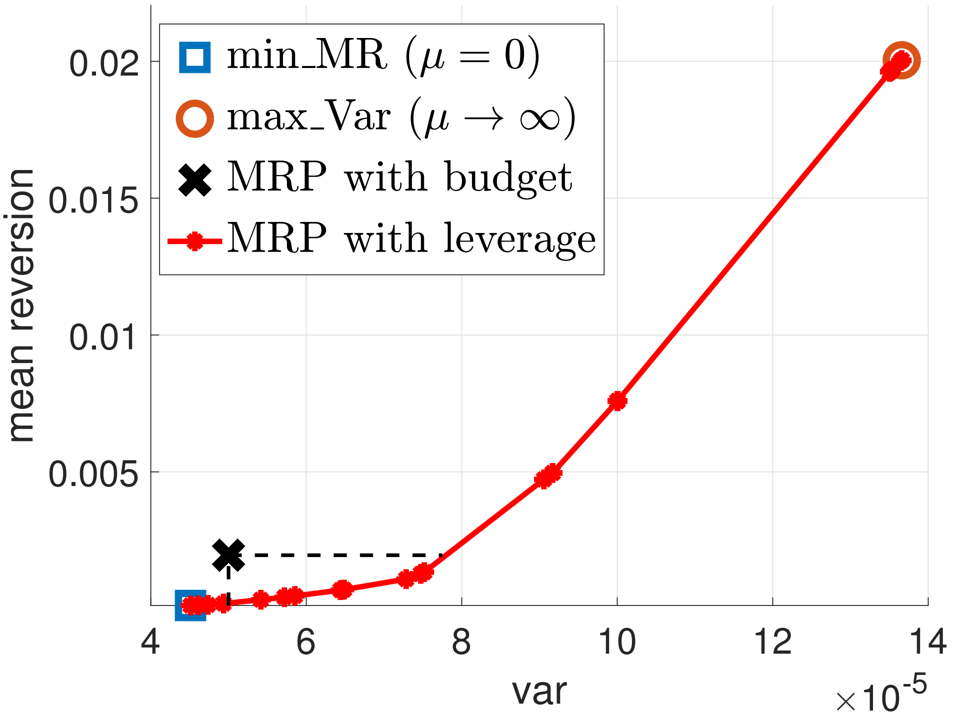


Figure 5.9: Comparison between the proposed methods and the methods in [13], [99] (Each point is averaged based on 100 Monte Carlo simulations with random initializations.).

## 5.9 Chapter Summary and Conclusions

The optimal mean-reverting portfolio design problem arising from statistical arbitrage has been considered in this chapter. We have proposed a general problem formulation for MRP design where a trade-off can be attained between the mean reversion and variance of an MRP. Asset selection criterion has been further considered in the problem formulation. A practical investment leverage constraint has been imposed for MRP design. To solve the problem, a unified SCA-based algorithm has been proposed with the inner subproblems efficiently solved by different algorithms. Numerical results have shown that our proposed problem formulation can generate consistent profits and outperform the benchmark methods.

## 5.10 Proof for Proposition 4

Given $\mathbf{B} \in \mathcal{R}(\mathbf{U})$, we assume the optimal MRP from Problem (5.3.1) is given by $\mathbf{w}_p^{\star} = \mathbf{B}\mathbf{w}^{\star} \in \mathcal{W}_p^{\star}$ with $\mathbf{w}^{\star} \in \mathcal{W}^{\star}$. Accordingly, for another $\mathbf{B}' \in \mathcal{R}(\mathbf{U})$, we have $\mathbf{w}_p'^{\star} = \mathbf{B}'\mathbf{w}'^{\star} \in \mathcal{W}_p'^{\star}$ with $\mathbf{w}'^{\star} \in \mathcal{W}'^{\star}$.

Since $\mathbf{B}, \mathbf{B}' \in \mathcal{R}(\mathbf{U})$, there always exists $\mathbf{Q} \succ \mathbf{0}$ such that $\mathbf{B}' = \mathbf{B}\mathbf{Q}$. Also notice that the estimation of parameters in $U(\mathbf{w})$ and $V(\mathbf{w})$ depend on $\mathbf{B}$. Substitute $\mathbf{B}' = \mathbf{B}\mathbf{Q}$ into

