ABSTRACT

Title of dissertation: MATRIX REDUCTION
IN NUMERICAL OPTIMIZATION

Sungwoo Park, Doctor of Philosophy, 2011

Dissertation directed by: Professor Dianne P. O'Leary
Department of Computer Science

Matrix reduction by eliminating some terms in the expansion of a matrix has been applied to a variety of numerical problems in many different areas. Since matrix reduction has different purposes for particular problems, the reduced matrices also have different meanings. In regression problems in statistics, the reduced parts of the matrix are considered to be noise or observation error, so the given raw data are purified by the matrix reduction. In factor analysis and principal component analysis (PCA), the reduced parts are regarded as idiosyncratic (unsystematic) factors, which are not shared by multiple variables in common. In solving constrained convex optimization problems, the reduced terms correspond to unnecessary (inactive) constraints which do not help in the search for an optimal solution.

In using matrix reduction, it is both critical and difficult to determine how and how much we will reduce the matrix. This decision is very important since it determines the quality of the reduced matrix and the final solution. If we reduce too much, fundamental properties will be lost. On the other hand, if we reduce too little, we cannot expect enough benefit from the reduction. It is also a difficult decision because the criteria for the reduction must be based on the particular type of problem.

In this study, we investigate matrix reduction for three numerical optimization problems. First, the total least squares problem uses matrix reduction to remove noise in observed data which follow an underlying linear model. We propose a new method to make the matrix reduction successful under relaxed noise assumptions. Second, we apply matrix reduction to the problem of estimating a covariance matrix of stock returns, used in financial portfolio optimization problem. We summarize all the previously proposed estimation methods in a common framework and present a new and effective Tikhonov method. Third, we present a new algorithm to solve semidefinite programming problems, adaptively reducing inactive constraints. In the constraint reduction, the Schur complement matrix for the Newton equations is the object of the matrix reduction. For all three problems, we propose appropriate criteria to determine the intensity of the matrix reduction. In addition, we verify the correctness of our criteria by experimental results and mathematical proof.

MATRIX REDUCTION IN NUMERICAL OPTIMIZATION

by

Sungwoo Park

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfilment
of the requirement for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Dianne P. O'Leary
Professor Howard Elman
Professor André Tits
Professor James A. Reggia
Professor Sung Lee

Acknowledgments

I would like to express immeasurable gratitude to my advisor, Professor Dianne O'Leary, for her efforts to lead me to a good researcher. She inspires me by professional insights and guides me to right directions whenever I struggled with difficult academic problems. She never saves her efforts to correct my manuscirpt many times. She also encourages my academic motivation with sincere and considerate advice. Without her guidance and support, my dissertation could not be completed.

I would like to thank my advisory committee members, Professor Howard Elman, Professor André Tits, Professor James A. Reggia, and Professor Sung Lee, for sharing their time to review the manuscript and for giving sincere comments to improve the dissertation. Especially, I deeply appreciate very careful comments by Professor André Tits on the study about semidefinite programming.

I heartily thank my parents for their endless sacrifice and love. I also thank all my friends in University of Maryland. Especially, I will never forget the time with Sukhyun Song, Eunhui Park, and Joonghoon Lee when we settled in Maryland. I could have wonderful time in Maryland thanks to my colleagues in CS and ECE departments: Dr. Jaeyoon Jung, Dr. Beomseok Nam, Dr. Ilchul Yoon, Dr. Minkyoung Cho, Dr. Youngmin Kim, Dr. Jik-soo Kim, Dr. Ji Sun Shin, Dr. Jin Hyuk Jung, Hyunyoung Song, Jaehwan Lee, Song Ie Noh, Dr. Soo Bum Lee, Inseok Choi, Kyungjin Yoo, Inkeun Cho, Ginnah Lee, and all other students.

I am grateful for support from NSF grant CCF 05 14213 and DOE grant DESC 0002218.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation develops the use of matrix reduction techniques to simplify and stabilize the solutions to various optimization problems.

1.1 Matrix Reduction

Matrix reduction approximates a matrix by removing some terms in its decomposition. Suppose that a matrix $\boldsymbol{M}$ can be expressed as a summation of matrices $\boldsymbol{M}_i$ as

$$\boldsymbol{M} = \sum_{i=1}^{k} \boldsymbol{M}_i = \boldsymbol{M}_1 + \cdots + \boldsymbol{M}_k.$$

This kind of expansion is common in matrix computation. For instance, any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has a singular value decomposition (SVD) [31, Chapter 2.5]

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\,\boldsymbol{V}^T,$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ is a diagonal matrix. We can write this decomposition as summation of rank one matrices

$$\boldsymbol{A} = \sum_{i=1}^{\min(m,n)} s_i \, \boldsymbol{u}_i \boldsymbol{v}_i^T ,$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the i-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, and s_i is the i-th diagonal element of $\boldsymbol{S}$.

If a matrix $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ is symmetric and positive definite, Cholesky decomposition [31, Chapter 4.2] also generates such an expansion as

$$\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^T = \sum_{i=1}^{m} \boldsymbol{l}_i \, \boldsymbol{l}_i^T ,$$

where $\boldsymbol{L}$ is a lower triangular matrix, and $\boldsymbol{l}_i$ is the i-th column of $\boldsymbol{L}$. This expansion is particularly important when $\boldsymbol{B}$ is updated by a low-rank correction since this can be accomplished by adding a small number of terms to the expression [27].

Broadly speaking, there are two different approaches to matrix reduction. First, if we know that only the first $\widehat{k}$ matrices $\boldsymbol{M}_i$ are important to us, we can construct a reduced matrix $\widehat{\boldsymbol{M}}$ as

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^{\widehat{k}} \boldsymbol{M}_i = \boldsymbol{M}_1 + \cdots + \boldsymbol{M}_{\widehat{k}}.$$

This reduction method is called *truncation*. Alternatively, we can apply a filtering factor $\phi_i \in [0, 1]$ to each matrix $\boldsymbol{M}_i$ for $i = 1, \ldots, k$. Then, the reduced matrix $\widehat{\boldsymbol{M}}$ becomes

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^{k} \phi_i \boldsymbol{M}_i = \phi_1 \boldsymbol{M}_1 + \cdots + \phi_k \boldsymbol{M}_k.$$

This filtering-based reduction can be regarded as a generalized version of *truncation* since *truncation* is a special case with $\phi_i \in \{0, 1\}$. Both reduction methods are used in many applications such as regularization of ill-posed problems and factor analysis.

We can also classify the matrix reduction approaches as *build-down* and *build-up*, depending on whether we remove some terms in a given matrix expansion or we construct the reduced matrix by adding terms until a certain goal is achieved. For example, while a complete matrix $\boldsymbol{M}$ is given to us in the problems of regression and factor analysis, we construct $\widehat{\boldsymbol{M}}$ by adding matrices $\boldsymbol{M}_i$ in constrained convex optimization.

The purposes of the matrix reduction are very different depending on particular problems. First, in regression problems in statistics, the truncated or filtered terms are considered to be noise or observation error, so matrix reduction purifies the given raw data. This can be useful in solving least squares problems for an over-determined linear system or regularizing the solution to an ill-posed problem. Second, in factor analysis and principal component analysis (PCA), the reduced parts are regarded as idiosyncratic (unsystematic) factors, which are not shared by multiple variables in common. Third, in constrained convex optimization problems, the reduced terms might correspond to unnecessary (inactive) constraints, which do not make significant contributions to the search for an optimal solution. So, we expect a benefit of decreased computational cost by using matrix reduction.

Whenever matrix reduction is applied, it is a very critical but difficult issue to decide how much to reduce the matrix. This important decision determines both the quality of the reduced matrix and that of the final result. If we reduce too much, we may fail to solve the problem. On the other hand, if we reduce too little, we cannot expect enough benefit from the reduction. It is a difficult decision because criteria for the reduction must be tailored to the problem and the circumstances. For example, in regularization of ill-

$\sigma_{\max}(\boldsymbol{X})$	The largest singular value of $\boldsymbol{X}$
$\sigma_{\min}(\boldsymbol{X})$	The smallest singular value of $\boldsymbol{X}$
$\sigma_i(\boldsymbol{X})$	The i-th largest singular value of $\boldsymbol{X}$
$\\|\boldsymbol{x}\\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$	2-norm for a vector $\boldsymbol{x}$
$\\|\boldsymbol{X}\\|_2 = \sigma_{\max}(\boldsymbol{X})$	2-norm for a matrix $\boldsymbol{X}$
$\\|\boldsymbol{X}\\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n x_{ij}^2}$	Frobenius norm for a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$
$\mathrm{tr}\,(\boldsymbol{X}) = \sum_{i=1}^n x_{ii}$	Trace of matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$
$\boldsymbol{I}_p$	An identity matrix of dimension p

Table 1.1: *Notation.*

posed problems, the criteria may change based on which distribution the embedded noise follows, or how the noise in different variables is correlated. Because of this difficulty, the criteria for constraint reduction has been studied in a variety of applications.

In this dissertation, we discuss matrix reduction in three numerical optimization problems. Our study focuses on how we can determine appropriate reduction intensity for successful matrix reduction in these problems. We introduce the problems in the next section.

Throughout this dissertation, we use the notation defined in Table 1.1. In addition, a few basic statistical definitions are frequently used. When a continuous random variable x has a probability density function $p_x(x)$, the expected value $\mathbb{E}(x)$ is defined as

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x\ p_x(x).$$

Then, the variance var(x) and the standard deviation std(x) are defined as

$$\begin{aligned} \mathrm{var}(x) &= \mathbb{E}\left((x - \mathbb{E}(x))^2\right) = \mathbb{E}(x^2) - \left(\mathbb{E}(x)\right)^2, \\ \mathrm{std}(x) &= \sqrt{\mathrm{var}(x)}. \end{aligned}$$

For two random variables x and y, the covariance cov(x, y) and the correlation corr(x, y)

are defined as

$$\begin{aligned}\text{cov}(x,y) &= \mathbb{E}\left(\, (x-\mathbb{E}(x))(y-\mathbb{E}(y))\,\right) \;= \mathbb{E}(xy)-\mathbb{E}(x)\mathbb{E}(y),\\ \text{corr}(x,y) &= \frac{\text{cov}(x,y)}{\text{std}(x)\text{std}(y)}.\end{aligned}$$

1.2 Overview of Numerical Optimization Problems

1.2.1 Total Least Squares Problems

Suppose that we have an underlying linear model,

$$(\boldsymbol{A}-\boldsymbol{E}_A)\boldsymbol{X} = (\boldsymbol{B}-\boldsymbol{E}_B),$$

where $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$ are unknown; they result from noise in the observed matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times d}$. To estimate the parameters $\boldsymbol{X}$, we construct a minimization problem

$$\min_{\boldsymbol{X},\Delta\boldsymbol{A},\Delta\boldsymbol{B}} \|[\Delta\boldsymbol{A},\Delta\boldsymbol{B}]\|_F,$$

subject to

$$(\boldsymbol{A}-\Delta\boldsymbol{A})\boldsymbol{X} = (\boldsymbol{B}-\Delta\boldsymbol{B}),$$

$$\text{rank}\left([(\boldsymbol{A}-\Delta\boldsymbol{A}),(\boldsymbol{B}-\Delta\boldsymbol{B})]\right) = r,$$

where r is the known rank of the noise-free data $(\boldsymbol{A}-\boldsymbol{E}_A)$.

The minimization problem above can be solved by matrix reduction on the SVD of $[\boldsymbol{A},\boldsymbol{B}]$. If there were no noise in $\boldsymbol{A}$ and $\boldsymbol{B}$, the concatenated matrix $[\boldsymbol{A},\boldsymbol{B}]$ would also have rank r since Range $(\boldsymbol{B}) \subseteq$ Range $(\boldsymbol{A})$. If the rank r of the noise-free data $(\boldsymbol{A} - \boldsymbol{E}_A)$ is given to us, we can truncate all but the r largest singular values of $[\boldsymbol{A}, \boldsymbol{B}]$. By the Eckart-Young-Mirsky Theorem, the resulting $(\boldsymbol{X}, \Delta\boldsymbol{A}, \Delta\boldsymbol{B})$ is the solution to the minimization problem. In addition, if the noise matrices $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$ are mutually uncorrelated and have zero mean and identical standard deviations, it is known that the minimization problem above gives us a consistent estimate $\boldsymbol{X}$ for the underlying linear model.

Our study starts from the question of how we can estimate $\boldsymbol{X}$ if we do not know the rank r or if the embedded noise matrices $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$ do not have identical standard deviations and the standard deviations are unknown. If the rank r is not given to us, we need to decide how many singular values to truncate. If the standard deviations of the noise are different and we do not know their values, we also need to find an appropriate weight α so that weighted data $\alpha\boldsymbol{A}$ and $(1 - \alpha)\boldsymbol{B}$ contain noise with identical standard deviations.

In Chapter 2, we propose a method to estimate the rank r and the weight α. We also present experimental results to evaluate the proposed method.

1.2.2 Covariance Matrix Estimation

In financial portfolio theory, Markowitz [59] proposed the Mean-Variance (MV) portfolio problem to find an optimal portfolio of N stocks satisfying given constraints. The MV portfolio problem requires an estimated covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ for the N stock returns. It is well known that the performance of the portfolio is very sensitive to the quality of the covariance matrix estimate, but a conventional sample covariance matrix is far from a good estimate.

The main difficulty is that the observed stock return data contain too much noise. Matrix reduction can be used to reduce the error in the covariance matrix estimate. Suppose that we have stock return data $\boldsymbol{R} \in \mathbb{R}^{N\times T}$ of N stocks for T time periods. For appropriate principal component analysis (PCA), we normalize each stock return, so that large return values for a few stocks do not overwhelm the other return values. Let $\boldsymbol{Z}$ denote the normalized data with zero-means and identical standard deviations. From the singular value decomposition of $\boldsymbol{Z}$, we have

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{S}\,\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{F} = \sum_{i=1}^{T} \boldsymbol{u}_i\,\boldsymbol{f}_i^T,$$

where $\boldsymbol{F} = \boldsymbol{S}\,\boldsymbol{V}^T$, $\boldsymbol{u}_i$ is the i-th column of $\boldsymbol{U}$, and $\boldsymbol{f}_i^T$ is the i-th row of $\boldsymbol{F}$. In PCA, the vector $\boldsymbol{f}_i$ is called the i-th principal component affecting the stock returns, and the vector $\boldsymbol{u}_i$ is called a load which determines how much each stock return is affected by the i-th component. Previously, many people proposed truncating a few smallest singular values, expecting that the principal components corresponding to the smallest singular values are more significantly contaminated by noise. However, no one has given a clear answer as to how many principal components should be truncated. This is a very difficult decision because we fundamentally do not know how many factors govern the stock returns.

In Chapter 3, we apply a Tikhonov filtering function to the principal components, a monotonically increasing function of the singular value. With this smooth filtering, we expect that the influence of important principal components is amplified while potential information in less important principal components is still preserved. Furthermore, we propose a method to determine filtering intensity. Experiments using stock return data in NYSE, AMEX, and NASDAQ from 1958 to 2007, show that the MV portfolio using Tikhonov filtered covariance matrix performs quite well.

1.2.3 Interior Point Method for Semidefinite Programming

The constrained convex optimization problem known as semidefinite programming (SDP) has the following primal and dual problems:

$$\begin{aligned} \text{Primal SDP:} \quad & \min_{\boldsymbol{X}} \boldsymbol{C} \bullet \boldsymbol{X} \quad \text{s.t.} \ \boldsymbol{A}_i \bullet \boldsymbol{X} = b_i \text{ for } i = 1, \ldots, m, \ \ \boldsymbol{X} \succeq \boldsymbol{0}, \\ \text{Dual SDP:} \quad & \max_{\boldsymbol{y}} \boldsymbol{b}^T \boldsymbol{y} \quad \text{s.t.} \ \sum_{i=1}^{m} y_i \boldsymbol{A}_i + \boldsymbol{Z} = \boldsymbol{C}, \ \ \boldsymbol{Z} \succeq \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{C}$, $\boldsymbol{A}_i$, $\boldsymbol{X}$, and $\boldsymbol{Z}$ are $n \times n$ symmetric matrices, $\boldsymbol{C} \bullet \boldsymbol{X} = \text{tr}\,(\boldsymbol{C}\boldsymbol{X})$ is the trace of the matrix, and $\boldsymbol{Z} \succeq \boldsymbol{0}$ means that $\boldsymbol{Z}$ is positive semidefinite.

In an interior point method (IPM) for solving the SDP, we use Newton's method to find a direction $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ leading toward an optimal solution and following a central path defined by the primal and dual constraints and complementarity equation. To make the computation of the direction efficient, the Newton equations are reduced to the linear system,

$$\boldsymbol{M}\Delta\boldsymbol{y} = \boldsymbol{g},$$

where the Schur complement matrix $\boldsymbol{M}$ is determined by the constraint matrices $\boldsymbol{A}_i$ and the current point $(\boldsymbol{X}, \boldsymbol{Z})$, and $\boldsymbol{g}$ is defined by current residuals. The IPM repeatedly solves this reduced equation until the iterate satisfies a given convergence tolerance.

It takes $O(mn^3 + m^2n^2)$ operations to compute $\boldsymbol{M}$, which is most expensive part for each iteration, so we can expect benefit by reducing its computational cost. In many applications of SDP such as the binary code problem, the quadratic assignment problem, and the traveling salesman problem, the matrices $\boldsymbol{A}_i$ and $\boldsymbol{C}$ have identical diagonal block structure. Using the block structure, $\boldsymbol{M}$ can be expanded to

$$\boldsymbol{M} = \sum_{j=1}^{p} \boldsymbol{M}_j,$$

where p is the number of diagonal blocks and matrix $\boldsymbol{M}_j$ is associated with the j-th constraint block. If some constraint blocks make insignificant or detrimental contributions to finding the search direction, we may be able to ignore the corresponding $\boldsymbol{M}_j$ when we compute $\boldsymbol{M}$. We call such blocks *inactive*. Similar to the previous problems, it is critical to determine which constraint blocks can be ignored while still guaranteeing that the iteration converges to the optimal solution.

In Chapter 4, we explain how constraint reduction can be applied to IPM for SDP problems and propose a basic predictor-corrector algorithm with constraint reduction. We demonstrate its performance by experiments with test problems. In Chapter 5, we develop a new predictor-corrector algorithm with adaptive criteria to determine *inactive* constraint blocks. We verify the correctness of the criteria by proving the global convergence of the proposed algorithm. Its polynomial complexity is also verified to be $O(n \ln(\epsilon_0/\epsilon))$, where ϵ_0 is an initial residual and ϵ is a required tolerance.

1.2.4 Summary

The work in this dissertation proposes matrix reduction methods for solving three important problems: total least squares problems, covariance matrix estimation, and semidefinite programming problems. We now consider each of these problems in turn, and present conclusions in Chapter 6.

Chapter 2

Implicitly-Weighted Total Least Squares

In a total least squares (TLS) problem, we estimate an optimal set of model parameters $\boldsymbol{X}$, so that $(\boldsymbol{A} - \Delta\boldsymbol{A})\boldsymbol{X} = \boldsymbol{B} - \Delta\boldsymbol{B}$, where $\boldsymbol{A}$ is the model matrix, $\boldsymbol{B}$ is the observed data, and $\Delta\boldsymbol{A}$ and $\Delta\boldsymbol{B}$ are corresponding corrections. Throughout the matrix reduction, we remove the noise terms in the concatenated matrix $[\boldsymbol{A}, \boldsymbol{B}]$, and estimate the parameter $\boldsymbol{X}$ from the remaining terms. For consistent estimation, it is necessary to adjust the scales of $\boldsymbol{A}$ and $\boldsymbol{B}$ to satisfy a noise assumption prior to applying matrix reduction. In addition, we also need to estimate the column rank of the noise-free model, which determines the number of reduced terms.

When $\boldsymbol{B}$ is a single vector, Rao [72] and Paige and Strakoš [64] suggested formulating standard least squares problems, for which $\Delta\boldsymbol{A} = \boldsymbol{0}$, and data least squares problems, for which $\Delta\boldsymbol{B} = \boldsymbol{0}$, as weighted and scaled TLS problems. In this work we define an implicitly-weighted TLS formulation (ITLS) that reparameterizes these formulations to make computation easier. We derive asymptotic properties of the estimates as the number of rows in the problem approaches infinity, handling the rank-deficient case as well. We discuss the role of the ratio between the variances of errors in $\boldsymbol{A}$ and $\boldsymbol{B}$ in choosing an appropriate parameter in ITLS. We also propose methods for computing the family of solutions efficiently and for choosing the appropriate solution if the ratio of variances is unknown. We provide experimental results on the usefulness of the ITLS family of solutions. This presentation closely follows that in [65].

2.1 Introduction

In formulating a linear model $\boldsymbol{AX} \approx \boldsymbol{B}$, there can be errors in the data $\boldsymbol{B}$, errors in the model matrix $\boldsymbol{A}$, or errors in both $\boldsymbol{B}$ and $\boldsymbol{A}$. This has led to the formulation of three distinct problems: given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times d}$, where usually $m > n$, find $\boldsymbol{X}$ and small correction matrices $\Delta\boldsymbol{A}$, and $\Delta\boldsymbol{B}$ satisfying

$$(\boldsymbol{A} - \Delta\boldsymbol{A})\boldsymbol{X} = \boldsymbol{B} - \Delta\boldsymbol{B}, \tag{2.1.1}$$

where

- $\Delta\boldsymbol{A} = \boldsymbol{0}$ for the *least squares* (LS) problem.
- $\Delta\boldsymbol{B} = \boldsymbol{0}$ for the *data least squares* (DLS) problem.
- both $\Delta\boldsymbol{A}$ and $\Delta\boldsymbol{B}$ are allowed to be nonzero for the *total least squares* (TLS) problem.

In least squares formulations, the values of $\boldsymbol{X}$, $\Delta\boldsymbol{A}$, and $\Delta\boldsymbol{B}$ are found by minimizing

$$\|[\Delta\boldsymbol{A},\Delta\boldsymbol{B}]\|_F. \tag{2.1.2}$$

Minimizing (2.1.2) makes sense, for example, if the errors in $\boldsymbol{A}$ and $\boldsymbol{B}$ are zero-mean, mutually uncorrelated, and drawn from the same distribution. If, on the other hand, the standard deviation of the errors in $\boldsymbol{A}$ is γ times the standard deviation of the errors in $\boldsymbol{B}$, then we should weight the terms in (2.1.2) as

$$\|[\Delta\boldsymbol{A},\gamma\Delta\boldsymbol{B}]\|_F.$$

For a single right-hand ($d = 1$), Rao [72] formulated a weighted TLS, and Paige and Strakoš [64] formulated a scaled TLS problem, which uses a scale factor γ to relate $\boldsymbol{A}$ and $\boldsymbol{B}$. The solution to their scaled problem is the TLS solution when $\gamma = 1$, approaches the solution to the LS problem as $\gamma \to 0$, and approaches the solution to the DLS problem as $\gamma \to \infty$. The underlying statistical assumption behind these methods is that the true error matrices for $\boldsymbol{A}$ and $\boldsymbol{B}$ are column-wise uncorrelated, and the columns of $\boldsymbol{A}$ have variance not necessarily identical to that of the columns of $\boldsymbol{B}$. In order to correctly obtain an estimate for $\boldsymbol{X}$, the covariance matrices must be known except for the single scaling constant γ that relates the two variances. However, neither [72] nor [64] discusses how to determine the scaling factor.

The main results of our work are as follows. We define in Section 2.2 an implicitly-weighted TLS formulation (ITLS) that reparameterizes these formulations to make computation easier. In particular, we use a scaling constant that ranges between 0 and 1 rather than the less convenient 0 and ∞. We propose in Section 2.3 an efficient method for computing the family of solutions. We prove asymptotic properties of the solution (as $m \to \infty$) in Section 2.4, holding even for rank-deficient problems. With this guidance, we propose algorithms for parameter choice in Section 2.5. We provide experimental results on the usefulness of ITLS in Section 2.6.

A simple notational convention will be helpful: A matrix $\boldsymbol{E}_C$ always denotes the true error in the matrix $\boldsymbol{C}$, and a matrix $\Delta\boldsymbol{C}$ always denotes our correction matrix for $\boldsymbol{C}$. We denote by $\widetilde{\boldsymbol{X}}$ the true parameters for our model, by $\boldsymbol{X}$ an estimated set of parameters, and by $\widehat{\boldsymbol{X}}$ a TLS estimate.

2.2 Implicitly Weighted Total Least Squares

In this section, we define the ITLS problem and show its relation to previous problem formulations. Perhaps most importantly, we discuss the error assumption that makes the ITLS formulation reasonable.

2.2.1 ITLS and Other Estimation Methods

Our underlying data model for ITLS is the following:

$$(\boldsymbol{A} - (1-\alpha)\boldsymbol{E}_{A_w})\widetilde{\boldsymbol{X}} = (\boldsymbol{B} - \alpha\boldsymbol{E}_{B_w}), \tag{2.2.1}$$

where matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times d}$ are given, α is a given weighting parameter satisfying $\alpha \in [0,1]$, and $\boldsymbol{E}_{A_w}$ and $\boldsymbol{E}_{B_w}$ are the scaled errors in $\boldsymbol{A}$ and $\boldsymbol{B}$. We want to estimate the matrix $\widetilde{\boldsymbol{X}}$, the true values of the model's parameters.

Given that model, we define the ITLS problem as follows:

$$\min_{\boldsymbol{X},\Delta\boldsymbol{A}_w,\Delta\boldsymbol{B}_w} \quad \|[\Delta\boldsymbol{A}_w, \Delta\boldsymbol{B}_w]\|_F \tag{2.2.2}$$

subject to

$$(\boldsymbol{A} - (1-\alpha)\Delta\boldsymbol{A}_w)\boldsymbol{X} = (\boldsymbol{B} - \alpha\Delta\boldsymbol{B}_w). \tag{2.2.3}$$

The matrices $\Delta\boldsymbol{A}_w$ and $\Delta\boldsymbol{B}_w$ are corrections corresponding to $\boldsymbol{E}_{A_w}$ and $\boldsymbol{E}_{B_w}$. The following lemma explains how the ITLS formulation unifies DLS, LS, and TLS.

Lemma 2.2.1. *The ITLS defined by (2.2.2) and (2.2.3) is equivalent to DLS when $\alpha = 0$, LS when $\alpha = 1$, and TLS when $\alpha = 1/2$.*

Proof. If $\alpha = 0$, then the matrix $\Delta\boldsymbol{B}_w$ does not contribute to (2.2.3), so its optimal value is $\Delta\boldsymbol{B}_w = \boldsymbol{0}$, and ITLS reduces to the data least squares problem DLS. Similarly, if $\alpha = 1$, then the optimal value of $\Delta\boldsymbol{A}_w$ is $\boldsymbol{0}$ and ITLS reduces to the least squares problem LS. If $\alpha = 1/2$, then we see by defining $\Delta\boldsymbol{A} = \Delta\boldsymbol{A}_w/2$ and $\Delta\boldsymbol{B} = \Delta\boldsymbol{B}_w/2$ that the problem is equivalent to TLS, and the value of our objective function (2.2.2) is two times the norm of the correction term $[\Delta\boldsymbol{A}, \Delta\boldsymbol{B}]$ in (2.1.2). □

In the case of a single right-hand side ($d = 1$), Paige and Strakoš [64] devised a scaled TLS (STLS) formulation. We can easily extend their formulation to the case of multiple right-hand-side data: For a given $\gamma \in (0, \infty)$,

$$\min_{\boldsymbol{X},\Delta\boldsymbol{A}_s,\Delta\boldsymbol{B}_s} \quad \|[\Delta\boldsymbol{A}_s, \Delta\boldsymbol{B}_s]\|_F \;\text{ s.t. }\; (\boldsymbol{A} - \Delta\boldsymbol{A}_s)\boldsymbol{X}\gamma = (\boldsymbol{B}\gamma - \Delta\boldsymbol{B}_s) \tag{2.2.4}$$

Paige and Strakoš proved that STLS becomes LS as $\gamma \to 0$, DLS as $\gamma \to \infty$, and TLS when $\gamma = 1$. The equivalence between ITLS and STLS for these three cases is summarized in Table 2.1. The following lemma establishes equivalence for other values of $\gamma \in (0, \infty)$.

Lemma 2.2.2 (Relation between α and γ)**.** *ITLS in (2.2.2) and STLS in (2.2.4) are equivalent to each other when the parameters α and γ satisfy*

$$\gamma = \frac{1-\alpha}{\alpha} \in (0, \infty). \tag{2.2.5}$$

Proof. Dividing the constraint equation in (2.2.4) by γ, we obtain

$$(\boldsymbol{A} - \Delta\boldsymbol{A}_s)\boldsymbol{X} = (\boldsymbol{B} - \frac{\Delta\boldsymbol{B}_s}{\gamma}).$$

By defining $\Delta\boldsymbol{A}_w$ and $\Delta\boldsymbol{B}_w$ by

$$\Delta\boldsymbol{A}_s = (1-\alpha)\Delta\boldsymbol{A}_w \quad \text{and} \quad \Delta\boldsymbol{B}_s = (1-\alpha)\Delta\boldsymbol{B}_w, \tag{2.2.6}$$

we can rewrite the equation above as

$$(\boldsymbol{A} - (1-\alpha)\Delta\boldsymbol{A}_w)\,\boldsymbol{X} = (\boldsymbol{B} - \frac{(1-\alpha)\Delta\boldsymbol{B}_w}{\gamma}).$$

By using (2.2.5) in the equation above, we obtain the constraint equation (2.2.3). Moreover, by substituting (2.2.6) in the minimization equation in (2.2.4), we obtain

$$\min_{\boldsymbol{X}, \Delta\boldsymbol{A}_w, \Delta\boldsymbol{B}_w} ||[(1-\alpha)\Delta\boldsymbol{A}_w, (1-\alpha)\Delta\boldsymbol{B}_w]||_F,$$

which is equivalent to (2.2.2) since $(1-\alpha)$ is a fixed constant. □

Even though ITLS and STLS are mathematically equivalent, notice that the parameter α in (2.2.3) ranges over $[0, 1]$ while γ in (2.2.4) ranges over $(0, \infty)$. A main theme in this work is the optimal choice of parameter value. Many robust algorithms (e.g., golden

Estimation method	ITLS	STLS
Data Least Squares	$\alpha = 0$	$\gamma \to \infty$
Total Least Squares	$\alpha = 0.5$	$\gamma = 1$
Least Squares	$\alpha = 1$	$\gamma \to 0$

Table 2.1: *Relations between ITLS and STLS.*

section search) can be applied only to optimization problems on bounded domains, so changing the parameterization from γ to α gives a key computational advantage. For this reason, the ITLS formulation is preferable to STLS.

2.2.2 ITLS and the Error Assumption

Now we develop an error assumption consistent with the ITLS formulation and explain the statistical meaning of the weight α. This will clarify when and how ITLS can be used.

Suppose we have a model $\boldsymbol{KZ} \approx \boldsymbol{Y}$, with errors in both the model matrix $\boldsymbol{K}$ and the observations $\boldsymbol{Y}$. As before, we want to estimate the variables $\boldsymbol{Z}$ and the correction matrices $\Delta\boldsymbol{K}$ and $\Delta\boldsymbol{Y}$ satisfying

$$(\boldsymbol{K} - \Delta\boldsymbol{K})\boldsymbol{Z} = (\boldsymbol{Y} - \Delta\boldsymbol{Y}). \tag{2.2.7}$$

We want to formulate this as an *errors-in-variable* (EIV) problem [90, Sec. 8.4]. Such a formulation, from the statistical literature, is closely related to TLS but makes some extra assumptions on the errors. In particular, the rows of the error matrices should be independent, uncorrelated, and identically distributed with finite variance. Under these assumptions, if the noise-free problem has a solution, then the solution to the ITLS problem converges to the true solution with probability 1 as $m \to \infty$, as we will show in Section 2.4.

The independence of the error rows can be imposed by pre-multiplying (2.2.7) by an appropriate matrix $\boldsymbol{D} \in \mathbb{R}^{m \times m}$. *We assume that this pre-multiplication has already been done, so that currently* $\mathbf{D} = \mathbf{I}_m$.

To make the columns of the error uncorrelated with constant variance, we need an estimate of the covariance matrix for the errors $[\boldsymbol{E}_K, \boldsymbol{E}_Y]$. We consider the case in which the errors in $\boldsymbol{K}$ are uncorrelated with the errors in $\boldsymbol{Y}$, so the covariance matrix is block diagonal:

$$\text{cov}[\boldsymbol{E}_K, \boldsymbol{E}_Y] = \begin{bmatrix} \sigma_A^2 \widehat{\boldsymbol{C}}_K & \mathbf{0} \\ \mathbf{0} & \sigma_B^2 \widehat{\boldsymbol{C}}_Y \end{bmatrix}. \tag{2.2.8}$$

We assume that we have good estimates of the nonsingular matrices $\widehat{\mathbf{C}}_K \in \mathbb{R}^{n \times n}$ *and* $\widehat{\mathbf{C}}_Y \in \mathbb{R}^{d \times d}$ *but that one or both of the scalars* σ_A^2 *and* σ_B^2 *may be unknown.* (Often, $\widehat{\boldsymbol{C}}_K$ and $\widehat{\boldsymbol{C}}_Y$ are estimated as identity matrices.)

Let $\widehat{\boldsymbol{C}}_K = \boldsymbol{L}_K \boldsymbol{L}_K^T$ and $\widehat{\boldsymbol{C}}_Y = \boldsymbol{L}_Y \boldsymbol{L}_Y^T$, where $\boldsymbol{L}_K$ and $\boldsymbol{L}_Y$ are Cholesky factors. Define

$$\boldsymbol{A} = \boldsymbol{K}\boldsymbol{L}_K^{-T}, \tag{2.2.9}$$

$$\boldsymbol{B} = \boldsymbol{Y}\boldsymbol{L}_Y^{-T}, \tag{2.2.10}$$

$$\boldsymbol{E}_A = \boldsymbol{E}_K\boldsymbol{L}_K^{-T}, \tag{2.2.11}$$

$$\boldsymbol{E}_B = \boldsymbol{E}_Y\boldsymbol{L}_Y^{-T}, \tag{2.2.12}$$

$$\boldsymbol{X} = \boldsymbol{L}_K^T\boldsymbol{Z}\boldsymbol{L}_Y^{-T}. \tag{2.2.13}$$

Under these definitions, it is easy to verify that the constraint (2.2.7) is equivalent to the constraint (2.1.1) studied above. By the construction of $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$, the covariance matrix for the transformed errors becomes

$$\text{cov}[\boldsymbol{E}_A, \boldsymbol{E}_B] = \begin{bmatrix} \sigma_A^2 \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_B^2 \boldsymbol{I}_d \end{bmatrix}.$$

To satisfy the assumptions in [28] for EIV convergence, we need only scale so that the variances are identical. To do this, we define

$$\begin{aligned} \sigma_E &= \sigma_A + \sigma_B, & (2.2.14) \\ \alpha &= \sigma_B / \sigma_E. & (2.2.15) \end{aligned}$$

Then $0 < \alpha < 1$ (as long as both σ_A^2 and σ_B^2 are positive), and $1 - \alpha = \sigma_A / \sigma_E$. Now let

$$\begin{aligned} \boldsymbol{A}_\alpha &= \alpha \boldsymbol{A}, & (2.2.16) \\ \boldsymbol{B}_\alpha &= (1 - \alpha) \boldsymbol{B}. & (2.2.17) \end{aligned}$$

Then the corresponding (true) errors $\boldsymbol{E}_{A_\alpha} = \alpha \boldsymbol{E}_A$ and $\boldsymbol{E}_{B_\alpha} = (1 - \alpha) \boldsymbol{E}_B$ are uncorrelated and have identical variances $\sigma_A^2 \sigma_B^2 / \sigma_E^2$. Finally, we obtain a linear model containing uncorrelated errors with identical variances:

$$(\boldsymbol{A}_\alpha - \boldsymbol{E}_{A_\alpha}) \boldsymbol{X}_\alpha = (\boldsymbol{B}_\alpha - \boldsymbol{E}_{B_\alpha}), \tag{2.2.18}$$

where

$$\boldsymbol{E}_{A_\alpha} = \frac{\sigma_B}{\sigma_E} \boldsymbol{E}_A, \quad \boldsymbol{E}_{B_\alpha} = \frac{\sigma_A}{\sigma_E} \boldsymbol{E}_B, \quad \text{and} \quad \boldsymbol{X}_\alpha = \left(\frac{\sigma_A}{\sigma_B} \right) \boldsymbol{X}. \tag{2.2.19}$$

The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be determined from the observed data matrices ($\boldsymbol{K}$, $\boldsymbol{Y}$) and the Cholesky factors ($\boldsymbol{L}_K, \boldsymbol{L}_Y$), but $\boldsymbol{A}_\alpha$ and $\boldsymbol{B}_\alpha$ contain the parameters σ_A^2 and σ_B^2.

Using the linear model (2.2.18), we can formulate a TLS problem, which includes the ratio σ_A^2/σ_B^2:

$$\min_{\boldsymbol{X}_\alpha, \Delta\boldsymbol{A}_\alpha, \Delta\boldsymbol{B}_\alpha} \|\Delta\boldsymbol{A}_\alpha, \Delta\boldsymbol{B}_\alpha\|_F \text{ s.t. } (\boldsymbol{A}_\alpha - \Delta\boldsymbol{A}_\alpha)\boldsymbol{X}_\alpha = (\boldsymbol{B}_\alpha - \Delta\boldsymbol{B}_\alpha). \tag{2.2.20}$$

We have thus proven the following lemma.

Lemma 2.2.3 (ITLS and equivalent TLS)**.** *If $\sigma_A^2 > 0$ and $\sigma_B^2 > 0$, then the TLS problem (2.2.20) is equivalent to ITLS (2.2.2)-(2.2.3) when $\alpha \in (0,1)$ satisfies*

$$\frac{\sigma_A}{\sigma_B} = \frac{1-\alpha}{\alpha}. \tag{2.2.21}$$

Paige and Strakoš [64] also made use of σ_A^2/σ_B^2 in defining γ for their STLS formulation.

We see that if we know the ratio of σ_A^2 to σ_B^2, then we can estimate the desired solution by solving the ITLS problem with $\alpha = \sigma_B/\sigma_E$. If $\sigma_A^2 = \sigma_B^2$, then $\alpha = 1/2$ and we have the standard TLS problem. For small values of the ratio, $\alpha \approx 1$ and we solve a problem close to LS. For large values, $\alpha \approx 0$ and we solve a problem close to DLS.

If the ratio σ_A^2/σ_B^2 is not known, then it is not clear what value of α should be used. We propose an answer to this dilemma in Section 2.5, using a method that varies α. In order to make this practical, we need an efficient algorithm for solving ITLS for multiple values of α. We develop such an algorithm in the next section.

2.3 Computing ITLS Solutions

In this section, we show that after an initial computation of the SVD of the $m \times (n+d)$ matrix $[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]$, we can compute the solution to the ITLS problem for any other value of α by working with a smaller upper-triangular matrix of dimension $(n+d) \times (n+d)$ when $m > n+d$.