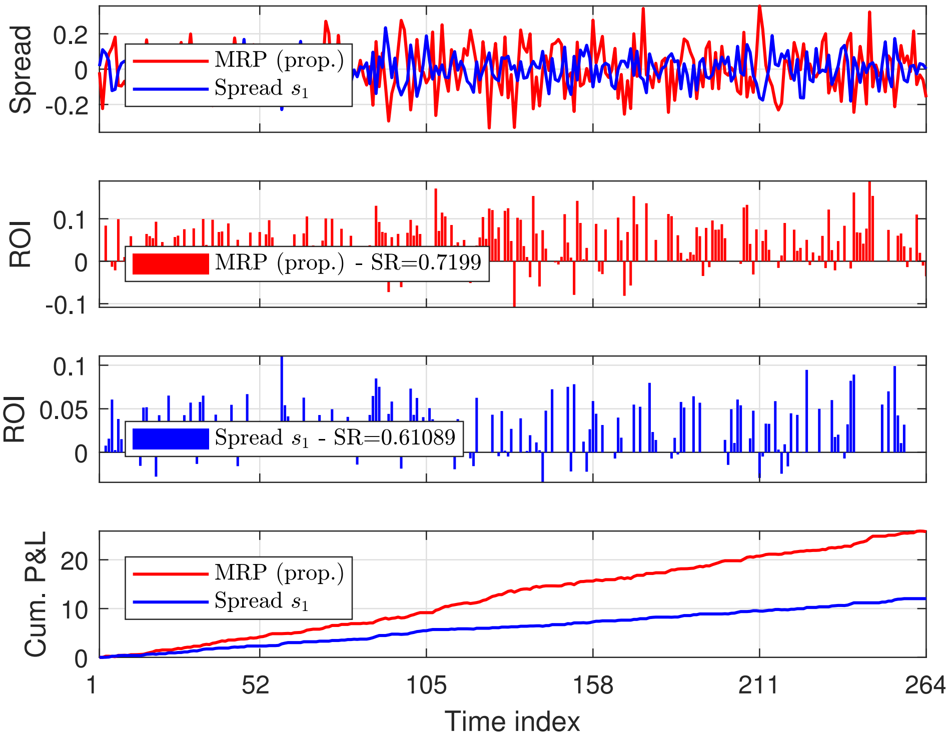


Figure 5.10: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP (prop.) and one underlying spread denoted as Spread $s_1$.

Problem (5.3.1) with variable $\mathbf{w}'$. Defining $\bar{\mathbf{w}} = \mathbf{Q}\mathbf{w}'$ with the optimal set $\bar{\mathcal{W}}^{\star}$, it is easy to see $\bar{\mathcal{W}}^{\star} = \mathcal{W}^{\star}$. Accordingly, we get $\forall \mathbf{w}'^{\star} \in \mathcal{W}'^{\star}$, $\mathbf{w}'^{\star} = \mathbf{Q}^{-1}\bar{\mathbf{w}}^{\star} = \mathbf{Q}^{-1}\mathbf{w}^{\star}$ with $\mathbf{w}^{\star} \in \mathcal{W}^{\star}$. Then we have $\forall \mathbf{w}_p'^{\star} \in \mathcal{W}_p'^{\star}$,

$$\mathbf{w}_p'^{\star} = \mathbf{B}'\mathbf{w}'^{\star} = (\mathbf{B}\mathbf{Q})\left(\mathbf{Q}^{-1}\mathbf{w}^{\star}\right) = \mathbf{B}\mathbf{w}^{\star} = \mathbf{w}_p^{\star},$$

which implies $\mathcal{W}_p'^{\star} = \mathcal{W}_p^{\star}$.

## 5.11 On The Derivation of $\tilde{u}_2^{(k)}(\mathbf{w})$

Given $u(\mathbf{w})$ in (5.5.5), we define the numerator quadratic function in $(\cdot)^2$ as $t \triangleq \mathbf{w}^T\mathbf{M}_i\mathbf{w}$. Then, with a little abuse of notation, we have

$$u(t, \mathbf{w}) = \left(\frac{t}{\mathbf{w}^T\mathbf{M}_0\mathbf{w} + \epsilon}\right)^2.$$

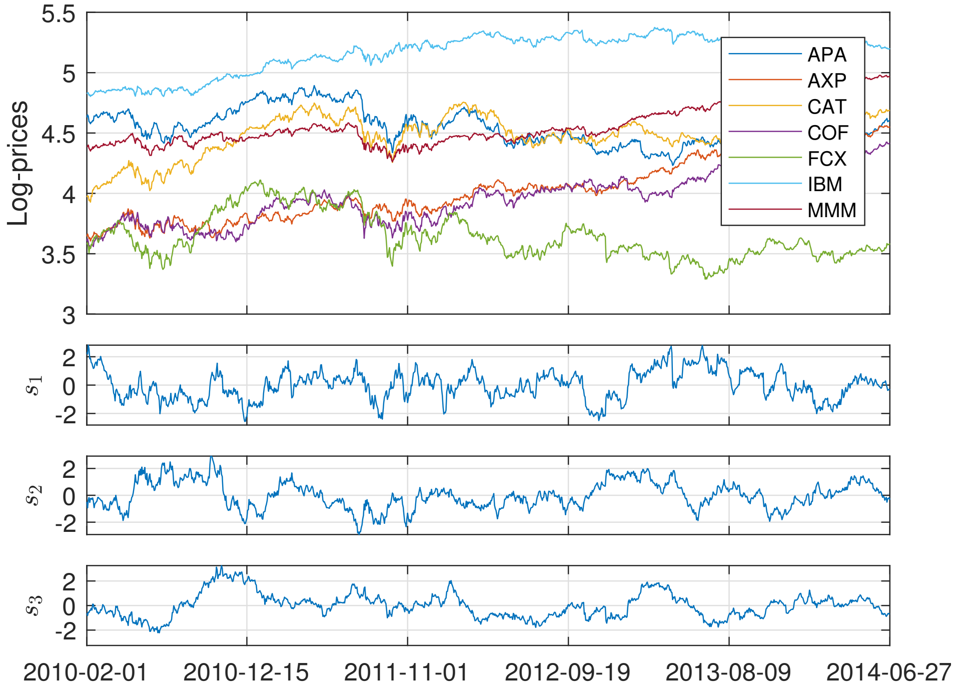


Figure 5.11: Log-prices for $\{\mathsf{APA}, \mathsf{AXP}, \mathsf{CAT}, \mathsf{COF}, \mathsf{FCX}, \mathsf{IBM}, \mathsf{MMM}\}$ and three estimated spreads $s_1$ , $s_2$, and $s_3$.