2.3.1 Reduction of the Problem

Following well-known results for the standard TLS problem, as described in [90, Chap. 2-3], we begin with some notation. Define the SVD of

$$[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha] = [\alpha \boldsymbol{A}, (1-\alpha)\boldsymbol{B}] \in \mathbb{R}^{m \times (n+d)}$$

by

$$[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha] = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = [\boldsymbol{U}_1, \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{bmatrix}, \tag{2.3.1}$$

where $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$ are partitioned by $\boldsymbol{U}_1 \in \mathbb{R}^{m \times t}$, $\boldsymbol{U}_2 \in \mathbb{R}^{m \times q}$, $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{t \times t}$, $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{q \times q}$, $\boldsymbol{V}_1 \in \mathbb{R}^{(n+d) \times t}$, and $\boldsymbol{V}_2 \in \mathbb{R}^{(n+d) \times q}$, and $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n+d}] \in \mathbb{R}^{m \times (n+d)}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n+d}] \in \mathbb{R}^{(n+d) \times (n+d)}$ have orthonormal columns, $\boldsymbol{\Sigma} = \text{diag}\left(()\, \sigma_1, ..., \sigma_{n+d}\right)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n+d} \geq 0$, and t is an integer in $[0, n+d]$ such that $t + q = n + d$.

Let $\widehat{\boldsymbol{A}}_\alpha$ and $\widehat{\boldsymbol{B}}_\alpha$ denote the corrected matrices

$$\widehat{\boldsymbol{A}}_\alpha = \boldsymbol{A}_\alpha - \Delta \boldsymbol{A}_\alpha \quad \text{and} \quad \widehat{\boldsymbol{B}}_\alpha = \boldsymbol{B}_\alpha - \Delta \boldsymbol{B}_\alpha, \tag{2.3.2}$$

for some correction matrices $\Delta \boldsymbol{A}_\alpha$ and $\Delta \boldsymbol{B}_\alpha$. Define $\widehat{\boldsymbol{X}}_\alpha$ to be the TLS solution (if it exists) associated with the corrected matrices $\widehat{\boldsymbol{A}}_\alpha$ and $\widehat{\boldsymbol{B}}_\alpha$, satisfying

$$\widehat{\boldsymbol{A}}_\alpha \widehat{\boldsymbol{X}}_\alpha = \widehat{\boldsymbol{B}}_\alpha \quad \text{or} \quad [\widehat{\boldsymbol{A}}_\alpha, \widehat{\boldsymbol{B}}_\alpha] \begin{bmatrix} \widehat{\boldsymbol{X}}_\alpha \\ -\boldsymbol{I}_d \end{bmatrix} = \boldsymbol{0}. \tag{2.3.3}$$

By the Eckart-Young-Mirsky Theorem, the solution to the problem

$$\min_{\text{rank}\left([\widehat{\boldsymbol{A}}_\alpha, \widehat{\boldsymbol{B}}_\alpha]\right)=t} \|[\Delta \boldsymbol{A}_\alpha, \Delta \boldsymbol{B}_\alpha]\|_F^2 \tag{2.3.4}$$

is

$$[\widehat{\boldsymbol{A}}_\alpha, \widehat{\boldsymbol{B}}_\alpha] = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^T, \tag{2.3.5}$$

and the value of the minimization function is

$$\sum_{i=t+1}^{n+d} \sigma_i^2.$$

The corresponding correction matrix $[\Delta \boldsymbol{A}_\alpha, \Delta \boldsymbol{B}_\alpha]$ is

$$[\Delta \boldsymbol{A}_\alpha, \Delta \boldsymbol{B}_\alpha] = \boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{V}_2^T. \tag{2.3.6}$$

Because of this, the solution $\widehat{\boldsymbol{X}}_\alpha$ of (2.3.3) must satisfy

$$\text{Range}\left(\begin{bmatrix} \widehat{\boldsymbol{X}}_\alpha \\ -\boldsymbol{I}_d \end{bmatrix}\right) \subseteq \text{Null}(\boldsymbol{V}_1^T) = \text{Range}\left(\boldsymbol{V}_2\right), \tag{2.3.7}$$

by orthogonality of the right singular matrix $\boldsymbol{V}$. In order to determine an appropriate partition size t, we need to consider (i) the existence of $\widehat{\boldsymbol{X}}_\alpha$ and (ii) the noise level. We partition $\boldsymbol{V}_2$ as

$$\boldsymbol{V}_2 = \begin{bmatrix} \boldsymbol{V}_{12} \\ \boldsymbol{V}_{22} \end{bmatrix}, \tag{2.3.8}$$

where $\boldsymbol{V}_{12} \in \mathbb{R}^{n\times q}$ and $\boldsymbol{V}_{22} \in \mathbb{R}^{d\times q}$. Further, let $\widehat{t}$ denote our choice of t and $\widehat{q}$ denote the corresponding q, so that $\widehat{t} + \widehat{q} = n + d$.

First, for a given t, such a $\widehat{\boldsymbol{X}}_\alpha$ may not exist unless the block matrix $\boldsymbol{V}_{22}$ of the last d rows in the corresponding matrix $\boldsymbol{V}_2$, has column rank d. Therefore we want

$$\widehat{t} \leq t_0 \quad \text{where} \quad t_0 = \max\{t : \text{rank}\,(\boldsymbol{V}_{22}) = d\}. \tag{2.3.9}$$

Second, we would like the magnitude of the correction term to be less than a given noise tolerance ϵ :

$$||\Delta \boldsymbol{A}_\alpha, \Delta \boldsymbol{B}_\alpha||_F^2 = \sum_{i=t+1}^{n+d} \sigma_i^2 < \epsilon. \tag{2.3.10}$$

Let r be the minimal value of t satisfying the inequality above, which is called the *numerical rank.* Then we choose

$$\widehat{t} = \min(t_0, r). \tag{2.3.11}$$

Note that, if such $\widehat{t}$ is less than n, there exist infinitely many solutions $\widehat{\boldsymbol{X}}_\alpha$ satisfying (2.3.3) or (2.3.7). In this case, we can single out a minimal norm solution among these candidates.

Let $\widetilde{\boldsymbol{V}}_2 \in \mathbb{R}^{(n+d)\times q}$ denote a matrix containing an orthonormal basis for Range $(\boldsymbol{V}_2)$, and partition $\widetilde{\boldsymbol{V}}_2$ as

$$\widetilde{\boldsymbol{V}}_2 = \begin{bmatrix} \widetilde{\boldsymbol{V}}_{12} \\ \widetilde{\boldsymbol{V}}_{22} \end{bmatrix},$$

where $\widetilde{\boldsymbol{V}}_{12} \in \mathbb{R}^{n\times q}$ and $\widetilde{\boldsymbol{V}}_{22} \in \mathbb{R}^{d\times q}$. For a chosen partition size $\widehat{t}$ and $\widehat{q}$, we can compute a minimal norm solution $\widehat{\boldsymbol{X}}_\alpha$ and the correction term $[\Delta\boldsymbol{A}_\alpha, \Delta\boldsymbol{B}_\alpha]$ as

$$\widehat{\boldsymbol{X}}_\alpha = -\widetilde{\boldsymbol{V}}_{12}\widetilde{\boldsymbol{V}}_{22}^{\dagger}, \tag{2.3.12}$$

$$[\Delta\boldsymbol{A}_\alpha, \Delta\boldsymbol{B}_\alpha] = [\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]\widetilde{\boldsymbol{V}}_2\widetilde{\boldsymbol{V}}_2^T. \tag{2.3.13}$$

Thus, we can compute the minimal norm TLS solution $\widehat{\boldsymbol{X}}_\alpha$ and the corresponding correction matrix $[\Delta\boldsymbol{A}_\alpha, \Delta\boldsymbol{B}_\alpha]$ solely from $\widetilde{\boldsymbol{V}}_2$, a matrix whose column space is the partial right singular subspace, without necessarily computing the right singular matrix $\boldsymbol{V}_2$.

2.3.2 Economical Computation of $\widetilde{\boldsymbol{V}}_2$

We now consider how the basis matrix $\widetilde{\boldsymbol{V}}_2$ can be computed. Clearly we could use the standard Golub-Kahan algorithm [30] to compute the SVD of $[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]$, obtaining the basis $\widetilde{\boldsymbol{V}}_2 = \boldsymbol{V}_2$, but there are more economical alternatives when multiple values of α are of interest. For example, the rank-revealing ULV algorithm [81] can accurately compute this basis without producing the SVD, and it was used in [25] to solve the TLS problem. Other alternatives include the partial SVD method (PSVD) [90, Sec. 4.3] and the implicitly-restarted Arnoldi algorithm [57].

If $m > n + d$, it is desirable to apply one of these algorithms to a smaller matrix. For example, we could first compute the $(n+d) \times (n+d)$ upper-triangular factor $\boldsymbol{R}_\alpha$ from the QR decomposition of $[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]$. According to [11], using QR before SVD reduces the computational cost when $m > \frac{5}{3}(n + d)$.

While searching for an appropriate value of α for ITLS, we need to compute the SVD of $[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]$ for different values of α. For a new parameter value α', the new upper-triangular factor is

$$\boldsymbol{R}'_\alpha = \boldsymbol{R}_\alpha \begin{bmatrix} \left(\frac{\alpha'}{\alpha}\right)\boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \left(\frac{1-\alpha'}{1-\alpha}\right)\boldsymbol{I}_d \end{bmatrix}. \tag{2.3.14}$$

The cost of this scaling is only $O((n+d)^2)$, rather than the $O(m(n+d)^2)$ cost needed to compute the QR decomposition of $[\boldsymbol{A}_{\alpha'}, \boldsymbol{B}_{\alpha'}]$. Thus, we will compute the right singular subspace of $\boldsymbol{R}'_\alpha$ instead of $[\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha]$ for different weights α.

In Section 2.5 we propose a method for choosing an optimal value of α, and this requires computing $\widetilde{\boldsymbol{V}}_2$ for many candidate values of α. In such an algorithm, it is especially important to economize by using (2.3.14) in conjunction with an algorithm such as the PSVD.

2.4 Asymptotic Behavior

In this section we keep α fixed but let m, the number of observations, vary, so our notation will change to reflect this. We study the behavior of the ITLS problem as $m \to \infty$. Our development follows that of Gleser[1] [28] except that *we also treat the rank-deficient case*.

Let $[\widetilde{\boldsymbol{A}}_m, \widetilde{\boldsymbol{B}}_m]$ denote the true but unknown matrix, and suppose it has rank $r \leq n$. (Since the columns of $\widetilde{\boldsymbol{B}}_m$ are in the range of $\widetilde{\boldsymbol{A}}_m$, the rank cannot be greater than n.) Let $\widetilde{\boldsymbol{X}}$ denote the unique true solution if $r = n$, or the unique minimum norm true solution otherwise, so that

$$\widetilde{\boldsymbol{A}}_m \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{B}}_m. \tag{2.4.1}$$

[1] Gleser's X^T, U^T, and B^T correspond to our $[\boldsymbol{A}, \boldsymbol{B}]$, $[\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}]$, and $\widetilde{\boldsymbol{X}}$ respectively, and we set his α to zero.

Then the observed data satisfies

$$\begin{aligned}[\boldsymbol{A}_m, \boldsymbol{B}_m] &= [\widetilde{\boldsymbol{A}}_m, \widetilde{\boldsymbol{B}}_m] + [\boldsymbol{E}_{A,m}, \boldsymbol{E}_{B,m}] \\ &= \widetilde{\boldsymbol{A}}_m[\boldsymbol{I}_n, \widetilde{\boldsymbol{X}}] + [\boldsymbol{E}_{A,m}, \boldsymbol{E}_{B,m}].\end{aligned}$$

Now the matrix $\widetilde{\boldsymbol{A}}_m[\boldsymbol{I}_n, \widetilde{\boldsymbol{X}}]$ also has rank r, so $[\boldsymbol{A}_m, \boldsymbol{B}_m]$ should have $(n + d - r)$ small singular values, resulting from the perturbations $[\boldsymbol{E}_{A,m}, \boldsymbol{E}_{B,m}]$. We need some insight into the behavior of these singular values.

We impose two assumptions.

Assumption 2.1. *Each row of $[\mathbf{E}_{A,m}, \mathbf{E}_{B,m}]$ is independent and identically distributed, with zero means and covariance matrix $\sigma_\epsilon^2 \mathbf{I}_{n+d}$.*

Assumption 2.2. *The matrices $(1/m)\, \widetilde{\mathbf{A}}_m^T \widetilde{\mathbf{A}}_m$ converge to a finite limit Δ:*

$$\lim_{m\to\infty} \frac{1}{m} \widetilde{\mathbf{A}}_m^T \widetilde{\mathbf{A}}_m = \Delta. \tag{2.4.2}$$

We define

$$\begin{aligned}\boldsymbol{W}_m &= [\boldsymbol{A}_m, \boldsymbol{B}_m]^T [\boldsymbol{A}_m, \boldsymbol{B}_m], & (2.4.3) \\ \widetilde{\boldsymbol{W}}_m &= [\widetilde{\boldsymbol{A}}_m, \widetilde{\boldsymbol{B}}_m]^T [\widetilde{\boldsymbol{A}}_m, \widetilde{\boldsymbol{B}}_m] = \begin{bmatrix} \boldsymbol{I}_n \\ \widetilde{\boldsymbol{X}}^T \end{bmatrix} \widetilde{\boldsymbol{A}}_m^T \widetilde{\boldsymbol{A}}_m [\boldsymbol{I}_n, \widetilde{\boldsymbol{X}}], & (2.4.4)\end{aligned}$$

and study the convergence of these matrices.

Lemma 2.4.1. *Under Assumptions 2.1 and 2.2, both $(1/m)\widetilde{\mathbf{W}}_m$ and $(1/m)\mathbf{W}_m$ converge* *to limits:*[2]

$$\begin{aligned} \lim_{m\to\infty} \frac{1}{m}\widetilde{\mathbf{W}}_m &= \begin{bmatrix} \mathbf{I}_n \\ \widetilde{\mathbf{X}}^T \end{bmatrix} \Delta[\mathbf{I}_n, \widetilde{\mathbf{X}}] \equiv \widetilde{\mathbf{\Theta}}, & (2.4.5) \\ \operatorname*{plim}_{m\to\infty} \frac{1}{m}\mathbf{W}_m &= \sigma_\epsilon^2 \mathbf{I}_{n+d} + \widetilde{\mathbf{\Theta}} \equiv \mathbf{\Theta}. & (2.4.6) \end{aligned}$$

Proof. The first result follows from using (2.4.2) in (2.4.5). For the second, see [28, Lemma 3.1]. □

Next, we need an eigendecomposition of $\mathbf{\Theta}$ and its relation to that of $\Delta(\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)$.

Lemma 2.4.2. *Denote the eigenvalues of* $\Delta(\mathbf{I}_n + \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T)$ *by* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, *and let the columns of* $\mathbf{\Psi}$ *be the corresponding eigenvectors. Then we have an eigendecomposition of* $\mathbf{\Theta}$ *as*

$$\mathbf{\Theta}[\mathbf{V}_{\Theta_1}, \mathbf{V}_{\Theta_2}] = [\mathbf{V}_{\Theta_1}, \mathbf{V}_{\Theta_2}] \begin{bmatrix} \sigma_\epsilon^2 \mathbf{I}_n + \mathbf{D}_\lambda & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 \mathbf{I}_d \end{bmatrix},$$

where $\mathbf{D}_\lambda = \operatorname{diag}(() \lambda_1, \ldots, \lambda_n)$ *and the columns of*

$$\mathbf{V}_{\Theta_1} = \begin{bmatrix} \mathbf{I}_n \\ \widetilde{\mathbf{X}}^T \end{bmatrix} \mathbf{\Psi}, \quad \mathbf{V}_{\Theta_2} = \begin{bmatrix} -\widetilde{\mathbf{X}} \\ \mathbf{I}_d \end{bmatrix} (\mathbf{I}_d + \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-\frac{1}{2}} \tag{2.4.7}$$

are mutually orthogonal and have norm 1.

Proof. See [28, page 35]. The symmetric positive semidefinite matrix $(\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)^{\frac{1}{2}} \Delta (\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)^{\frac{1}{2}}$ has eigenvalues that are real and non-negative and has an eigenvector matrix, denoted by $(\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)^{\frac{1}{2}} \mathbf{\Psi}$, that is orthonormal:

$$\mathbf{\Psi}^T (\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T) \mathbf{\Psi} = \boldsymbol{I}_n, \tag{2.4.8}$$

[2] We denote “convergence with probability one” using the notation “plim”.

The matrix $\Delta(\boldsymbol{I}_n + \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)$ is similar to this matrix and has eigenvectors $\boldsymbol{\Psi}$.

The eigendecomposition of $\boldsymbol{\Theta}$ and the orthonormality of its eigenbasis are verified by direct computation. □

Using this eigendecomposition, we can understand the convergence of the singular values σ_i from (2.3.1).

Lemma 2.4.3. *Let* $\sigma_{1,m} \geq \sigma_{2,m} \geq \cdots \geq \sigma_{n+d,m} \geq 0$ *denote the singular values of* $[\mathbf{A}_m, \mathbf{B}_m]$. *Under Assumptions 2.1 and 2.2,*

$$\operatorname*{plim}_{m\to\infty} \frac{1}{m}\sigma_{i,m}^2 = \begin{cases} \sigma_\epsilon^2 + \lambda_i, & i = 1,\dots,n, \\ \sigma_\epsilon^2, & i = n+1,\dots,n+d. \end{cases}$$

Proof. This is a direct consequence of the definition of $\boldsymbol{W}_m$ in (2.4.3), the convergence of $(1/m)\boldsymbol{W}_m$ to $\boldsymbol{\Theta}$ (Lemma 2.4.1), and Lemma 2.4.2. □

Gleser [28, Assumption C] assumes that Δ is positive definite, but we are able to omit that assumption. We denote the rank of the symmetric positive semidefinite matrix Δ by $r \leq n$. Then $\lambda_i = 0$ for $i = r+1, ..., n$, so by Lemma 2.4.3,

$$\operatorname*{plim}_{m\to\infty} \frac{1}{m}\sigma_{i,m}^2 = \sigma_\epsilon^2 \text{ for } i = r+1, \dots, n+d. \tag{2.4.9}$$

This gives us a way to estimate σ_ϵ^2, as shown in the following lemma.

Lemma 2.4.4. *Let*

$$\widehat{\sigma}_{\epsilon,m}^2 = \frac{1}{n+d-r}\sum_{i=r+1}^{n+d} \sigma_{i,m}^2. \tag{2.4.10}$$

Under Assumptions 2.1 and 2.2,

$$\operatorname*{plim}_{m\to\infty} \frac{1}{m}\widehat{\sigma}_{\epsilon,m}^2 = \sigma_\epsilon^2.$$

Proof. This is a direct result of Lemma 2.4.3 and the fact that $\lambda_i = 0$ for $i = r + 1, ..., n$. □

In order to use $\widehat{\sigma}^2_{\epsilon,m}$ in an algorithm, we need to know that we can reliably estimate the rank r as $m \to \infty$.

Lemma 2.4.5. *Under Assumptions 2.1 and 2.2,*

$$\lim_{m\to\infty} \Pr\{\frac{1}{m}(\sigma^2_{r,m} - \sigma^2_{r+1,m}) < \frac{\lambda_r}{2}\} = 0.$$

Proof. The result follows since $(1/m)(\sigma^2_{r,m} - \sigma^2_{r+1,m})$ converges with probability one to $\lambda_r > 0$. □

With this result and (2.4.10), we see that, with appropriate choice of ϵ in (2.3.10), our rank estimation algorithm in (2.3.11) gives the correct result (with probability one) as $m \to \infty$, and from this we can establish convergence of the solution estimates, just as Gleser did in the full-rank case [28, Lemma 3.3].