A linear approximation function for $u\left(t, \mathbf{w}\right)$ at $\left(t^{(k)}, \mathbf{w}^{(k)}\right)$ is given as follows:

$$\begin{aligned}\widetilde{u}_2^{(k)}\left(t, \mathbf{w}\right) = &\left(\frac{t^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2\\ &+2\left(\frac{1}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2 t^{(k)}(t-t^{(k)})\\ &-4(t^{(k)})^2\left(\frac{1}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^3\mathbf{w}^{(k)T}\mathbf{M}_0(\mathbf{w}-\mathbf{w}^{(k)}).\end{aligned}$$

Changing the variables back to $\mathbf{w}$ (i.e., $t = \mathbf{w}^T\mathbf{M}_i\mathbf{w}$ and $t^{(k)} = \mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}$), we have

$$\begin{aligned}\widetilde{u}_2^{(k)}\left(\mathbf{w}\right) = &\left(\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2+\frac{1}{\left(\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon\right)^2}\\ &\times 2\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}(\mathbf{w}^T\mathbf{M}_i\mathbf{w}-\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)})\\ &-4\left(\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2\frac{\mathbf{w}^{(k)T}\mathbf{M}_0(\mathbf{w}-\mathbf{w}^{(k)})}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\\ =&\,4\left(\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^3-\left(\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2\\ &-4\left(\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\right)^2\frac{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\\ &+2\frac{\mathbf{w}^{(k)T}\mathbf{M}_i\mathbf{w}^{(k)}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}.\end{aligned}$$

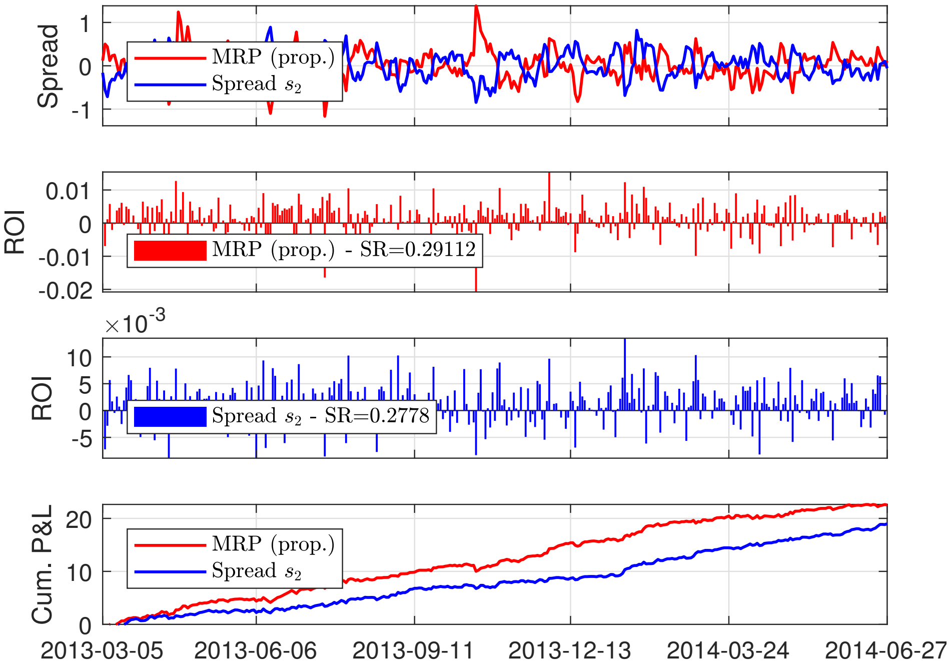


Figure 5.12: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP (prop.) and one underlying spread denoted as Spread $s_2$.

Based on the definitions in (5.5.4), we further have

$$\begin{aligned}\widetilde{u}_2^{(k)}\left(\mathbf{w}\right) =& 4\left(r_i^{(k)}\right)^3 - \left(r_i^{(k)}\right)^2 - 4r_i^{(k)}\left(\mathbf{d}_{i,0}^{(k)}\right)^T\mathbf{w} \\ &+ 2r_i^{(k)}\frac{\mathbf{w}^T\mathbf{M}_i\mathbf{w}}{\mathbf{w}^{(k)T}\mathbf{M}_0\mathbf{w}^{(k)}+\epsilon}.\end{aligned}$$

## 5.12 Proof for The Variable Transformation

Since $\tilde{\mathbf{w}} = \mathbf{B}\mathbf{w}$ (where $\mathbf{B} \in \mathbb{R}^{M\times N}$ with $M \geq N$), we have

$$\tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{B} & \mathbf{B}^{\perp} \end{bmatrix}\begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix},$$

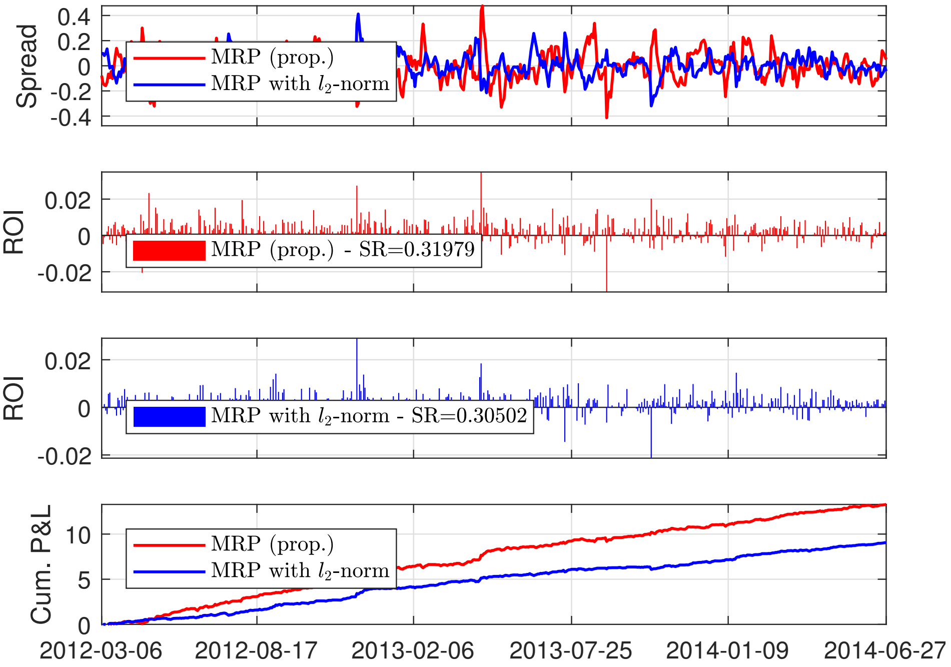


Figure 5.13: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP (prop.) and the MRP design from [68], [69] denoted as MRP with $\ell_2$-norm.

where the columns of $\mathbf{B}^{\perp}$ span the orthogonal complementary of $\mathbf{B}$. Multiplying both sides of the above equation by $\begin{bmatrix} \mathbf{B} & \mathbf{B}^{\perp} \end{bmatrix}^T$, we get

$$\begin{aligned} \begin{bmatrix} \mathbf{B}^T \\ \left(\mathbf{B}^{\perp}\right)^T \end{bmatrix} \tilde{\mathbf{w}} &= \begin{bmatrix} \mathbf{B}^T \\ \left(\mathbf{B}^{\perp}\right)^T \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{B}^{\perp} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{B}^T\mathbf{B} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{B}^{\perp}\right)^T \mathbf{B}^{\perp} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix}, \end{aligned}$$

and then we have

$$\begin{bmatrix} \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T \\ \left(\left(\mathbf{B}^{\perp}\right)^T \mathbf{B}^{\perp}\right)^{-1} \left(\mathbf{B}^{\perp}\right)^T \end{bmatrix} \tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix}.$$

Notice that $\left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T$ is the Moore-Penrose pseudo-inverse of $\mathbf{B}$ which can be written as $\mathbf{B}^{\dagger}$. We get the following equivalence relation

$$\tilde{\mathbf{w}} = \mathbf{B}\mathbf{w} \iff \begin{cases} \mathbf{B}^{\dagger}\tilde{\mathbf{w}} = \mathbf{w} \\ \left(\mathbf{B}^{\perp}\right)^T \tilde{\mathbf{w}} = \mathbf{0} \end{cases}.$$