Lemma 2.4.6. *Under Assumptions 2.1 and 2.2,*

$$\operatorname*{plim}_{m\to\infty} \widehat{\mathbf{X}}_m = \widetilde{\mathbf{X}}$$

where $\widetilde{\mathbf{X}}$ is the minimal norm true solution satisfying (2.4.1) and ϵ in (2.3.10) satisfies

$$m(n+d-r)\sigma^2_\epsilon \le \epsilon \le m\left((n+d-r+1)\sigma^2_\epsilon + \frac{\lambda_r}{2}\right). \tag{2.4.11}$$

Proof. With this choice of ϵ, by Lemma 2.4.5, our estimated rank converges to the true rank r with probability one. Since $(1/m)\boldsymbol{W}_m$ converges with probability one to $\boldsymbol{\Theta}$, and since there is, by Lemma 2.4.2, a gap in the spectrum of $\boldsymbol{\Theta}$, the invariant subspace corresponding to the smallest $n+d-r$ eigenvalues of $(1/m)\boldsymbol{W}_m$ converges with probability one to the span of the last $n+d-r$ columns of $\boldsymbol{V}_\theta$. Since our estimate $\hat{\boldsymbol{X}}_m$ is independent of the choice of basis for this invariant subspace, it also must converge with probability one to $\widetilde{\boldsymbol{X}}$, which, by (2.4.7), and the formula (2.3.12), is the desired minimum norm solution. □

We have now laid the groundwork for algorithms for choosing ITLS parameters. From Lemma 2.4.1, we know that the sequence of $\boldsymbol{W}$ matrices converges with probability one to $\boldsymbol{\Theta}$, and from (2.4.7) we know that $\boldsymbol{V}_{\Theta_2}$ is full rank. Therefore, our parameter t_0 in (2.3.9) converges with probability one to n, so $\widehat{t}$ in (2.3.11) converges to r. From now on, we assume, based on Lemma 2.4.4 and Lemma 2.4.5, that we have enough observations so that in (2.3.11) we have $\widehat{t}=r$, when ϵ in (2.3.10) satisfies (2.4.11).

2.5 Choice of Parameters

In this section, we propose two heuristic methods to determine the ITLS parameters based on the asymptotic convergence properties established in the previous section. We consider two cases: (1) either σ_A^2 or σ_B^2 is known, or (2) neither is known, in which case we require $n+d-r>1$.

2.5.1 Prior Information on σ_A^2 or σ_B^2

If the weight parameter α perfectly adjusts the variance of $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$, then $\boldsymbol{E}_{A_\alpha}$ and $\boldsymbol{E}_{B_\alpha}$ have identical variances, so that

$$\alpha^2\sigma_A^2 = (1-\alpha)^2\sigma_B^2 = \sigma_\epsilon^2. \tag{2.5.1}$$

By Lemma 2.4.4, $\widehat{\sigma}_\epsilon^2$ is a consistent estimate for σ_ϵ^2. Therefore, if we know σ_A^2, for example, then it is reasonable to find the α that minimizes a relative gap between $\alpha^2\sigma_A^2$ and $\widehat{\sigma}_\epsilon^2$:

$$\min_\alpha \left|\log \frac{\widehat{\sigma}_\epsilon^2}{\alpha^2\sigma_A^2}\right|. \tag{2.5.2}$$

Similarly, if we know σ_B^2, we could choose the value of α that solves the problem

$$\min_\alpha \left|\log \frac{\widehat{\sigma}_\epsilon^2}{(1-\alpha)^2\sigma_B^2}\right|. \tag{2.5.3}$$

Figure 2.1 illustrates how the estimated error variance $\widehat{\sigma}_\epsilon^2$ changes with α. The red and blue dashed lines represent the change of $\alpha^2\sigma_A^2$ and $(1-\alpha)^2\sigma_B^2$, and their intersection gives the true α and the true σ_ϵ^2, by (2.5.1). We can see that the estimate $\widehat{\sigma}_\epsilon^2$ approaches the true error variance σ_ϵ^2 as α approaches the true value, illustrating the usefulness of a choice of α based on the minimization problem (2.5.2) or (2.5.3).

In order to compute $\widehat{\sigma}_\epsilon^2$, the rank r of Δ is required. If the rank is given to us, we can immediately apply the optimization methods above. If not, we also need to estimate the rank. We examine how $\widehat{\sigma}_\epsilon^2$ and the resulting objective function values are influenced by the estimate $\widehat{r}$ of the rank. First, when $\widehat{r}$ is overestimated, we expect that the minimum value of (2.5.2) is still close to 0. This is because $\widehat{\sigma}_\epsilon^2$ is still a consistent estimator of $\alpha^2\sigma_A^2$

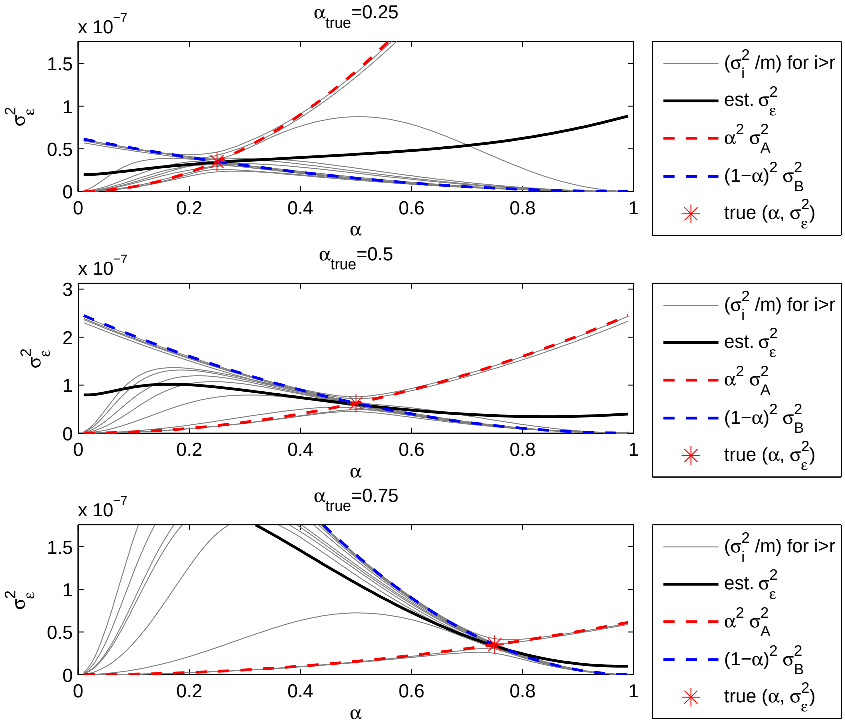


Figure 2.1: *The estimated error variance* $\widehat{\sigma}_\epsilon^2$ *as a function of* α, *for* $\alpha_{true} = 0.25, 0.5$, *and* 0.75. *The true value of* $(\alpha, \sigma_\epsilon^2)$ *is the intersection of the* $\alpha^2\sigma_A^2$ *curve (red dashed) and the* $(1-\alpha)^2\sigma_B^2$ *curve (blue dashed), marked with a star. The behavior of the small singular values as a function of* α *is traced by the grayish curves. The test problem is specified in Section 2.6, with* $m = 200$, $n = 8$, $r = 6$, $d = 10$, $\sigma_E = 0.01$.

when α is well estimated, as shown in (2.4.9). Second, if $\widehat{r}$ is underestimated, the resulting $\widehat{\sigma}_\epsilon^2$ is overwhelmed by incorrectly adding large σ_i^2 for $i < r$. From these observations, we can determine $\widehat{r}$ by solving (2.5.2), decreasing $\widehat{r}$ from n to 1. We recognize the correct rank by looking for a jump and then a plateau in the optimal objective function value. A similar argument holds for (2.5.3). In contrast to (2.5.2), though, the denominator will force the minimizer α to be lower than its true value, so looking for a jump in the optimal

α as $\hat{r}$ is changed is an alternative way to recognize an underestimated rank. We will exhibit these phenomena with sample problems in Section 2.6.

2.5.2 No Prior Information on σ_A^2 or σ_B^2

If we do not have any prior information about error variances σ_A^2 or σ_B^2, we cannot use (2.5.2) or (2.5.3). Instead, we use the convergence property (2.4.9) to evaluate a given α. Since all $(n + d - r)$ smallest singular values converge to a single constant value as the number of observations increases, we choose α to minimize their dispersion. Note that this convergence property holds only when Assumption 2.1 applies to our problem, which will be satisfied by the correct value of α. As an example, the grayish curves in Figure 2.1 show how the smallest singular values change as α varies. We can see that the singular values get closer to each other near α_{true}.

We measure the dispersion using the coefficient of variation c_v, defined as

$$c_v(\boldsymbol{y}) = \frac{\text{std}(\boldsymbol{y})}{\text{mean}(\boldsymbol{y})},$$

where mean($\boldsymbol{y}$) and std($\boldsymbol{y}$) denote the mean and standard deviation of the data vector $\boldsymbol{y}$. Thus, we choose α as the solution to

$$\min_{\alpha} \; c_v\left([\sigma_{r+1}^2(\alpha), \ldots, \sigma_{n+d}^2(\alpha)]\right). \tag{2.5.4}$$

There are other dispersion measures, such as standard deviation or variance. However, as the estimated α decreases to 0, the smallest singular values approach zero regardless of the true α, so these dispersion measures can be misleading. The coefficient of variation is dimensionless and therefore not subject to this limitation.

As in the optimization methods of Section 2.5.1, estimating c_v in (2.5.4) requires knowledge of the rank r. If the rank is not available, we can apply a similar rank-estimation strategy. For the true α, when $\widehat{r}$ is an overestimate, c_v remains acceptably small by (2.4.9). On the other hand, when $\widehat{r}$ is an underestimate, c_v grows significantly. Therefore, if we repeatedly solve (2.5.4) decreasing $\widehat{r}$ from n, we can find an appropriate $\widehat{r}$ by recognizing a jump in the corresponding value of c_v. In contrast to the rank and α estimation method of Section 2.5.1, this method requires $n + d - r > 1$, since we need at least two singular values to compute the coefficient of variation. Thus, we cannot use this method for a full-rank, single right-hand side TLS problem ($d = 1$ and $r = n$).

2.6 Experiments

We now present the results of some simple experiments exploring whether ITLS can be useful in data fitting problems. Since the "correct" choice of α depends on the error distributions for $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$, our questions are these:

- How sensitive is the solution $\boldsymbol{X}$ to the ITLS problem as α varies?

- Can the "correct" value of α be determined computationally?

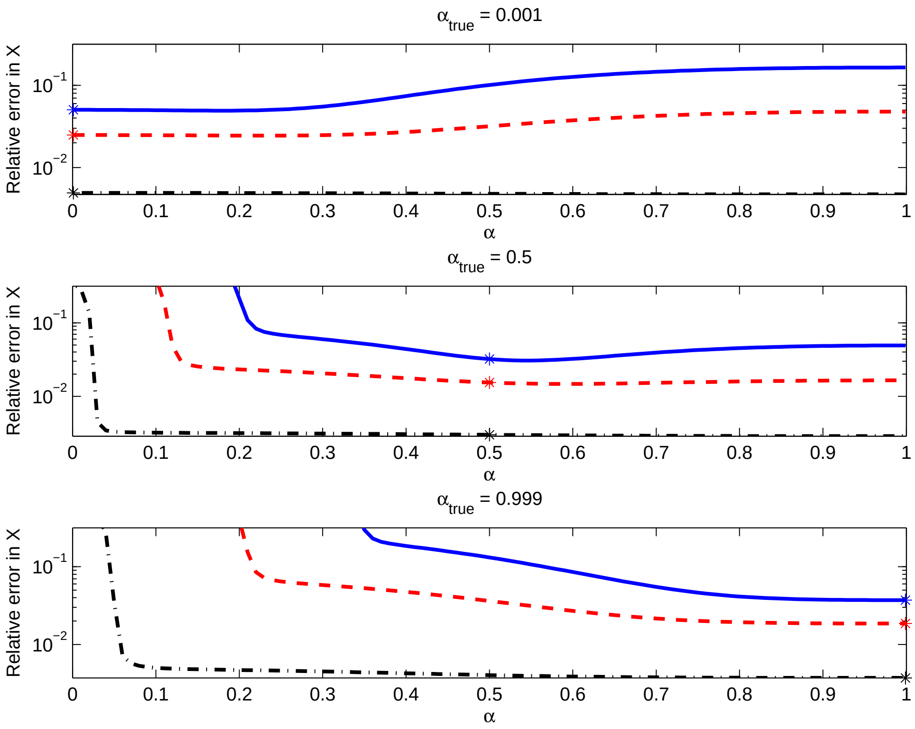


Figure 2.2: *Relative errors in* $\mathbf{X}$ *as a function of* α *for* $\alpha_{true} = 0.001, 0.5,$ *and* 0.999 *with different noise levels* σ_E*:* 0.01 *(blue solid),* 0.005 *(red dashed), and* 0.001 *(black dash-dotted). The star on each curve marks* α_{true}*.*

For given weight parameter α, rank r, and noise level σ_E, we generate a sample problem in the following way:

1. Generate $\widetilde{\boldsymbol{A}}$ and $\boldsymbol{X}$ using Matlab's `randn()`.

2. Modify $\widetilde{\boldsymbol{A}}$ to have rank r.

3. Generate $\widetilde{\boldsymbol{B}}$ as $\widetilde{\boldsymbol{B}} = \widetilde{\boldsymbol{A}}\boldsymbol{X}$.

4. Compute a minimal norm solution $\widetilde{\boldsymbol{X}}$ of $\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{B}}$.

5. Generate noise $\boldsymbol{E}_A \sim N(\boldsymbol{0}, \sigma_E^2(1-\alpha)^2\boldsymbol{I}_n)$ and $\boldsymbol{E}_B \sim N(\boldsymbol{0}, \sigma_E^2\alpha^2\boldsymbol{I}_d)$. [3]

6. Add the noise to $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{B}}$ to form $\boldsymbol{A} = \widetilde{\boldsymbol{A}} + \boldsymbol{E}_A$ and $\boldsymbol{B} = \widetilde{\boldsymbol{B}} + \boldsymbol{E}_B$.

Now, given $\boldsymbol{A}$ and $\boldsymbol{B}$, our goal is to estimate the hidden parameters (α, r, σ_E) as well as the true TLS solution $\widetilde{\boldsymbol{X}}$. Note that the noise level σ_E is related to the noise variance of σ_ϵ^2 in Assumption 2.1 by

$$\sigma_\epsilon = \alpha(1-\alpha)\sigma_E.$$

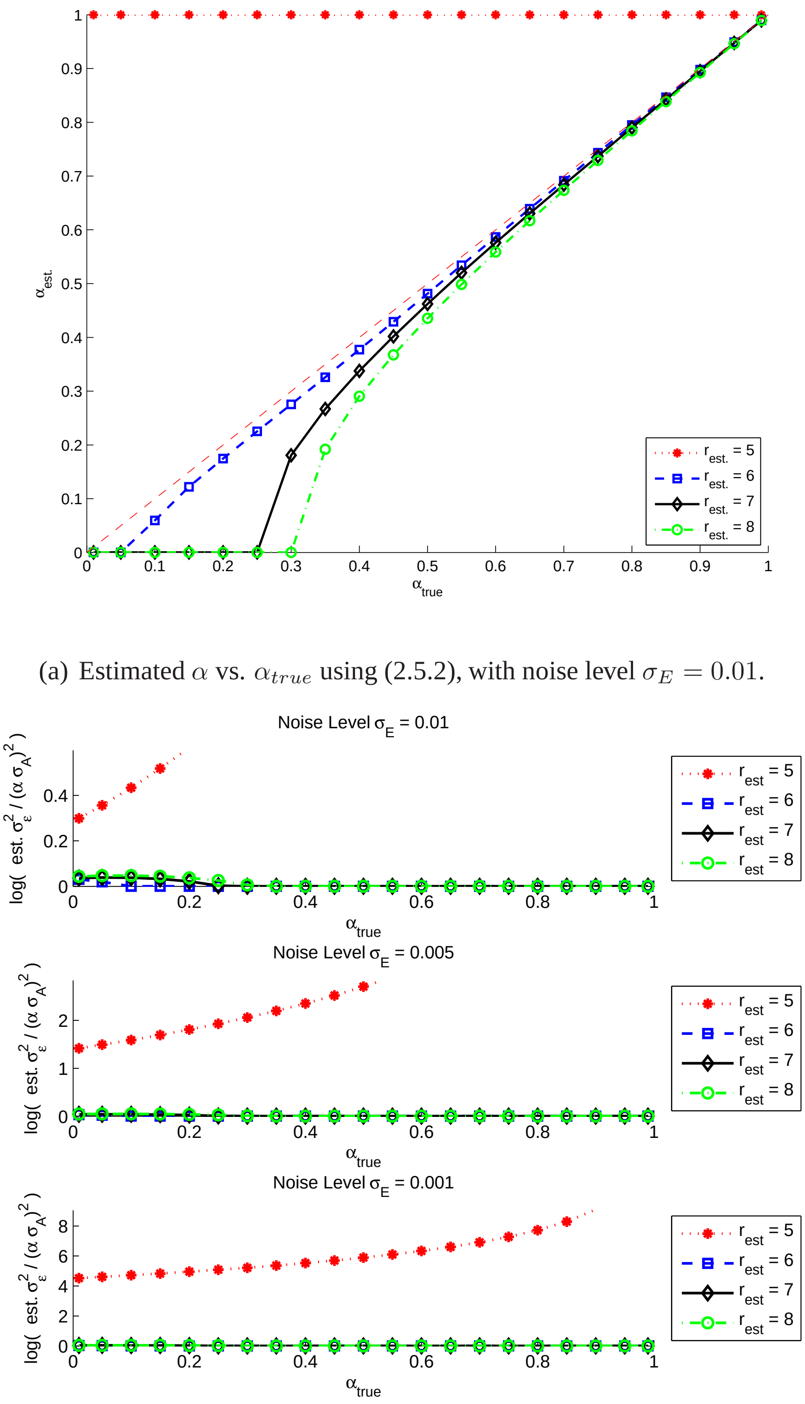
In our first experiment, we set $m = 200$, $n = 8$, $d = 4$, $r = 6$, and varied the noise level σ_E as 0.01, 0.005, and 0.001. We obtained similar results for other choices of the problem, including non-random matrices.

First, we examine the sensitivity of the TLS solution to the choice of α. Figure 2.2 plots the relative error in $\boldsymbol{X}$ as a function of α, for three different true values $\alpha_{true} = 0.001, 0.5,$ and 0.999 (which are marked by a star on the curve) with varying noise level. We can see that the sensitivity increases as the noise level increases, so the more noise, the more important it is to determine α correctly.

Next, we evaluate the performance of our methods for determining α. We apply the methods described in Section 2.5 to find a minimizer α for (2.5.2), (2.5.3), and (2.5.4), using Matlab's `fminbnd` [8], performing function evaluations using the partial SVD. The results are shown in Figures 2.3(a) - 2.5(b).

Figure 2.3(a) shows the results of estimating α when σ_A is known, using the minimizer of (2.5.2) for different values of $\widehat{r}$ with $\sigma_E = 0.01$. The estimated α approaches

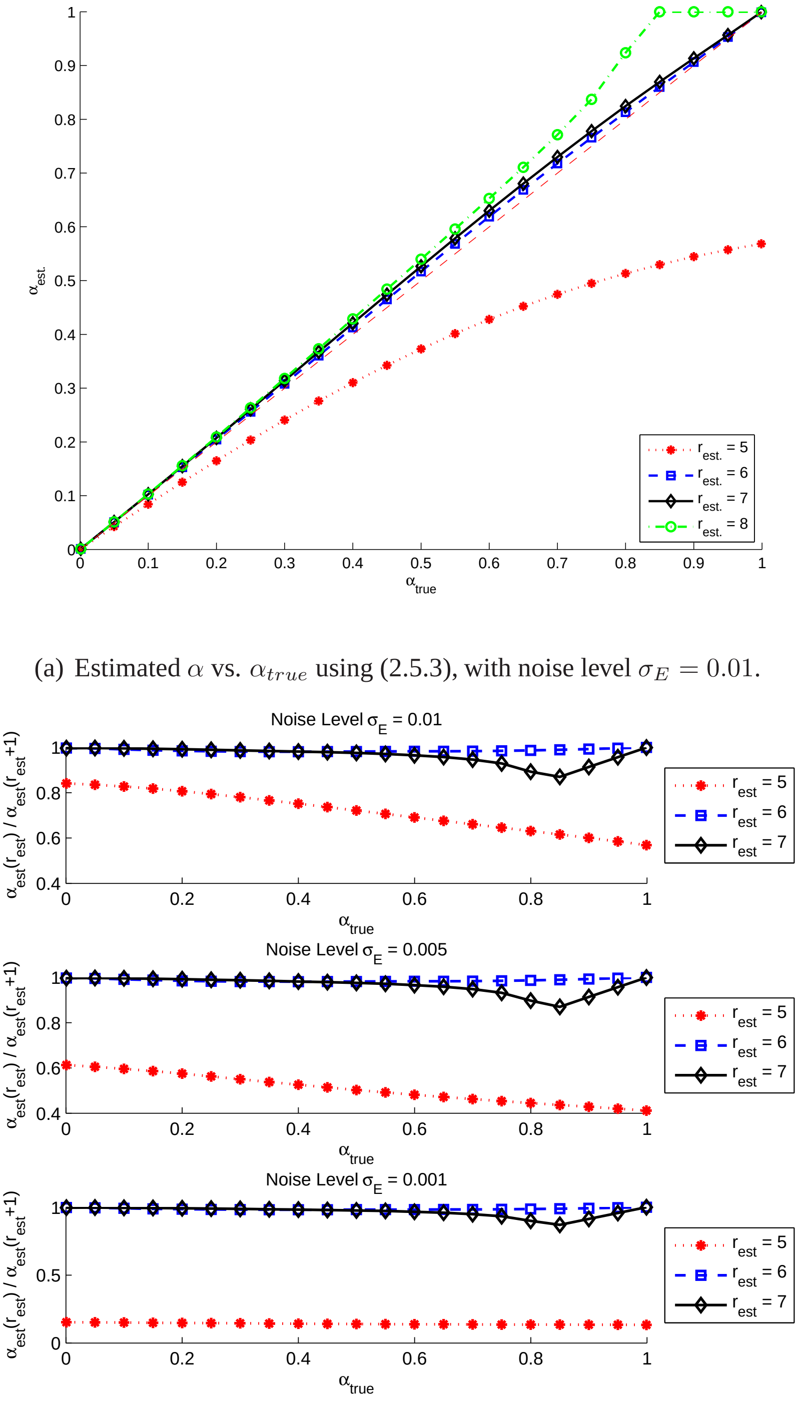
[3] Even though we generate normally-distributed errors $\boldsymbol{E}_A$ and $\boldsymbol{E}_B$ for the experiments, our methods are not restricted to a particular distribution as long as the errors are uncorrelated with identical variances.



(a) Estimated α vs. α_{true} using (2.5.2), with noise level $\sigma_E = 0.01$.

(b) Function value from (2.5.2)

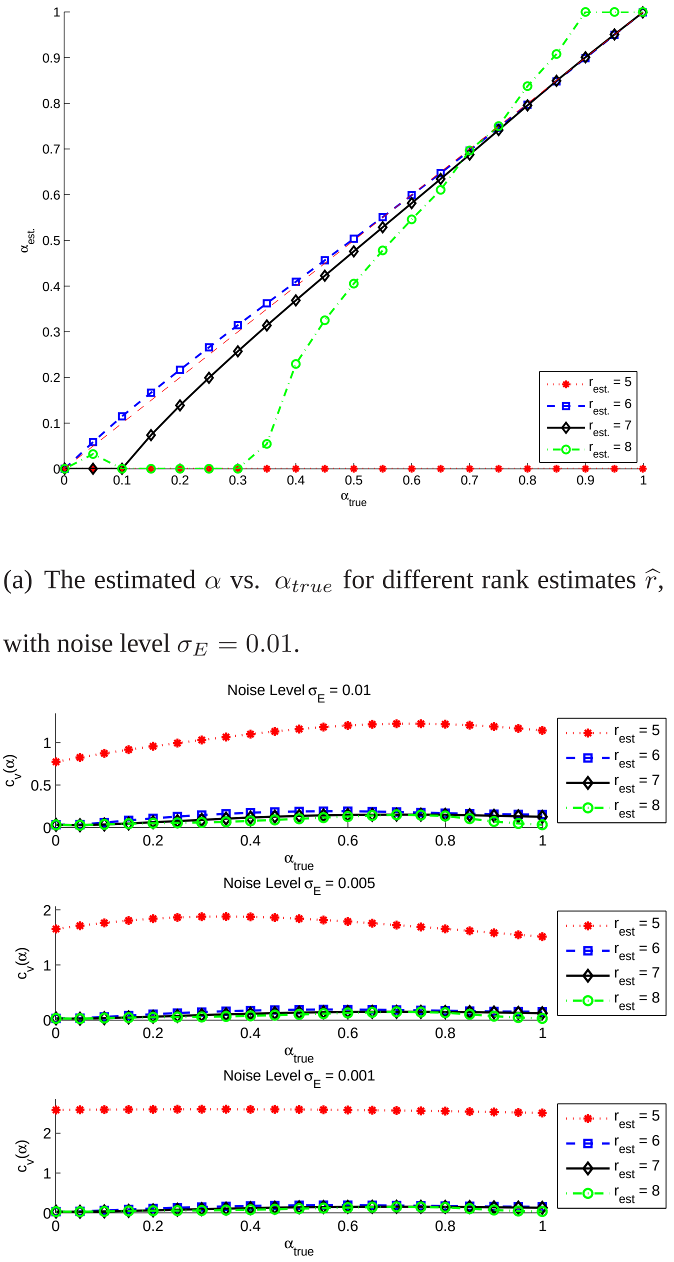
Figure 2.3: *Results when σ_A is known:* $m = 200$, $n = 8$, $d = 4$, $r = 6$.



(a) Estimated α vs. α_{true} using (2.5.3), with noise level $\sigma_E = 0.01$.

(b) Function value from (2.5.3)

Figure 2.4: *Results when σ_B is known:* $m = 200$, $n = 8$, $d = 4$, $r = 6$.



(a) The estimated α vs. α_{true} for different rank estimates $\widehat{r}$, with noise level $\sigma_E = 0.01$.

(b) The estimated coefficient of variation, c_v, for different choices of $\widehat{r}$ and noise level.

Figure 2.5: *The result from using (2.5.4) to determine* α *for* $m = 200$, $n = 8$, $d = 4$, $r = 6$.

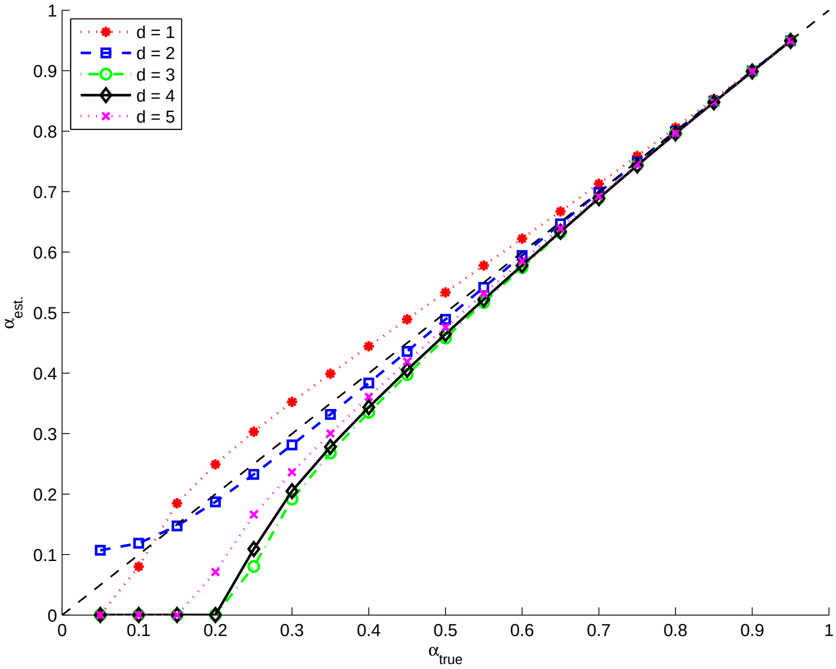


Figure 2.6: *Estimated α vs. α_{true} using the coefficient of variation for $m = 200$, $n = 8$, and $r = 7$, $\sigma_E = 0.01$, varying the number of right-hand sides d from 1 to 5.*

α_{true} as $\widehat{r}$ decreases to the true rank $r = 6$. Once the rank is underestimated, the estimated α diverges from the true α. Figure 2.3(b) shows the optimal function values for (2.5.2). The values remain close to 0 while $\widehat{r} \geq r$, but vary greatly when $\widehat{r} < r$. This phenomenon becomes more pronounced as the noise level σ_E decreases. Thus, this could be one clue to choosing an appropriate rank r when the noise level is low.

Figure 2.4(a) shows the corresponding results for (2.5.3) when σ_B is known. The α estimation is even more stable than in the previous case when $\widehat{r}$ is overestimated. Interestingly, when $\widehat{r}$ is underestimated, so is α (red dotted line). Figure 2.4(b) represents the ratio of the estimated $\alpha(\widehat{r})$ to the estimated $\alpha(\widehat{r}+1)$. The ratio stays close to 1 while $\widehat{r} \geq r$, but is much smaller when $\widehat{r} < r$. Even when the noise level is relatively high

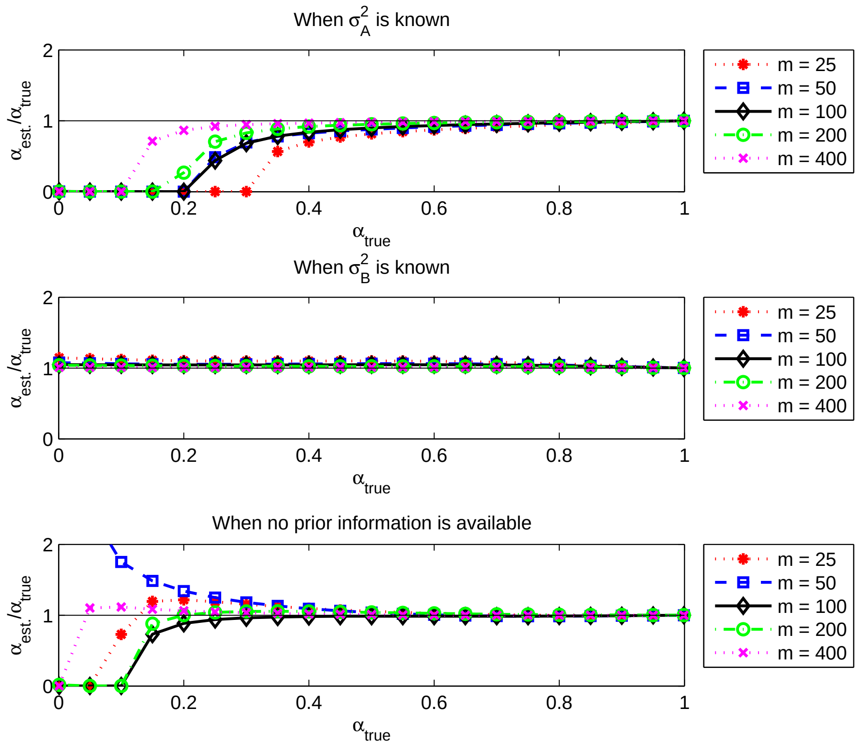


Figure 2.7: *Known rank:* α*-ratio for* $n = 8$*,* $r = 6$*, and* $d = 4$*,* $\sigma_E = 0.01$*, varying the number of observations* m*.*

($\sigma_E = 0.01$), this decrease is distinguishable, but it is larger as the noise level decreases. Therefore, this ratio of the minimizers α could be an alternative criterion to determine the rank r.

Figure 2.5(a) shows the estimated α based on (2.5.4), used when neither σ_A nor σ_B is known. Similar to the previous cases, the estimated α approaches the true α as $\widehat{r}$ approaches the true rank r from above, but the estimation of α fails when $\widehat{r} < r$. Figure 2.5(b) shows the minimized coefficient of variation, for different noise levels. While the minimized dispersion remains close to zero when $\widehat{r} \geq r$, the dispersion jumps to a large value (greater than 0.5) when $\widehat{r} < r$. The jump becomes more prominent as the noise level decreases. Hence, this is another criterion to determine the rank r. Extensive experiments revealed that rank-determination using c_v is more reliable than the other methods. Since it requires no prior information about σ_A and σ_B, we recommend using this rank-determining strategy to confirm the rank determined by other methods, whenever $n + d - r > 1$.

Next we examine the effect of sample size $(n + d - r)$ in the (2.5.4) method. We may suspect that the dispersion measure may not be reliable if $n + d - r$ is too small, so we set $m = 200$, $n = 8$, $r = 7$, and vary d from 1 to 5. Figure 2.6 shows the estimated α for different values of d. As d increases, the estimate tends to improve, but it is generally good (for moderately large values of α) even for small $n + d - r$.

Finally, we test how the number of observations m affects the estimation of α. Since all of our methods are based on an asymptotic property of the smallest singular values, we expect that increasing m should improve the quality of the estimate of α. Figure 2.7 shows the relative error in the α estimates as m varies between 25 and 400. The estimation does improve with larger m for all proposed methods, and estimation by (2.5.3) (with a known σ_B^2) shows the most reliable performance even with small m.

2.7 Discussion and Conclusions

We have defined an implicitly-weighted TLS formulation (ITLS) that includes LS, TLS, and DLS as special cases as a parameter varies between 0 and 1. We have discussed the role of the ratio between the variances of errors in $\boldsymbol{A}$ and $\boldsymbol{B}$ in choosing an appropriate parameter in ITLS. We derived asymptotic properties of the estimate as the number of observations $m \to \infty$, even when the model is rank deficient. We also proposed methods for computing the family of solutions efficiently. We developed algorithms for choosing the appropriate solution when only σ_A^2 or σ_B^2 is known, or neither is known, in which case we require $n+d-r > 1$. We provided experimental results on the usefulness of the ITLS (or, equivalently the STLS) family of solutions, and on our algorithms for estimating α and r.

It would be easy to add a regularization term to the ITLS problem, in order to handle discrete ill-posed problems.

This work leaves two important open questions. First, the concept of a *core problem* [39, 64, 70], so useful for a single right-hand side, does not completely explain the character of TLS problems when $d > 1$, and more work is needed. This is related to the choice of $\hat{t}$. Second, our parameter choice algorithm requires an estimate of either σ_A or σ_B when $n + d - r = 1$, a single right-hand-side problem with full rank, so more work on that case is needed.

Chapter 3

Portfolio Selection Using Tikhonov Filtering to Estimate the Covariance Matrix

Markowitz's portfolio selection problem chooses weights for stocks in a portfolio based on an estimated covariance matrix for stock returns. Since the performance of the resulting portfolio is very sensitive to the quality of the covariance matrix, its estimation is very critical for the portfolio selection to be successful. A conventional sample covariance matrix is not a good estimate since it takes all transient information and observation noise as important factors. Matrix reduction on the covariance matrix removes the unsystematic factors generated by the noise.

Our study proposes to reduce noise in the estimation using a Tikhonov filter function. In addition, we prevent rank deficiency of the estimated covariance matrix and propose a method for effectively choosing the Tikhonov parameter, which determines the filtering intensity. We put previous estimators into a common framework and compare their filtering functions for eigenvalues of the correlation matrix. We demonstrate the effectiveness of our estimator using stock return data from 1958 through 2007. This presentation closely follows that in [66].

3.1 Introduction

A stock investor might want to construct a portfolio of stocks whose return has a small variance, because large variance implies high risk. Given a target portfolio return q, a mean-variance problem (MV) [59] finds a stock weight vector $\boldsymbol{w}$ to determine a portfolio that minimizes the variance of the return. Let $\boldsymbol{\mu}$ be a vector of expected returns for each of N stocks, and let $\boldsymbol{\Sigma}$ be an $N \times N$ covariance matrix for the returns. The problem can be written as

$$\min_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \text{ subject to } \boldsymbol{w}^T \mathbf{1} = 1, \quad \boldsymbol{w}^T \boldsymbol{\mu} = q, \tag{3.1.1}$$

where $\mathbf{1}$ is a vector of N ones. On the other hand, a global minimum variance problem (GMV) finds a portfolio that minimizes the variances of the portfolio returns without the return constraint:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \text{ subject to } \boldsymbol{w}^T \mathbf{1} = 1. \tag{3.1.2}$$

Even though these optimization problems play a central role in a modern portfolio theory, it has been observed that the solutions are very sensitive to their input parameters [6, 10, 12, 13]. Thus, in order to construct a good portfolio using these formulations, the covariance matrix Σ must be well-estimated. We let $\widetilde{\Sigma}$ denote an estimate of Σ, and $\widetilde{\Sigma}_{method}$ denote a resulting estimate by a particular method.

Let $\boldsymbol{R} = [\boldsymbol{r}(1), \cdots, \boldsymbol{r}(T)]$ be an $N \times T$ matrix containing observations on N stocks' returns for each of T times. A conventional estimator – a sample covariance matrix $\widetilde{\Sigma}_{sample}$ – can be computed from the stock return matrix $\boldsymbol{R}$ as

$$\widetilde{\Sigma}_{sample} = \frac{1}{T}\boldsymbol{R}(\boldsymbol{I}_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T)\boldsymbol{R}^T. \tag{3.1.3}$$

From classical statistics, $\widetilde{\Sigma}_{sample}$ is a consistent estimate for fixed N; in our case, since T is fixed and of the same order as N, this result is not so useful. Moreover, since the stock return matrix $\boldsymbol{R}$ contains noise, the sample covariance matrix $\widetilde{\Sigma}_{sample}$ might not estimate the true covariance matrix well. We use principal component analysis and reduce the noise in the covariance matrix estimate by using a Tikhonov regularization method. We demonstrate experimentally that this improves the portfolio weight $\boldsymbol{w}$ obtained from (3.1.2).

Our study is closely related to factor analysis and principal component analysis, which were previously applied to explain interdependency of stock returns and classify the securities into appropriate subgroups. Sharpe [79] first proposed a single-factor model in this context using market returns. King [49] analyzed stock behaviors with both multiple factors and multiple principal components. These factor models established a basis for the asset pricing models CAPM [58, 62, 80, 87] and APT [73, 74].