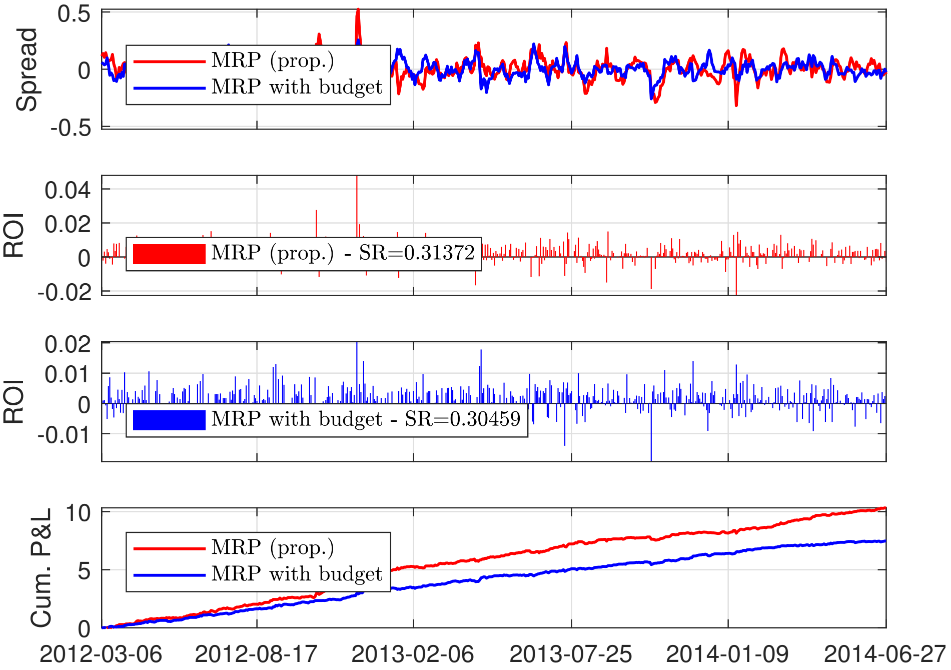


Figure 5.14: Comparisons of ROIs, Sharpe ratios, and cumulative P&Ls between the MRP designed using our proposed method denoted as MRP (prop.) and the MRP design from [13], [99] denoted as MRP with budget.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

Nowadays, financial engineering has been an active research area which receives extensive attention and interest and statistical arbitrage as a risk neutral strategy becomes more and more popular in the financial industry. Statistical arbitrage, also known as pairs trading, is usually involved in a series of stages, say, assets selection, model estimation, portfolio design, and mean reversion trading. In this thesis, we have focused the model estimation and portfolio design parts.

After the introduction given in Chapter 1, we have explored each work one by one and the main results are summarized as follows.

- Chapter 2 considers the efficient estimation of sparse reduced-rank regression model via nonconvex optimization.
- Chapter 3 considers the robust maximum likelihood estimation of vector error correction model problem.
- Chapter 4 studies the mean-reverting portfolio design problem with a budget constraint.
- Chapter 5 studies the mean-reverting portfolio design problem with a leverage constraint.

To summarize, in this dissertation we have applied the signal processing and machine learning methods for several problems related to statistical arbitrage in finance.

## 6.2 Future Work

For each work presented in this thesis, there are some interesting future extensions.

Regarding to the content of RRR and VECM estimation problems, one promising problem is

- to improve the estimation of the model parameters by taking into account the prior covariance structures;

As to the mean-reverting portfolio design, one interesting problem is

- to find more desirable optimization criteria for the design of portfolios.

Through the demonstration of this dissertation, we have focused on using optimization techniques to solve several statistical arbitrage problems in finance. In fact, we can also consider developing an end-to-end approach to conduct the statistical arbitrage strategy, and as to research in this direction the deep learning approach may be an interesting idea to explore in the future.

# Bibliography

[1] T. W. Anderson, “Estimating linear restrictions on regression coefficients for multivariate normal distributions,” *The annals of mathematical statistics*, pp. 327–351, 1951.

[2] T. W. Anderson, Ed., *An introduction to multivariate statistical analysis*. Wiley, 1984.

[3] A. J. Izenman, “Reduced-rank regression for the multivariate linear model,” *Journal of multivariate analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[4] M. Viberg, P. Stoica, and B. Ottersten, “Maximum likelihood array processing in spatially correlated noise fields using parameterized signals,” *Ieee transactions on signal processing*, vol. 45, no. 4, pp. 996–1004, 1997.

[5] P. Stoica and M. Jansson, “MIMO system identification: State-space and subspace approximations versus transfer function and instrumental variables,” *Ieee transactions on signal processing*, vol. 48, no. 11, pp. 3087–3099, 2000.

[6] J. H. Manton and Y. Hua, “Convolutive reduced rank wiener filtering,” in *Proc. the 2001 ieee international conference on acoustics, speech, and signal processing (icassp’01)*, IEEE, vol. 6, 2001, pp. 4001–4004.

[7] E. Lindskog and C. Tidestav, “Reduced rank channel estimation,” in *Proc. 1999 ieee 49th vehicular technology conference,*, IEEE, vol. 2, 1999, pp. 1126–1130.

[8] Y. Hua, M. Nikpour, and P. Stoica, “Optimal reduced-rank estimation and filtering,” *Ieee transactions on signal processing*, vol. 49, no. 3, pp. 457–469, 2001.

[9] M. Nicoli and U. Spagnolini, “Reduced-rank channel estimation for time-slotted mobile communication systems,” *Ieee transactions on signal processing*, vol. 53, no. 3, pp. 926–944, 2005.

[10] G. Zhou, “Small sample rank tests with applications to asset pricing,” *Journal of empirical finance*, vol. 2, no. 1, pp. 71–93, 1995.

[11] P. Bekker, P. Dobbelstein, and T. Wansbeek, “The APT model as reduced-rank regression,” *Journal of business & economic statistics*, vol. 14, no. 2, pp. 199–202, 1996.

[12] Z. Zhao and D. P. Palomar, “Robust maximum likelihood estimation of sparse vector error correction model,” in *Proc. the 2017 5th ieee global conference on signal and information processing*, Montreal, C, Canada, Nov. 2017, pp. 913–917. DOI: `10.1109/GlobalSIP.2017.8309093`.

[13] Z. Zhao and D. P. Palomar, “Mean-reverting portfolio with budget constraint,” *Ieee trans. signal processing*, vol. 66, no. 9, pp. 2342–2357, Jan. 2018.

[14] R. Velu and G. C. Reinsel, *Multivariate reduced-rank regression: Theory and applications*. Springer Science & Business Media, 2013, vol. 136.

[15] L. Chen and J. Z. Huang, “Sparse reduced-rank regression for simultaneous dimension reduction and variable selection,” *Journal of the american statistical association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[16] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[17] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[18] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the american statistical association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[19] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *Ieee trans. signal process.*, vol. 65, no. 3, pp. 794–816, 2016.

[20] Q. Yao and J. T. Kwok, “Efficient learning with nonconvex regularizers by nonconvexity redistribution.,” *Journal of machine learning research*, 2018.

[21] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, *et al.*, “Optimization with sparsity-inducing penalties,” *Foundations and trends® in machine learning*, vol. 4, no. 1, pp. 1–106, 2012.

[22] D. Geman and G. Reynolds, “Constrained restoration and the recovery of discontinuities,” *Ieee transactions on pattern analysis and machine intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[23] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. Oxford University Press Oxford, 2004, vol. 3.

[24] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, “A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing,” *Ieee signal processing magazine*, vol. 33, no. 1, pp. 57–77, 2016.

[25] N. Parikh, S. Boyd, *et al.*, “Proximal algorithms,” *Foundations and trends® in optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[26] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer, 2007.

[27] C. W. Granger, “Cointegrated variables and error correction models,” Unpublished USCD Discussion Paper 83-13a, Tech. Rep., 1983.

[28] R. F. Engle and C. W. Granger, “Co-integration and error correction: Representation, estimation, and testing,” *Econometrica: Journal of the econometric society*, pp. 251–276, 1987.

[29] S. Johansen, “Statistical analysis of cointegration vectors,” *Journal of economic dynamics and control*, vol. 12, no. 2, pp. 231–254, 1988.

[30] ——, “Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models,” *Econometrica: Journal of the econometric society*, pp. 1551–1580, 1991.

[31] ——, “Identifying restrictions of linear equations with applications to simultaneous equations and cointegration,” *Journal of econometrics*, vol. 69, no. 1, pp. 111–132, 1995.

[32] A. Pole, *Statistical arbitrage: Algorithmic trading insights and techniques*. John Wiley & Sons, 2011, vol. 411.

[33] S. T. Rachev, C. Menn, and F. J. Fabozzi, *Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing*. John Wiley & Sons, 2005, vol. 139.

[34] P. H. Franses and N. Haldrup, "The effects of additive outliers on tests for unit roots and cointegration," *Journal of business & economic statistics*, vol. 12, no. 4, pp. 471–478, 1994.

[35] P. H. Franses, T. Kloek, and A. Lucas, "Outlier robust analysis of long-run marketing effects for weekly scanning data," *Journal of econometrics*, vol. 89, no. 1, pp. 293–315, 1998.

[36] H. B. Nielsen, "Cointegration analysis in the presence of outliers," *The econometrics journal*, vol. 7, no. 1, pp. 249–271, 2004.

[37] A. Lucas, "Unit root tests based on m estimators," *Econometric theory*, vol. 11, no. 02, pp. 331–346, 1995.

[38] ——, "An outlier robust unit root test with an application to the extended nelson-plosser data," *Journal of econometrics*, vol. 66, no. 1, pp. 153–173, 1995.

[39] P. H. Franses and A. Lucas, "Outlier detection in cointegration analysis," *Journal of business & economic statistics*, vol. 16, no. 4, pp. 459–468, 1998.

[40] A. Lucas, "Cointegration testing using pseudolikelihood ratio tests," *Econometric theory*, pp. 149–169, 1997.

[41] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the royal statistical society. series b (methodological)*, pp. 267–288, 1996.

[42] I. Wilms and C. Croux, "Forecasting using sparse cointegration," *International journal of forecasting*, vol. 32, no. 4, pp. 1256–1267, 2016.

[43] L. Chen and J. Z. Huang, "Sparse reduced-rank regression with covariance estimation," *Statistics and computing*, vol. 26, no. 1-2, pp. 461–470, 2016.

[44] B. Bosco, L. Parisio, M. Pelagatti, and F. Baldi, "Long-run relations in European electricity prices," *Journal of applied econometrics*, vol. 25, no. 5, pp. 805–832, 2010.

[45] M. A. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *Ieee transactions on image processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[46] G. Vidyamurthy, *Pairs trading: Quantitative methods and analysis*. John Wiley & Sons, 2004, vol. 217.

[47] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, "Pairs trading: Performance of a relative-value arbitrage rule," *Review of financial studies*, vol. 19, no. 3, pp. 797–827, 2006.

[48] D. S. Ehrman, *The handbook of pairs trading: Strategies using equities, options, and futures*. John Wiley & Sons, 2006, vol. 240.