There have been previous efforts, which we discuss in detail later in this chapter, to improve the estimate of Σ. Sharpe [79] proposed a market-index covariance matrix

$\widetilde{\boldsymbol{\Sigma}}_{market}$ derived from a single-factor model of market returns. Ledoit et al. [55] introduced a shrinkage method that averages $\widetilde{\boldsymbol{\Sigma}}_{sample}$ and $\widetilde{\boldsymbol{\Sigma}}_{market}$. They [56] also applied the shrinkage method with a different target, an identity matrix. Later, it was shown by DeMiguel et al. [19] that their shrinkage methods have the same effect as adding the constraint $||\boldsymbol{w}||_{\boldsymbol{A}} \leq \delta$ to the GMV problem (3.1.2), where $\boldsymbol{A}$ is the shrinkage target matrix ($\widetilde{\boldsymbol{\Sigma}}_{market}$ or $\boldsymbol{I}_N$) and δ is a given threshold. Elton and Gruber [24] estimated $\boldsymbol{\Sigma}$ using a few principal components from a correlation matrix. More recently, Plerou et al. [69], Laloux et al. [53], Conlon et al. [14], and Kwapień [52] applied random matrix theory [60] to this problem. They found that most eigenvalues of correlation matrices from stock return data lie within the bound for a random correlation matrix and hypothesized that eigencomponents (principal components) outside this interval contain true information. Bengtsson and Holst [5] generalized the approach of Ledoit et al. [55] by damping all but the k largest eigenvalues by a single rate. In summary, the estimator of Sharpe [79] uses $\widetilde{\boldsymbol{\Sigma}}_{market}$, the estimator of Ledoit et al. [55, 56] takes the weighted average of $\widetilde{\boldsymbol{\Sigma}}_{sample}$ and different target matrices, the estimator of Elton and Gruber [24] truncates the smallest eigenvalues, the estimators of Plerou et al. [69], Laloux et al. [53], Conlon et al. [14], and Kwapień [52] adjust principal components in some interval, and the estimator of Bengtsson and Holst [5] attenuates the smallest eigenvalues by a single rate.

Jagannathan and Ma [44] showed that a short-sale constraint ($w \geq 0$) is equivalent to shrinking the input covariance matrix $\boldsymbol{\Sigma}$ by subtracting $(\boldsymbol{\lambda}\mathbf{1}^T + \mathbf{1}^T\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers for the constraints. DeMiguel et al. [19] showed that adding the short-sale constraint to GMV is equivalent to adding a 1-norm constraint $||\boldsymbol{w}||_1 \leq 1$, and generalized this constraint to $||\boldsymbol{w}||_1 \leq \delta$ for a certain threshold δ which determines a short-sale budget.

Our study focuses on estimating a good covariance matrix. We propose to decrease the contribution of the smaller eigenvalues of a correlation matrix gradually by using a *Tikhonov filtering function*. To derive the Tikhonov filtering, we construct a linear model based on principal component analysis and formulate an optimization problem that finds appropriately noise-filtered factors. Using the filtered factor data, we estimate a Tikhonov covariance matrix.

In Section 3.2, we introduce Tikhonov regularization to reduce noise in the stock return data. In Section 3.3, we show that applying Tikhonov regularization results in filtering the eigenvalues of the correlation matrix for the stock returns. In Section 3.4, we discuss how we can choose a Tikhonov parameter that determines the intensity of Tikhonov filtering. In Section 3.5, we put all of the factor-based estimators into a common framework, and compare the characteristics of their filtering functions for the eigenvalues of the correlation matrix. In Section 3.6, we show the results of numerical experiments comparing the covariance estimators for portfolio construction using monthly return data of 100 randomly chosen stocks from the CRSP. In Section 3.7, we highlight the differences between Tikhonov filtering and the other methods.

3.2 Tikhonov Filtering

To estimate the covariance matrix, we apply a principal component analysis to find an orthogonal basis that maximizes the variance of the projected data into the basis. Based on the analysis, we use the Tikhonov regularization method to filter out the noise from the data. Next, we explain the feature of gradual down-weighting, which is the key difference between Tikhonov filtering and other methods.

3.2.1 Principal Component Analysis

First, we establish some notation. For a random process $\boldsymbol{x}(t)$, let $\mathbb{E}[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times 1}$, $\text{var}[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times 1}$, $\text{cov}[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times N}$, and $\text{corr}[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times N}$ denote a mean, a variance, a covariance matrix, and a correlation matrix. For a given collection of observations $\boldsymbol{X} = [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)]$ for N objects during T times, let $\mathbb{E}_s[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times 1}$, $\text{var}_s[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times 1}$, $\text{cov}_s[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times N}$, and $\text{corr}_s[\boldsymbol{x}(t)] \in \mathbb{R}^{N\times N}$ denote the corresponding sample statistics, defined, for example, in [37, Section 3.3].

Now we apply principal component analysis (PCA)[1] to the stock return data $\boldsymbol{R}$. Let $\boldsymbol{Z} = [\boldsymbol{z}(1), \ldots, \boldsymbol{z}(T)]$ be an $N \times T$ matrix of normalized stock returns derived from $\boldsymbol{R}$, defined so that

$$\mathbb{E}_s[\boldsymbol{z}(t)] = \boldsymbol{0}, \quad \text{var}_s[\boldsymbol{z}(t)] = \boldsymbol{1}, \tag{3.2.1}$$

where $\boldsymbol{0}$ is a vector of N zeros. We can compute $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \boldsymbol{D}_V^{-\frac{1}{2}}(\boldsymbol{R} - \frac{1}{T}\boldsymbol{R}\,\boldsymbol{1}\boldsymbol{1}^T), \tag{3.2.2}$$

where $\boldsymbol{D}_V = \text{diag}\,(\text{var}_s[\boldsymbol{r}(t)]) \in \mathbb{R}^{N\times N}$ is a diagonal matrix containing the N sample variances for the N stock returns. By using the normalized stock return matrix $\boldsymbol{Z}$ rather

[1] In this chapter, the term PCA always refers to applying PCA to the matrix $\boldsymbol{R}$ of sample stock returns. For convergence properties of the sample PCA toward its population PCA, refer to [43, Chapter 4].

than $\boldsymbol{R}$, we can make the PCA independent of the different variance of each stock return [43, pp.64-66].

PCA finds an orthogonal basis $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{N\times k}$ for $\boldsymbol{Z}$ where $k = \text{rank}\,(\boldsymbol{Z})$. Each basis vector $\boldsymbol{u}_i$ maximizes the variance of the projected data $\boldsymbol{u}_i^T\boldsymbol{Z}$, while maintaining orthogonality to all the preceding basis vectors $\boldsymbol{u}_j$ ($j < i$). By PCA, we can represent the given data $\boldsymbol{Z} = [\boldsymbol{z}(1), \ldots, \boldsymbol{z}(T)]$ as

$$\boldsymbol{Z} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]\,\boldsymbol{F} = \boldsymbol{U}\boldsymbol{F}, \tag{3.2.3}$$

$$\boldsymbol{z}(t) = \boldsymbol{U}\boldsymbol{f}(t) = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]\,\boldsymbol{f}(t) = \sum_{i=1}^{k} f_i(t)\boldsymbol{u}_i, \tag{3.2.4}$$

where $\boldsymbol{f}(t) = [f_1(t), \ldots, f_k(t)]^T$, a column of $\boldsymbol{F}$, is the projected data at time t, and $\text{var}_s[f_1(t)] \geq \text{var}_s[f_2(t)] \geq \cdots \geq \text{var}_s[f_k(t)]$. The projected data $f_i(t)$ is called the i-th principal component in PCA or the i-th factor in the factor analysis. Larger $\text{var}_s[f_i(t)]$ implies that the corresponding $\boldsymbol{f}_i(t)$ plays a more important role in representing $\boldsymbol{Z}$. The orthogonal basis $\boldsymbol{U}$ and the projected data $\boldsymbol{F}$ can be obtained by the singular value decomposition (SVD) of $\boldsymbol{Z}$,

$$\boldsymbol{Z} = \boldsymbol{U}_k\,\boldsymbol{S}_k\,\boldsymbol{V}_k^T, \tag{3.2.5}$$

where k is the rank of Z,

$\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{N\times k}$ is a matrix of left orthogonal singular vectors,

$\boldsymbol{S}_k = \text{diag}\,(s_1, \ldots, s_k) \in \mathbb{R}^{k\times k}$ is a diagonal matrix of singular values s_i,

and $\boldsymbol{V}_k = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{T\times k}$ is a matrix of right orthogonal singular vectors.

In PCA, the orthogonal basis matrix $\boldsymbol{U}$ corresponds to $\boldsymbol{U}_k$, and the projected data $\boldsymbol{F}$ corresponds to $(\boldsymbol{S}_k\boldsymbol{V}_k^T)$ [43, p.193]. Moreover, the variance of the projected data $f_i(t)$ is proportional to the square of singular value s_i^2 as we now show. $\mathbb{E}_s[\boldsymbol{z}(t)] = \mathbf{0}$ means that $\boldsymbol{Z}\mathbf{1} = \mathbf{0}$. Therefore, since $\boldsymbol{z}(t) = \boldsymbol{U}\boldsymbol{f}(t)$,

$$\mathbb{E}_s[\boldsymbol{f}(t)] = \boldsymbol{U}^T\boldsymbol{Z}\mathbf{1} = \mathbf{0}, \tag{3.2.6}$$

so $\boldsymbol{f}(t)$ also has zero-mean. Therefore,

$$\text{var}_s[f_i(t)] = \frac{1}{T}\sum_{t=1}^{T}\left(f_i(t) - \mathbb{E}_s[f_i(t)]\right)^2 = \frac{1}{T}\sum_{t=1}^{T} f_i^2(t).$$

Since $\boldsymbol{F}$ is equal to $\boldsymbol{S}_k\boldsymbol{V}_k^T$,

$$f_i(t) = s_i v_i(t), \tag{3.2.7}$$

where $v_i(t)$ is the (t, i) element of $\boldsymbol{V}_k$. Thus,

$$\text{var}_s[f_i(t)] = \frac{1}{T}\sum_{t=1}^{T}(s_i v_i(t))^2 = \frac{1}{T}s_i^2(\boldsymbol{v}_i^T\boldsymbol{v}_i) = \frac{s_i^2}{T}, \tag{3.2.8}$$

by the orthonormality of $\boldsymbol{v}_i$. Thus, the singular value s_i determines the magnitude of $\text{var}_s[f_i(t)]$, so it measures the contribution of the projected data $f_i(t)$ to $\boldsymbol{z}(t)$.

3.2.2 Tikhonov Regularization

$\boldsymbol{U}$ and $\boldsymbol{f}(t)$ in (3.2.4) form a linear model with a k–dimensional orthogonal basis for the normalized stock return $\boldsymbol{Z}$, where $k = \text{rank}(\boldsymbol{Z})$. As mentioned in the previous section, the singular value s_i determines how much the principal component $f_i(t)$ contributes to $\boldsymbol{z}(t)$. However, since noise is included in $\boldsymbol{z}(t)$, the k–dimensional model is overfitted, containing unimportant principal components possibly corresponding to the noise. We use a Tikhonov regularization method [67, 83, 89], sometimes called ridge regression [40, 41], to reduce the contribution of unimportant principal components to the normalized stock return $\boldsymbol{Z}$. Eventually, we construct a filtered principal component $\widetilde{\boldsymbol{f}}(t)$ and a filtered market return $\widetilde{\boldsymbol{Z}}$.

Originally, regularization methods were developed to reduce the influence of noise when solving a discrete ill-posed problem $\boldsymbol{b} \approx \boldsymbol{Ax}$, where the $M \times N$ matrix $\boldsymbol{A}$ has some singular values close to 0 [34, pp.71-86]. If we write the SVD of $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\,\boldsymbol{V}^T = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N] \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_N \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_N^T \end{bmatrix},$$

then the minimum norm least square solution $\boldsymbol{x}_{LS}$ to $\boldsymbol{b} \approx \boldsymbol{Af}$ is

$$\boldsymbol{x}_{LS} = \boldsymbol{A}^{\dagger}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{S}^{\dagger}\,\boldsymbol{U}^T\boldsymbol{b} = \sum_{i=1}^{\text{rank}(\boldsymbol{A})} \frac{\boldsymbol{u}_i^T\boldsymbol{b}}{s_i}\boldsymbol{v}_i. \tag{3.2.9}$$

If $\boldsymbol{A}$ has some small singular values, then $\boldsymbol{x}_{LS}$ is dominated by the corresponding singular vectors $\boldsymbol{v}_i$. Two popular methods are used for regularization to reduce the influence of components $\boldsymbol{v}_i$ corresponding to small singular values: a truncated SVD method (TSVD) [30, 36] and a Tikhonov method [83]. Briefly speaking, the TSVD simply truncates terms in (3.2.9) corresponding to singular values close to 0. In contrast, Tikhonov regularization solves the least squares problem

$$\min_{\boldsymbol{f}} ||\boldsymbol{b} - \boldsymbol{Ax}||^2 + \alpha^2||\boldsymbol{Px}||^2, \tag{3.2.10}$$

where α and $\boldsymbol{P}$ are predetermined. The penalty term $||\boldsymbol{Px}||^2$ restricts the magnitude of the solution $\boldsymbol{x}$ so that the effects of small singular values are reduced.

Returning to our original problem, we use regularization in order to filter out the noise from the principal component $\boldsymbol{f}(t)$. We formulate the linear problem to find a filtered principal component $\widetilde{\boldsymbol{f}}(t)$ as

$$\widetilde{\boldsymbol{z}}(t) = \boldsymbol{U}\widetilde{\boldsymbol{f}}(t), \tag{3.2.11}$$

$$\boldsymbol{z}(t) = \widetilde{\boldsymbol{z}}(t) + \boldsymbol{\epsilon}_z(t) = \boldsymbol{U}\widetilde{\boldsymbol{f}}(t) + \boldsymbol{\epsilon}_z(t), \tag{3.2.12}$$

where $\widetilde{\boldsymbol{z}}(t)$ is the resulting filtered data and $\boldsymbol{\epsilon}_z(t)$ is the extracted noise. In (3.2.4), $\boldsymbol{f}(t)$ is the exact solution of (3.2.12) when $\boldsymbol{\epsilon}_z(t) = 0$. By (3.2.7), we can express $\boldsymbol{f}(t)$ as

$$\boldsymbol{f}(t) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_k(t) \end{bmatrix} = \begin{bmatrix} s_1\ v_1(t) \\ \vdots \\ s_k v_k(t) \end{bmatrix} = \sum_{i=1}^{k} (s_i v_i(t))\boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is the i-th column of the identity matrix. Since we expect that the unimportant principal components $f_i(t)$ are more contaminated by the noise, we reduce the contribution of these principal components. We apply a filtering matrix $\boldsymbol{\Phi} = \text{diag}\,(\phi_1, \ldots, \phi_k)$ to $\boldsymbol{f}(t)$ with each $\phi_i \in [0, 1]$ so that

$$\widetilde{\boldsymbol{f}}(t) = \boldsymbol{\Phi}\,\boldsymbol{f}(t).$$

The element ϕ_i should be small when s_i is small. The resulting filtered data are

$$\widetilde{\boldsymbol{z}}(t) = \boldsymbol{U}\,\boldsymbol{\Phi}\,\boldsymbol{f}(t), \tag{3.2.13}$$

$$\widetilde{\boldsymbol{Z}} = \boldsymbol{U}\,\boldsymbol{\Phi}\boldsymbol{F}. \tag{3.2.14}$$

We introduce two different filtering matrices, $\boldsymbol{\Phi}_{trun}(\widehat{k})$ and $\boldsymbol{\Phi}_{tikh}(\alpha)$, which correspond to truncated SVD and Tikhonov regularization. First, we can simply truncate all but $\widehat{k}$ most important components as Elton and Gruber [24] did by using a filtering matrix of $\mathbf{\Phi}_{trun}(\widehat{k}) = \text{diag}\left(\underbrace{1,\ldots,1}_{\widehat{k}},\underbrace{0,\ldots,0}_{k-\widehat{k}}\right)$, so the truncated principal component $\widetilde{\boldsymbol{f}}_{trun}(t)$ is

$$\widetilde{\boldsymbol{f}}_{trun}(t) = \mathbf{\Phi}_{trun}(\widehat{k})\boldsymbol{f}(t).$$

By (3.2.13) and (3.2.14), the resulting filtered data are $\widetilde{\boldsymbol{z}}_{trun}(t) = \boldsymbol{U}\mathbf{\Phi}_{trun}(\widehat{k})\boldsymbol{f}(t)$ and $\widetilde{\boldsymbol{Z}}_{trun} = \boldsymbol{U}\,\mathbf{\Phi}_{trun}(\widehat{k})\boldsymbol{F}$. Since $\boldsymbol{F} = \boldsymbol{S}_{\,k}\boldsymbol{V}_k^T$, we can rewrite $\widetilde{\boldsymbol{Z}}_{trun}$ as

$$\widetilde{\boldsymbol{Z}}_{trun} = \boldsymbol{U}\,\mathbf{\Phi}_{trun}(\widehat{k})(\boldsymbol{S}_{\,k}\boldsymbol{V}_k^T) = \sum_{i=1}^{\widehat{k}} s_i\boldsymbol{u}_i\boldsymbol{v}_i^T. \tag{3.2.15}$$

From (3.2.15), we can see that this truncation method corresponds to the truncated SVD regularization (TSVD) [30, 36].

Second, we can apply the Tikhonov method, and this is our approach to estimating the covariance matrix. We formulate the regularized least squares problem to solve (3.2.10) as

$$\min_{\widetilde{\boldsymbol{f}}(t)} M(\widetilde{\boldsymbol{f}}(t)) \tag{3.2.16}$$

with

$$M(\widetilde{\boldsymbol{f}}(t)) = ||\boldsymbol{z}(t) - \boldsymbol{U}\widetilde{\boldsymbol{f}}(t)||^2 + \alpha^2||\boldsymbol{P}\widetilde{\boldsymbol{f}}(t)||^2,$$

where α^2 is a penalty parameter and $\boldsymbol{P}$ is a penalty matrix. The first term $||\boldsymbol{z}(t) - \boldsymbol{U}\widetilde{\boldsymbol{f}}(t)||^2$ forces $\widetilde{\boldsymbol{f}}(t)$ to be close to the exact solution $\boldsymbol{f}(t)$. The second term $||\boldsymbol{P}\widetilde{\boldsymbol{f}}(t)||^2$ controls the size of $\widetilde{\boldsymbol{f}}(t)$. We can choose, for example,

$$\boldsymbol{P} = \text{diag}\left(s_1^{-1},\ldots,s_k^{-1}\right).$$

Let $\widetilde{f}_i(t)$ denote the i-th element of $\widetilde{\boldsymbol{f}}(t)$. The matrix $\boldsymbol{P}$ scales each $\widetilde{f}_i(t)$ by s_i^{-1}, so the unimportant principal components corresponding to small s_i are penalized more than the more important principal components, since we expect that the unimportant principal components $f_i(t)$ are more contaminated by the noise. Thus, the penalty term prevents $\widetilde{\boldsymbol{f}}(t)$ from containing large amounts of unimportant principal components. As we showed before, s_i^2 is proportional to the variance of the i-th principal component $f_i(t)$. Therefore, this penalty matrix $\boldsymbol{P}$ is statistically meaningful considering that the values of $\widetilde{f}_i(t)/s_i$ in $\boldsymbol{P}\widetilde{\boldsymbol{f}}(t)$ are in proportion to the normalized principal components $\widetilde{f}_i(t)/\sqrt{\text{var}_s[f_i(t)]}$.

The penalty parameter α balances the minimization between the error term $||\boldsymbol{z}(t) - \boldsymbol{U}\widetilde{\boldsymbol{f}}(t)||^2$ and the penalty term $||\boldsymbol{P}\widetilde{\boldsymbol{f}}(t)||^2$. Therefore, as α increases, the regularized solution $\widetilde{\boldsymbol{f}}(t)$ moves away from the exact solution $\boldsymbol{f}(t)$ but should discard more of $\boldsymbol{f}(t)$ as noise. We can quantify this property by determining the solution to (3.2.16). At the minimizer of (3.2.16), the gradient of $M(\widetilde{\boldsymbol{f}}(t))$ with respect to each $\widetilde{f}_i(t)$ becomes zero, so

$$\nabla M(\widetilde{\boldsymbol{f}}(t)) = 2\boldsymbol{U}^T\boldsymbol{U}\widetilde{\boldsymbol{f}}(t) - 2\boldsymbol{U}^T\boldsymbol{z}(t) + 2\alpha^2\boldsymbol{P}^T\boldsymbol{P}\widetilde{\boldsymbol{f}}(t) = 0,$$

and thus

$$(\boldsymbol{U}^T\boldsymbol{U} + \alpha^2\boldsymbol{P}^T\boldsymbol{P})\widetilde{\boldsymbol{f}}(t) = \boldsymbol{U}^T\boldsymbol{z}(t).$$

Since $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k$, $\boldsymbol{P} = \text{diag}\left(s_1^{-1}, \ldots, s_k^{-1}\right)$, and $\boldsymbol{z}(t) = \boldsymbol{U}\boldsymbol{f}(t)$, this becomes

$$\left(\boldsymbol{I}_k + \alpha^2\text{diag}\left(s_1^{-2}, \ldots, s_k^{-2}\right)\right)\widetilde{\boldsymbol{f}}(t) = \boldsymbol{U}^T\left(\boldsymbol{U}\boldsymbol{f}(t)\right).$$

Therefore,

$$\text{diag}\left(\frac{s_1^2 + \alpha^2}{s_1^2}, \ldots, \frac{s_k^2 + \alpha^2}{s_k^2}\right)\widetilde{\boldsymbol{f}}(t) = \boldsymbol{f}(t),$$

and

$$\widetilde{\boldsymbol{f}}(t) = \text{diag}\left(\frac{s_1^2}{s_1^2 + \alpha^2}, \ldots, \frac{s_k^2}{s_k^2 + \alpha^2}\right)\boldsymbol{f}(t).$$

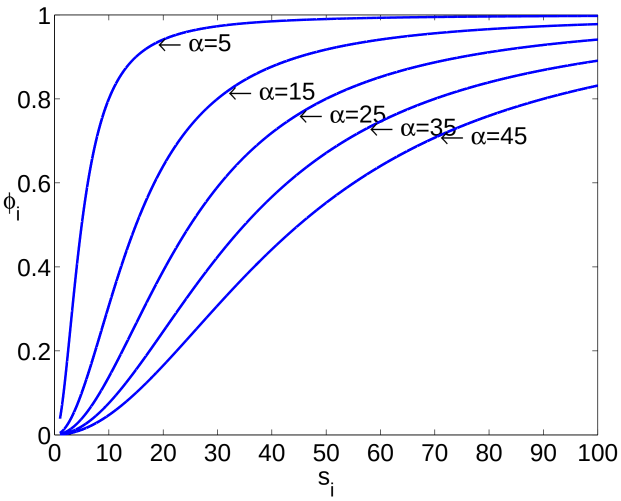


Figure 3.1: *Tikhonov filtering as a function of s_i for various values of α.*

So, our Tikhonov estimate is

$$\widetilde{\boldsymbol{f}}_{tikh}(t) = \boldsymbol{\Phi}_{tikh}(\alpha)\,\boldsymbol{f}(t),$$

where $\boldsymbol{\Phi}_{tikh}(\alpha)$, called the Tikhonov filtering matrix, denotes $(\boldsymbol{S}_k^2 + \alpha^2\boldsymbol{I}_k)^{-1}\boldsymbol{S}_k^2$. Thus, we can see that the regularized principal component $\widetilde{\boldsymbol{f}}_{tikh}(t)$ is the result after filtering the original principal component $\boldsymbol{f}(t)$ with the diagonal matrix $\boldsymbol{\Phi}_{tikh}(\alpha)$, whose diagonal elements $\phi_i^{tikh}(\alpha) = \dfrac{s_i^2}{s_i^2 + \alpha^2}$ lie in $[0, 1]$. By (3.2.13) and (3.2.14), the resulting filtered data become $\widetilde{\boldsymbol{z}}_{tikh}(t) = \boldsymbol{U}\boldsymbol{\Phi}_{tikh}(\alpha)\,\boldsymbol{f}(t)$ and $\widetilde{\boldsymbol{Z}}_{tikh} = \boldsymbol{U}\,\boldsymbol{\Phi}_{tikh}(\alpha)\,\boldsymbol{F}$.

Let us see how $\phi_i^{tikh}(\alpha)$ changes as α and s_i vary. First, as α increases, $\phi_i^{tikh}(\alpha)$ decreases, as illustrated in Figure 3.1. This is reasonable since α balances the error term and the penalty term. Later in Section 3.4, we will propose how we can determine an appropriate parameter α. Second, $\phi_i^{tikh}(\alpha)$ monotonically increases as s_i increases, so the Tikhonov filter matrix reduces the less important principal components more intensely. The main difference between the Tikhonov method and TSVD is that Tikhonov preserves some information from the least important principal components while TSVD discards all of it.

3.2.3 The Relation Between Filtered PCA and a Factor Model

Some asset pricing models (e.g., [74, 80]) model asset returns with a factor model:

$$\boldsymbol{r}(t) = \mathbb{E}[\boldsymbol{r}(t)] + \mathcal{B}\boldsymbol{\varphi}(t) + \boldsymbol{\epsilon}(t). \tag{3.2.17}$$

The assumptions are that

$$\mathbb{E}[\boldsymbol{\varphi}(t)] = \mathbb{E}[\boldsymbol{\epsilon}(t)] = \mathbf{0}, \tag{3.2.18}$$

$$\mathbb{E}(\epsilon_i(t)\epsilon_j(t)) = \mathbb{E}(\epsilon_i(t)\varphi_\ell(t)) = \mathbb{E}(\varphi_i(t)\varphi_j(t)) = 0 \text{ for all } i \neq j, \tag{3.2.19}$$

where $\boldsymbol{\varphi}(t) = [\varphi_1(t), ..., \varphi_\ell(t)]^T$ and $\boldsymbol{\epsilon}(t) = [\epsilon_1(t), ..., \epsilon_N(t)]^T$. The common factors $\varphi_i(t)$ are referred to as systematic factors, and $\epsilon_i(t)$ is called an unsystematic (idiosyncratic) factor. The matrix $\mathcal{B} = (\beta_{ik})$ is called a factor-loading matrix, and β_{ik} represents the sensitivity of the i-th asset to the k-th factor.

We can interpret our linear model (3.2.12) as a factor model. By (3.2.2) and (3.2.12), we have a linear equation for $\boldsymbol{r}(t)$ as

$$\begin{aligned} \boldsymbol{r}(t) &= \mathbb{E}_s[\boldsymbol{r}(t)] + \boldsymbol{D}_V^{\frac{1}{2}} \left(\boldsymbol{U}\widetilde{\boldsymbol{f}}(t) + \boldsymbol{\epsilon}_z(t)\right) & (3.2.20) \\ &= \mathbb{E}_s[\boldsymbol{r}(t)] + \boldsymbol{B}\widetilde{\boldsymbol{f}}(t) + \boldsymbol{\epsilon}_r(t), & (3.2.21) \end{aligned}$$

where

$$\boldsymbol{B} = \boldsymbol{D}_V^{\frac{1}{2}}\boldsymbol{U} \quad \text{and} \quad \boldsymbol{\epsilon}_r(t) = \boldsymbol{D}_V^{\frac{1}{2}}\boldsymbol{\epsilon}_z(t). \tag{3.2.22}$$

Comparing (3.2.17) and (3.2.21), if we assume that $\widetilde{\boldsymbol{f}}(t)$ represents the systematic factors $\boldsymbol{\varphi}(t)$ well, we can interpret $\boldsymbol{B}$ and $\boldsymbol{\epsilon}_r(t)$ as estimates of the loading matrix $\mathcal{B}$ and the unsystematic factor $\boldsymbol{\epsilon}(t)$ in (3.2.17). Since $\boldsymbol{\epsilon}_z(t) = \boldsymbol{z}(t) - \boldsymbol{U}\widetilde{\boldsymbol{f}}(t)$, $\boldsymbol{\epsilon}_r(t)$ becomes

$$\boldsymbol{\epsilon}_r(t) = \boldsymbol{D}_V^{\frac{1}{2}}\boldsymbol{\epsilon}_z(t) = \boldsymbol{D}_V^{\frac{1}{2}}(\boldsymbol{z}(t) - \boldsymbol{U}\widetilde{\boldsymbol{f}}(t)). \tag{3.2.23}$$

Because $\boldsymbol{z}(t) = \boldsymbol{U}\boldsymbol{f}(t)$ and $\widetilde{\boldsymbol{f}}(t) = \boldsymbol{\Phi}\boldsymbol{f}(t)$, the factor models result in the estimate

$$\boldsymbol{\epsilon}_r(t) = \boldsymbol{D}_V^{\frac{1}{2}}(\boldsymbol{U}\boldsymbol{f}(t) - \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{f}(t)) = (\boldsymbol{D}_V^{\frac{1}{2}}\boldsymbol{U})(\boldsymbol{I}_k - \boldsymbol{\Phi})\boldsymbol{f}(t) = \boldsymbol{B}(\boldsymbol{I}_k - \boldsymbol{\Phi})\boldsymbol{f}(t). \tag{3.2.24}$$

3.3 Estimate of the Covariance Matrix Σ

In this section we study how filtering changes the covariance and correlation estimates and the estimate of risk exposure, and how to ensure that the estimated covariance matrix has full rank.

3.3.1 A Covariance Estimate

Now we derive a covariance matrix estimate $\widetilde{\boldsymbol{\Sigma}}$ from (3.2.21), respecting the structure of the factor model (3.2.17). By (3.2.19), the covariance matrix $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \mathcal{B}\text{cov}[\boldsymbol{\varphi}(t)]\mathcal{B}^T + \text{cov}[\boldsymbol{\epsilon}(t)] = \boldsymbol{\Sigma}_s + \boldsymbol{D}_\epsilon, \tag{3.3.1}$$

where $\boldsymbol{\Sigma}_s$ denotes the systematic component $\mathcal{B}\text{cov}[\boldsymbol{\varphi}(t)]\mathcal{B}^T$ and $\boldsymbol{D}_\epsilon$ denotes the unsystematic component $\text{cov}[\boldsymbol{\epsilon}(t)]$. We estimate the systematic part $\boldsymbol{\Sigma}_s$ by $\widetilde{\boldsymbol{\Sigma}}_s = \boldsymbol{B}\text{cov}_s[\widetilde{\boldsymbol{f}}(t)]\boldsymbol{B}^T$. Because $\boldsymbol{f}(t)$ has zero-mean, $\widetilde{\boldsymbol{f}}(t) = \boldsymbol{\Phi}\boldsymbol{f}(t)$ also has zero-mean, so

$$\text{cov}_s[\widetilde{\boldsymbol{f}}(t)] = \frac{1}{T}(\boldsymbol{\Phi}\boldsymbol{F})(\boldsymbol{\Phi}\boldsymbol{F})^T = \frac{1}{T}(\boldsymbol{\Phi}^2\boldsymbol{S}_k^2). \tag{3.3.2}$$

Therefore, the estimate of Σ_s becomes

$$\widetilde{\Sigma}_s = \boldsymbol{B}\text{cov}_s[\widetilde{\boldsymbol{f}}(t)]\boldsymbol{B}^T = \frac{1}{T}\boldsymbol{B}(\Phi^2\boldsymbol{S}_k^2)\boldsymbol{B}^T. \tag{3.3.3}$$

The unsystematic part $\boldsymbol{D}_\epsilon$ in (3.3.1) is diagonal since the unsystematic factors $\epsilon_i(t)$ are mutually uncorrelated. Thus, we estimate cov$[\boldsymbol{\epsilon}(t)]$ by the diagonal part of the difference $\widetilde{\boldsymbol{D}}_\epsilon$ between

$$\widetilde{\Sigma}_{sample} = \text{cov}_s[\boldsymbol{r}(t)] = \frac{1}{T}\boldsymbol{B}\boldsymbol{S}_k^2\boldsymbol{B}^T, \tag{3.3.4}$$

and $\widetilde{\Sigma}_s$. Hence,

$$\widetilde{\boldsymbol{D}}_\epsilon = \text{diag}\left(\widetilde{\Sigma}_{sample} - \widetilde{\Sigma}_s\right) = \text{diag}\left(\frac{1}{T}(\boldsymbol{B}(\boldsymbol{I}_k - \Phi^2)\boldsymbol{S}_k^2\boldsymbol{B}^T)\right). \tag{3.3.5}$$

Finally, the filtered covariance matrix $\widetilde{\Sigma}$ will be

$$\widetilde{\Sigma} = \widetilde{\Sigma}_s + \widetilde{\boldsymbol{D}}_\epsilon, \tag{3.3.6}$$

where $\widetilde{\Sigma}_s$ and $\widetilde{\boldsymbol{D}}_\epsilon$ are defined by (3.3.3) and (3.3.5). By the definition of $\widetilde{\boldsymbol{D}}_\epsilon$, the diagonal of $\widetilde{\Sigma}$ equals var$_s[\boldsymbol{r}(t)]$.

Now we analyze how the filtering function Φ affects the sample correlation matrix corr$_s[\boldsymbol{r}(t)]$. By (3.3.6), the filtered correlation matrix $\widetilde{\Omega}$ can be calculated as

$$\widetilde{\Omega} = \boldsymbol{D}_V^{-\frac{1}{2}}\widetilde{\Sigma}\boldsymbol{D}_V^{-\frac{1}{2}} = \frac{1}{T}\boldsymbol{U}\Phi^2\boldsymbol{S}_k^2\boldsymbol{U}^T + \boldsymbol{D}_V^{-\frac{1}{2}}\tilde{\boldsymbol{D}}_\epsilon\boldsymbol{D}_V^{-\frac{1}{2}}, \tag{3.3.7}$$

where the second term makes the diagonal elements of $\widetilde{\Omega}$ equal one. On the other hand, the sample correlation matrix corr$_s[\boldsymbol{r}(t)]$ can be calculated as

$$\text{corr}_s[\boldsymbol{r}(t)] = \boldsymbol{D}_V^{-\frac{1}{2}}\widetilde{\Sigma}_{sample}\boldsymbol{D}_V^{-\frac{1}{2}}.$$

```
Step 1. Estimate the systematic component of the covariance $\frac{1}{T}\boldsymbol{B}(\boldsymbol{\Phi}^2\boldsymbol{S}_k^2)\boldsymbol{B}^T$
        where $\boldsymbol{\Phi}$ is the diagonal matrix of filter factors.
Step 2. Change the main diagonal to be the sample variances.
```

Table 3.1: *The algorithm to compute the covariance estimate* $\widetilde{\boldsymbol{\Sigma}}$*. For Tikhonov, the filter factors are* $\boldsymbol{\Phi}_{tikh} = \mathrm{diag}\left(\frac{s_1^2}{s_1^2+\alpha^2}, \ldots, \frac{s_k^2}{s_k^2+\alpha^2}\right)$.

By (3.2.22) and (3.3.4), this becomes

$$\mathrm{corr}_s[\boldsymbol{r}(t)] = \boldsymbol{D}_V^{-\frac{1}{2}} \left(\frac{1}{T}\boldsymbol{B}\boldsymbol{S}_k^2\boldsymbol{B}^T\right) \boldsymbol{D}_V^{-\frac{1}{2}} = \frac{1}{T}\boldsymbol{U}\boldsymbol{S}_k^2\boldsymbol{U}^T. \tag{3.3.8}$$

Comparing $\widetilde{\boldsymbol{\Omega}}$ in (3.3.7) and $\mathrm{corr}_s[\boldsymbol{r}(t)]$ in (3.3.8), we can see that $\widetilde{\boldsymbol{\Omega}}$ is the result of applying the filtering matrix $\boldsymbol{\Phi}^2$ to $\boldsymbol{S}_k^2$ in $\mathrm{corr}_s[\boldsymbol{r}(t)]$ and replacing the diagonal elements with one. Since each diagonal element of $\boldsymbol{S}_k^2$ corresponds to an eigenvalue of $\mathrm{corr}_s[\boldsymbol{r}(t)]$, the filtering matrix $\boldsymbol{\Phi}^2$ attenuates the eigenvalues of $\mathrm{corr}_s[\boldsymbol{r}(t)]$. In the previous section, we introduced two filtering matrices :

$$\boldsymbol{\Phi}_{trun}(\widehat{k}) = \mathrm{diag}\left(\underbrace{1,\ldots,1}_{\widehat{k}}, \underbrace{0,\ldots,0}_{k-\widehat{k}}\right), \tag{3.3.9}$$

$$\text{and} \quad \boldsymbol{\Phi}_{tikh}(\alpha) = \mathrm{diag}\left(\frac{s_1^2}{s_1^2+\alpha^2}, \ldots, \frac{s_k^2}{s_k^2+\alpha^2}\right). \tag{3.3.10}$$

Therefore, $\boldsymbol{\Phi}_{trun}^2(\widehat{k})$ truncates the eigencomponents corresponding to the $(k-\widehat{k})$ smallest eigenvalues, and $\boldsymbol{\Phi}_{tikh}^2(\alpha)$ down-weights all the eigenvalues at a rate $\left(\frac{s_i^2}{s_i^2+\alpha^2}\right)^2 = \left(\frac{\lambda_i}{\lambda_i+\alpha^2}\right)^2$ where λ_i is the i-th largest eigenvalue of $\mathrm{cov}_s[\boldsymbol{z}(t)]$. Hence, the truncated

SVD filtering functions $\phi^2_{trun}(\lambda_i)$ for eigenvalues λ_i become

$$\phi^2_{trun}(\lambda_i) = \begin{cases} 1, & \text{if } i \leq \widehat{k}, \\ 0, & \text{otherwise}, \end{cases}$$

and the Tikhonov filtering functions $\phi^2_{tikh}(\lambda_i)$ are

$$\phi^2_{tikh}(\lambda_i) = \left(\frac{\lambda_i}{\lambda_i + \alpha^2}\right)^2.$$

We let $\widetilde{\boldsymbol{\Sigma}}_{trun}$ and $\widetilde{\boldsymbol{\Sigma}}_{tikh}$ denote the estimates resulting from applying $\boldsymbol{\Phi}^2_{trun}(\widehat{k})$ and $\boldsymbol{\Phi}^2_{tikh}(\alpha)$ to (3.3.6). Finally, we can summarize the process of estimating the covariance matrix as Table 3.1.

3.3.2 Risk Exposure to Factors

By (3.3.1), the variance of a portfolio return can be expressed as

$$\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} = \boldsymbol{w}^T\left(\boldsymbol{\Sigma}_s + \boldsymbol{D}_\epsilon\right)\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{\Sigma}_s\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{D}_\epsilon\boldsymbol{w}. \tag{3.3.11}$$

The systematic risk is

$$\boldsymbol{w}^T\boldsymbol{\Sigma}_s\boldsymbol{w} = \boldsymbol{w}^T\left(\mathcal{B}\text{cov}[\boldsymbol{\varphi}(t)]\mathcal{B}^T\right)\boldsymbol{w} = \boldsymbol{w}^T\left(\mathcal{B}\text{diag}\left(\text{var}[\boldsymbol{\varphi}(t)]\right)\mathcal{B}^T\right)\boldsymbol{w}, \tag{3.3.12}$$

because $\varphi_i(t)$ are mutually uncorrelated by (3.2.19). This can be expanded as

$$\boldsymbol{w}^T\boldsymbol{\Sigma}_s\boldsymbol{w} = \sum_{i=1}^{k}\text{var}[\varphi_i(t)](\boldsymbol{w}^T\boldsymbol{\beta}_i)^2, \tag{3.3.13}$$

where $\boldsymbol{\beta}_i$ is the i-th column of $\mathcal{B}$. The i-th term in (3.3.13) represents the risk exposure of the portfolio to the i-th factor.