[49] R. Bookstaber, *A demon of our own design: Markets, hedge funds, and the perils of financial innovation*. John Wiley & Sons, 2007, ISBN: 9781118045589.

[50] D. Butterworth and P. Holmes, "Inter-market spread trading: Evidence from UK index futures markets," *Applied financial economics*, vol. 12, no. 11, pp. 783–790, 2002.

[51] S.-J. Kim, J. Primbs, and S. Boyd, "Dynamic spread trading," *Unpublished working paper*, 2008.

[52] T. Kanamura, S. T. Rachev, and F. J. Fabozzi, "A profit model for spread trading with an application to energy futures," *The journal of trading*, vol. 5, no. 1, pp. 48–62, 2010.

[53] M. Cummins and A. Bucca, "Quantitative spread trading on crude oil and refined products markets," *Quantitative finance*, vol. 12, no. 12, pp. 1857–1875, 2012.

[54] M. Whistler, *Trading pairs: Capturing profits and hedging risk with statistical arbitrage strategies*. John Wiley & Sons, 2004, vol. 216.

[55] C. Alexander and A. Dimitriu, "Indexing and statistical arbitrage," *The journal of portfolio management*, vol. 31, no. 2, pp. 50–63, 2005.

[56] S. F. LeRoy and J. Werner, *Principles of financial economics*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[57] J. G. Nicholas, *Market neutral investing: Long/short hedge fund strategies*. Bloomberg Press, 2000.

[58] B. I. Jacobs and K. N. Levy, *Market neutral strategies*. John Wiley & Sons, 2005, vol. 112.

[59] C. Krauss, "Statistical arbitrage pairs trading strategies: Review and outlook," *Journal of economic surveys*, vol. 31, no. 2, pp. 513–545, 2017.

[60] P. Draper and J. K. Fung, "A study of arbitrage efficiency between the FTSE-100 index futures and options contracts," *Journal of futures markets*, vol. 22, no. 1, pp. 31–58, 2002.

[61] G. Hong and R. Susmel, "Pairs-trading in the Asian ADR market," *University of houston, unpublished manuscript*, 2003.

[62] M. S. Perlin, "Evaluation of pairs-trading strategy at the Brazilian financial market," *Journal of derivatives & hedge funds*, vol. 15, no. 2, pp. 122–136, 2009.

[63] M. Avellaneda and J.-H. Lee, "Statistical arbitrage in the US equities market," *Quantitative finance*, vol. 10, no. 7, pp. 761–782, 2010.

[64] S. Drakos, "Statistical arbitrage in S&P500," *Journal of mathematical finance*, vol. 6, no. 01, p. 166, 2016.

[65] J. L. Farrell and W. J. Reinhart, *Portfolio management: Theory and application*. McGraw-Hill, 1997.

[66] H. M. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.

[67] A. d'Aspremont, "Identifying small mean-reverting portfolios," *Quantitative finance*, vol. 11, no. 3, pp. 351–364, 2011.

[68] M. Cuturi and A. d'Aspremont, "Mean reversion with a variance threshold," in *Proc. of the 30th int. conf. on machine learning (icml-13)*, Atlanta, GA, USA, Jun. 2013, pp. 271–279.

[69] ——, "Mean-reverting portfolios," in *Financial signal processing and machine learning*, A. N. Akansu, S. R. Kulkarni, and D. M. Malioutov, Eds., John Wiley & Sons, 2016, ch. 3, pp. 23–40.

[70] F. J. Fabozzi, S. M. Focardi, and P. N. Kolm, *Quantitative equity investing: Techniques and strategies*. John Wiley & Sons, 2010.

[71] Y. Huang and D. P. Palomar, "Randomized algorithms for optimal solutions of double-sided QCQP with applications in signal processing," *Ieee trans. signal process.*, vol. 62, no. 5, pp. 1093–1108, Jan. 2014.

[72] S. Johansen, "Modelling of cointegration in the vector autoregressive model," *Economic modelling*, vol. 17, no. 3, pp. 359–373, 2000.

[73] H. M. Markowitz, “The optimization of a quadratic function subject to linear constraints,” *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 111–133, 1956.

[74] G. E. Box and G. C. Tiao, “A canonical analysis of multiple time series,” *Biometrika*, vol. 64, no. 2, pp. 355–365, 1977.

[75] G. E. Box and D. A. Pierce, “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *J. amer. statist. assoc.*, vol. 65, no. 332, pp. 1509–1526, 1970.

[76] N. D. Ylvisaker, “The expected number of zeros of a stationary gaussian process,” *Ann. math. statist.*, vol. 36, no. 3, pp. 1043–1046, 1965.

[77] B. Kedem and S. Yakowitz, *Time series analysis by higher order crossings*. Piscataway, NJ, USA: IEEE Press, 1994.

[78] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[79] A. Beck and Y. C. Eldar, “Strong duality in nonconvex quadratic optimization with two quadratic constraints,” *Siam journal on optimization*, vol. 17, no. 3, pp. 844–860, 2006.

[80] Y. Huang and D. P. Palomar, “Rank-constrained separable semidefinite programming with applications to optimal beamforming,” *Ieee trans. signal process.*, vol. 58, no. 2, pp. 664–678, Sep. 2010.

[81] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[82] J. Song, P. Babu, and D. P. Palomar, “Sparse generalized eigenvalue problem via smooth optimization,” *Ieee trans. signal process.*, vol. 63, no. 7, pp. 1627–1642, Jan. 2015.

[83] J. J. Moré, “Generalizations of the trust region problem,” *Optimization methods and software*, vol. 2, no. 3-4, pp. 189–209, 1993.

[84] T. K. Pong and H. Wolkowicz, “The generalized trust region subproblem,” *Computational optimization and applications*, vol. 58, no. 2, pp. 273–322, 2014.

[85] D. R. Hunter and K. Lange, “A tutorial on MM algorithms,” *Amer. statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[86] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *Siam j. optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[87] J. Pang, "Partially B-regular optimization and equilibrium problems," *Math. oper. res.*, vol. 32, no. 3, pp. 687–699, 2007.

[88] J.-S. Pang, M. Razaviyayn, and A. Alvarado, "Computing B-stationary points of nonsmooth DC programs," *Math. oper. res.*, vol. 42, no. 1, pp. 95–118, Feb. 2017.

[89] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *J. amer. statist. assoc.*, vol. 74, no. 366a, pp. 427–431, 1979.

[90] P. C. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika*, vol. 75, no. 2, pp. 335–346, 1988.

[91] W. F. Sharpe, "The Sharpe ratio," *The journal of portfolio management*, vol. 21, no. 1, pp. 49–58, 1994.

[92] 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, 2016.

[93] W. Xiong, "Convergence trading with wealth effects: An amplification mechanism in financial markets," *Journal of financial economics*, vol. 62, no. 2, pp. 247–292, 2001.

[94] Y. Stander, D. Marais, and I. Botha, "Trading strategies with Copulas," *Journal of economic and financial sciences*, vol. 6, no. 1, pp. 83–107, 2013.

[95] R. J. Elliott, J. Van Der Hoek*, and W. P. Malcolm, "Pairs trading," *Quantitative finance*, vol. 5, no. 3, pp. 271–276, 2005.

[96] T. Leung and X. Li, *Optimal mean reversion trading: Mathematical analysis and practical applications*. World Scientific, 2016.

[97] H. Rad, R. K. Y. Low, and R. Faff, "The profitability of pairs trading strategies: Distance, cointegration and copula methods," *Quantitative finance*, vol. 16, no. 10, pp. 1541–1558, 2016.

[98] H. Zhang and Q. Zhang, "Trading a mean-reverting asset: Buy low and sell high," *Automatica*, vol. 44, no. 6, pp. 1511–1518, 2008.

[99] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio design via majorization-minimization method," in *2016 50th asilomar conf. signals proc. systems and computers*, Pacific Grove, CA, USA, Nov. 2016, pp. 1530–1534. DOI: 10.1109/ACSSC.2016.7869634.

[100] Z. Zhao, R. Zhou, Z. Wang, and D. P. Palomar, "Optimal portfolio design for statistical arbitrage in finance," in *Proc. the 20th ieee statistical signal processing workshop (ssp 2018)*, Freiburg, Germany, Jun. 2018, pp. 801–805.

[101] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris, "Sparse and stable Markowitz portfolios," *Proceedings of the national academy of sciences*, vol. 106, no. 30, pp. 12 267–12 272, 2009.

[102] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed $\ell_0$ norm," *Ieee transactions on signal processing*, vol. 57, no. 1, pp. 289–301, 2009.

[103] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations research*, vol. 26, no. 4, pp. 681–683, 1978.

[104] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *Ieee transactions on signal processing*, vol. 62, no. 3, pp. 641–656, 2014.

[105] S. J. Wright and J. Nocedal, "Numerical optimization," *Springer science*, vol. 35, no. 67-68, p. 7, 1999.

[106] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization-Part I: Theory.," *Ieee transactions on signal processing*, vol. 65, no. 8, pp. 1929–1944, 2017.

[107] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[108] K.-C. Toh, M. J. Todd, and R. H. 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*, Springer, 2012, pp. 715–754.

[109] A. MOSEK, *The MOSEK optimization toolbox for MATLAB manual version 7.1 (revision 28)*, 2015, p. 17.

[110] J. Lofberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Computer aided control systems design, 2004 ieee international symposium on*, IEEE, 2004, pp. 284–289.

[111] M. Grant, S. Boyd, and Y. Ye, *CVX: Matlab software for disciplined convex programming*, 2008.

[112] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and trends® in machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[113] D. P. Palomar, "Convex primal decomposition for multicarrier linear MIMO transceivers," *Ieee transactions on signal processing*, vol. 53, no. 12, pp. 4661–4674, 2005.

[114] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the $\ell_1$-ball for learning in high dimensions," in *Proceedings of the 25th international conference on machine learning*, ACM, 2008, pp. 272–279.

[115] M. Li, D. Sun, and K.-C. Toh, "A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization," *Siam journal on optimization*, vol. 26, no. 2, pp. 922–950, 2016.