On the other hand, the estimated matrix $\widetilde{\boldsymbol{\Sigma}}_s$ in (3.3.3) can be rewritten as

$$\widetilde{\boldsymbol{\Sigma}}_s = \frac{1}{T}\boldsymbol{B}\boldsymbol{\Phi}^2\boldsymbol{S}_k^2\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{\Phi}^2\left(\frac{\boldsymbol{S}_k^2}{T}\right)\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{\Phi}^2\mathrm{diag}\left(\mathrm{var}_s[\boldsymbol{f}(t)]\right)\,\boldsymbol{B}^T, \tag{3.3.14}$$

because $\mathrm{var}_s[\boldsymbol{f}(t)] = \mathrm{diag}\left(\boldsymbol{S}_k^2/T\right)$ by (3.2.8). Hence, we can calculate the estimated systematic risk as

$$\boldsymbol{w}^T\widetilde{\boldsymbol{\Sigma}}_s\boldsymbol{w} = \sum_{i=1}^{k}\phi_i^2\left(\mathrm{var}_s[\boldsymbol{f}_i(t)](\boldsymbol{w}^T\boldsymbol{b}_i)^2\right), \tag{3.3.15}$$

where $\boldsymbol{b}_i$ is the i-th column of $\boldsymbol{B}$. Therefore, we can see that our estimate of the risk exposure to the i-th factor is reduced by ϕ_i^2. This equation explains how the estimated covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ affects the estimated risk measure of a portfolio, downweighting risk factors corresponding to small values of $\phi_i(\alpha)$.

3.3.3 Rank Deficiency of the Covariance Matrix

Since the covariance matrix is positive semidefinite, the MV problem (3.1.1) and the GMV problem (3.1.2) always have a minimizer $\boldsymbol{w}$. However, when the covariance matrix is rank deficient, the minimizer $\boldsymbol{w}$ is not unique, which might not be desirable for investors who want to choose one portfolio. The sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{sample}$ from (3.1.3) has rank $(T-1)$ at most. Therefore, whenever the number of observations T is less than or equal to the number of stocks N, $\widetilde{\boldsymbol{\Sigma}}_{sample}$ is rank deficient. To insure a full rank and high quality estimate, we must have at least $(N+1)$ recent observations of returns, derived from at least $(N+1)$ recent trades, and this is not always possible.

Recall that the covariance matrix estimate $\widetilde{\boldsymbol{\Sigma}}$ is the sum of the systematic part $\widetilde{\boldsymbol{\Sigma}}_s$ and the unsystematic part $\widetilde{\boldsymbol{D}}_\epsilon$. By (3.3.3), we can see that $\widetilde{\boldsymbol{\Sigma}}_s$ has non-negative eigenvalues.

On the other hand, by (3.3.5),

$$\text{(The } i\text{-th diagonal element of } \widetilde{\boldsymbol{D}}_\epsilon) = \boldsymbol{e}_i^T \left(\frac{1}{T} \boldsymbol{B}(\boldsymbol{I}_k - \Phi^2) \boldsymbol{S}_k^2 \boldsymbol{B}^T \right) \boldsymbol{e}_i. \tag{3.3.16}$$

It is reasonable to assume that $\boldsymbol{e}_i^T \boldsymbol{B}$ is not zero for any i since it becomes zero only when the i-th stock has zero variance of returns by (3.2.22). Thus, the diagonal matrix $\widetilde{\boldsymbol{D}}_\epsilon$ is positive definite whenever all $\phi_i < 1$. In the case of Tikhonov filtering, whenever $\alpha > 0$,

$$\phi_i^{tikh}(\alpha) = \frac{s_i^2}{s_i^2 + \alpha^2} < 1,$$

so $\widetilde{\boldsymbol{D}}_\epsilon$ is positive definite. Therefore, since $\widetilde{\boldsymbol{\Sigma}}_s$ is positive semidefinite, adding a positive definite matrix ensures that that Tikhonov covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{tikh}$ is positive definite and therefore full-rank.

Sharpe [79], Ledoit et al. [55], Bengtsson and Holst [5], and Plerou et al. [69] also overcome the rank-deficiency problem by replacing the diagonals of their estimate with the sample variances like Step 2 in Table 3.1. However, some of their filtering values ϕ_i could have a value of 1 as we will see in Section 3.5. This implies that the resulting estimate $\widetilde{\boldsymbol{\Sigma}}$ could be rank-deficient or very ill-conditioned even after adding $\widetilde{\boldsymbol{D}}_\epsilon$, because $\widetilde{\boldsymbol{D}}_\epsilon$ is positive semidefinite. In the case that the estimate still has a large condition number even after the Step 2, we can fix the problem by a small modification as follows:

$$\widetilde{\boldsymbol{\Sigma}}_{ii} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{ii} + \delta_i \quad \text{for } i = 1, \ldots, N\,, \tag{3.3.17}$$

where δ_i is a small positive number.

Theorem 3.3.1 (Condition number modification)**.** *Replacing the main diagonal of the covariance estimate $\widetilde{\mathbf{\Sigma}}$ as specified in (3.3.17) guarantees that*

$$\textit{cond}(\widetilde{\mathbf{\Sigma}}) \leq \frac{\lambda_{\max}(\widetilde{\mathbf{\Sigma}}) + \max(\delta_i)}{\min(\delta_i)},$$

where $\lambda_{\max}(\cdot)$ is the maximum eigenvalue of the matrix.

Proof. This is a direct consequence of the eigenvalue interlacing theorem [82, p.203] and the positive semidefiniteness of $\widetilde{\mathbf{\Sigma}}$. □

This modification is useful especially for the sample covariance matrix $\widetilde{\mathbf{\Sigma}}_{sample}$ when $T \leq N$, and for the truncation-based estimators whose filtering factors ϕ_i equal 1 for some i.

3.4 Choice of Tikhonov Parameter α

So far, we have seen how to filter noise from the covariance matrix using regularization and how to fix the rank deficiency of the resulting covariance matrix. In order to use Tikhonov regularization, we need to determine the Tikhonov parameter α. In regularization methods for discrete ill-posed problems, there are intensive studies about choosing α using methods such as Generalized Cross Validation [29], L-curves [33, 35], and residual periodograms [75, 76].

In factor analysis and principal component analysis, there are analogous studies to determine the number of factors such as Bartlett's test [3], SCREE test [9], average root [32], partial correlation procedure [91], and cross-validation [94]. More recently, Plerou

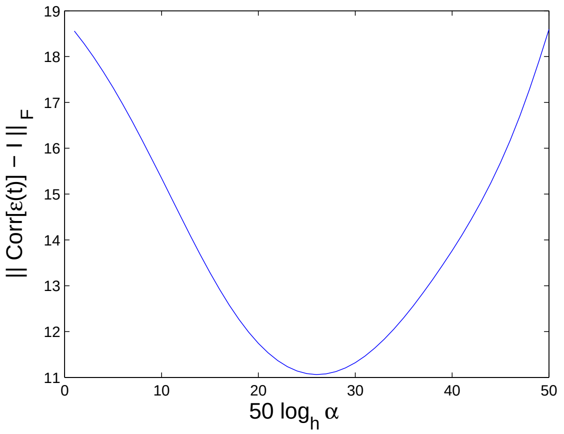


Figure 3.2: *The difference* $||\mathrm{corr}_s[\boldsymbol{\epsilon}_r(t)] - \mathbf{I}_N||_F$ *as a function of log-scaled* α *where* $h = \max{(s_i)}$.

et al. [68, 69] applied random matrix theory, which will be described in Section 3.5.6. In the context of arbitrage pricing theory, some different approaches were proposed to determine the number of factors: Trzcinka [88] studied the behavior of eigenvalues as the number of assets increases, and Connor and Korajczyk [15] studied the probabilistic behavior of noise factors.

The use of these methods requires various statistical properties for $\boldsymbol{\epsilon}_r(t)$ in the linear model (3.2.21). We note that since $\mathbb{E}_s[\boldsymbol{f}(t)] = \mathbf{0}$ by (3.2.6), the noise $\boldsymbol{\epsilon}_r(t)$ in (3.2.21) has zero-mean: By (3.2.24),

$$\mathbb{E}_s[\boldsymbol{\epsilon}_r(t)] = \boldsymbol{B}(\boldsymbol{I}_k - \boldsymbol{\Phi})\, \mathbb{E}_s[\boldsymbol{f}(t)] = \mathbf{0}. \tag{3.4.1}$$

For our Tikhonov estimation, we propose a new method adopting a mutually uncorrelated noise assumption in a factor model (3.2.19), so $\mathrm{corr}_s[\boldsymbol{\epsilon}_r(t)] \simeq \boldsymbol{I}_N$. Hence, as a criterion to determine an appropriate parameter α, we formulate an optimization problem minimizing the correlations among the noise,

$$\min_{\alpha \in [s_k, s_1]} || \operatorname{corr}_s[\boldsymbol{\epsilon}_r(t)] - \boldsymbol{I}_N \, ||_F, \tag{3.4.2}$$

where s_1 and s_k are the largest and the smallest singular values of $\boldsymbol{Z}$ as defined in (3.2.5). This is similar to Velicer's partial correlation procedure [91] to determine the number of principal components. Figure 3.2 illustrates an example of $||\operatorname{corr}_s[\boldsymbol{\epsilon}_r(t)] - \boldsymbol{I}_N||_F$ as a function of α in the range $[s_k, s_1]$. The parameter might alternatively be determined by an asymptotic analysis proposed by Ledoit and Wolf [55, 56] or a cross validation used by DeMiguel et al. [19].

3.5 Comparison to Other Estimators

In this section, we compare other covariance estimators to our Tikhonov estimator and put them all in a common framework. We summarize how they filter the eigenvalues of the sample correlation matrix with filtering functions $\phi^2(\lambda_i)$. Most of these methods use a two step procedure as shown in Table 3.1: filter the eigenvalues, and then adjust the main diagonal. We note any exceptions in our descriptions.

3.5.1 $\widetilde{\boldsymbol{\Sigma}}_{sample}$: Sample Covariance Matrix

A sample covariance matrix is the filtering target of most covariance estimators including our Tikhonov estimator. Thus, the sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{sample}$ can be thought of as an unfiltered covariance matrix, so the filtering function $\phi_s^2(\lambda_i)$ for eigenvalues of

$\mathrm{cov}_s[\boldsymbol{z}(t)]$ is

$$\phi_s^2(\lambda_i) = 1 \quad \text{for } i = 1, \ldots, \text{rank}\left(\widetilde{\Sigma}_{sample}\right).$$

3.5.2 $\widetilde{\Sigma}_{market}$ from the Single Market Index Model [79]

Sharpe [79] proposed a single index market model

$$\boldsymbol{r}\,(t) = \mathbb{E}[\boldsymbol{r}(t)] + \boldsymbol{b}\, r_m(t) + \boldsymbol{\epsilon}(t), \tag{3.5.1}$$

where $\boldsymbol{r}\,(t) \in \mathbb{R}^{N\times 1}$ is stock return at time t,

$r_m(t)$ is market return at time t,

$\boldsymbol{\epsilon}(t)$ is zero-mean uncorrelated error at time t,

and $\boldsymbol{b} \in \mathbb{R}^{N\times 1}$.

Unlike the factor model (3.2.17), this model assumes that the stock returns $\boldsymbol{r}(t)$ have only one common factor, the market return $r_m(t)$. Interestingly, Plerou et al. [69, p.8] observed that the principal component corresponding to the largest eigenvalue of the correlation matrix $\mathrm{corr}_s[\boldsymbol{r}(t)](= \mathrm{cov}_s[\boldsymbol{z}(t)])$ is proportional to the entire market returns. This observation is natural in that most stocks are highly affected by the market situation. Based on their observation, we expect that the most important principal component $f_1(t)$ in (3.2.4) represents the market return $r_m(t)$. Thus, we can represent the relation between $\widetilde{\boldsymbol{f}}(t) = [\tilde{f}_1(t), \ldots, \tilde{f}_k(t)]$ in (3.2.21) and $r_m(t)$ as

$$\tilde{f}_i(t) \simeq \begin{cases} C\, r_m(t), & \text{when } i = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.5.2}$$

for some constant C. Hence, the corresponding filtering function $\phi_m^2(\lambda_i)$ for $\widetilde{\boldsymbol{\Sigma}}_{market}$ becomes

$$\phi_m^2(\lambda_i) \simeq \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.5.3}$$

Therefore, the filter function implicitly truncates all but the largest eigencomponent of $\text{corr}_s[\boldsymbol{r}(t)]$.

3.5.3 $\widetilde{\boldsymbol{\Sigma}}_{s\to m}$: Shrinkage toward $\widetilde{\boldsymbol{\Sigma}}_{market}$ [55]

Ledoit et al. propose a shrinkage method from $\widetilde{\boldsymbol{\Sigma}}_{sample}$ to $\widetilde{\boldsymbol{\Sigma}}_{market}$ as

$$\widetilde{\boldsymbol{\Sigma}}_{s\to m} = \gamma\, \widetilde{\boldsymbol{\Sigma}}_{market} + (1-\gamma)\widetilde{\boldsymbol{\Sigma}}_{sample}, \tag{3.5.4}$$

where $0 \le \gamma \le 1$. Thus, the shrinkage estimator is the weighed average of $\widetilde{\boldsymbol{\Sigma}}_{sample}$ and $\widetilde{\boldsymbol{\Sigma}}_{market}$. In order to find an optimal weight γ, they minimize the distance between $\widetilde{\boldsymbol{\Sigma}}_{s\to m}$ and the true covariance matrix $\boldsymbol{\Sigma}$:

$$\min_{\gamma} ||\widetilde{\boldsymbol{\Sigma}}_{s\to m} - \boldsymbol{\Sigma}||_F^2.$$

Since the true covariance matrix $\boldsymbol{\Sigma}$ is unknown, they use an asymptotic variance to determine an optimal γ. (Refer to [55, Section 2.5-6] for a detailed description.) Considering that $\widetilde{\boldsymbol{\Sigma}}_{market}$ is the result of the implicit truncation method, we can think of this shrinkage method as implicitly down-weighting all eigenvalues but the largest at a rate $(1-\gamma)$. Therefore, we can represent the filtering function $\phi_{s\to m}^2(\lambda_i)$ as

$$\phi_{s\to m}^2(\lambda_i) \simeq \begin{cases} 1, & \text{if } i = 1, \\ 1-\gamma, \text{ where } 0 \le \gamma \le 1 & \text{otherwise.} \end{cases} \tag{3.5.5}$$

3.5.4 Truncated Covariance Matrix $\widetilde{\mathbf{\Sigma}}_{trun}$ [24]

As mentioned in Section 3.3.1, the truncated covariance matrix $\widetilde{\mathbf{\Sigma}}_{trun}$ has the filtering function $\phi^2_{trun}(\lambda_i)$ for the eigenvalues λ_i of cov$_s[\boldsymbol{z}(t)]$, where

$$\phi^2_{trun}(\lambda_i) = \begin{cases} 1, & \text{if } i = 1, \ldots, \widehat{k}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.5.6}$$

Thus, the model of Elton and Gruber [24] truncates all but the $\widehat{k}$ largest eigencomponents of cov$_s[\boldsymbol{z}(t)]$.

3.5.5 $\widetilde{\mathbf{\Sigma}}_{s\to trun}$: Shrinkage toward $\widetilde{\mathbf{\Sigma}}_{trun}$ [5]

Bengtsson and Holst propose a shrinkage estimator from $\widetilde{\mathbf{\Sigma}}_{sample}$ to $\widetilde{\mathbf{\Sigma}}_{trun}$ as

$$\widetilde{\mathbf{\Sigma}}_{s\to trun} = \gamma\, \widetilde{\mathbf{\Sigma}}_{trun} + (1-\gamma)\widetilde{\mathbf{\Sigma}}_{sample}, \tag{3.5.7}$$

where $0 \leq \gamma \leq 1$. They determine the parameter γ in a way similar to [55]. (Refer to [5, Section 4.1-4.2] for detailed description.) Therefore, $\widetilde{\mathbf{\Sigma}}_{s\to trun}$ is a variant of the shrinkage method toward $\widetilde{\mathbf{\Sigma}}_{trun}$. Because $\widetilde{\mathbf{\Sigma}}_{trun}$ is the truncated covariance matrix containing the $\widehat{k}$ most significant eigencomponents of cov$_s[\boldsymbol{z}(t)]$, we can regard $\widetilde{\mathbf{\Sigma}}_{s\to trun}$ as damping the smallest eigenvalues by $(1-\gamma)$. Thus, the filtering function corresponding to this approach is

$$\phi^2_{s\to trun}(\lambda_i) = \begin{cases} 1, & \text{if } i = 1, \ldots, \widehat{k}\ , \\ 1-\gamma, \text{ where } 0 \leq \gamma \leq 1, & \text{otherwise.} \end{cases} \tag{3.5.8}$$

Rather than removing all the least important principal components as Elton and Gruber did, Bengtsson and Holst try to preserve the potential information of unimportant principal components by this single-rate attenuation. Bengtsson and Holst conclude that their shrinkage matrix $\widetilde{\boldsymbol{\Sigma}}_{s\rightarrow trun}$ performed best in the Swedish stock market when the shrinkage target $\widetilde{\boldsymbol{\Sigma}}_{trun}$ takes only the most significant principal component ($\widehat{k} = 1$). They also mention that the result is consistent with RMT because only the largest eigenvalue deviates far from the range of $[\lambda_{\min}, \lambda_{\max}]$.

3.5.6 $\widetilde{\boldsymbol{\Sigma}}_{RMT:trun}$ Truncation by Random Matrix Theory [69]

Plerou et al. [69] apply random matrix theory (RMT) [60] which shows that the eigenvalues of a random correlation matrix have a distribution within an interval determined by the ratio of N and T. Let corr$_{random}$ be a random correlation matrix

$$\text{corr}_{random} = \frac{1}{T}\boldsymbol{A}\boldsymbol{A}^T, \tag{3.5.9}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times T}$ contains mutually independent random elements $a_{i,t}$ with zero-mean and unit variance. When $Q = T/N \geq 1$ is fixed, the eigenvalues λ of corr$_{random}$ have a limiting distribution (as $N \rightarrow \infty$)

$$f(\lambda) = \begin{cases} \dfrac{Q}{2\pi\sigma^2}\dfrac{\sqrt{(\lambda_{\max}-\lambda)(\lambda_{\min}-\lambda)}}{\lambda}, & \lambda_{\min} \leq \lambda \leq \lambda_{\max}, \\ \\ 0, & \text{otherwise,} \end{cases} \tag{3.5.10}$$

where σ^2 is the variance of the elements of $\boldsymbol{A}$, $\lambda_{\min} \leq \lambda \leq \lambda_{\max}$, and

$$\lambda_{\min}^{\max} = \sigma^2\left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right).$$

By comparing the eigenvalue distribution of corr$_s[\boldsymbol{r}(t)]$ with $f(\lambda)$, Plerou et al. show that most eigenvalues are within $[\lambda_{\min}, \lambda_{\max}]$. They conclude that only a few large eigenvalues deviating from $[\lambda_{\min}, \lambda_{\max}]$ correspond to eigenvalues of the real correlation matrix, so the other eigencomponents should be removed from $\mathrm{corr}_s[\boldsymbol{r}(t)]$. Thus, the filtering function $\phi^2_{RMT:trun}(\lambda_i)$ for the eigenvalue λ_i of $\mathrm{corr}_s[\boldsymbol{r}(t)]$ is

$$\phi^2_{RMT:trun}(\lambda_i) = \begin{cases} 1, & \text{if } \lambda_i \geq \lambda_{\max} \text{ ,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.5.11}$$

3.5.7 $\widetilde{\Sigma}_{RMT:repl}$ Replacing the RMT Eigenvalues [53]

Laloux et al. apply RMT to this problem in a way somewhat different from Plerou et al. First, they find the best fitting σ^2 in (3.5.10) to the eigenvalue distribution of the observed correlation matrix rather than assuming that $\sigma^2 = 1$. Second, they replace each eigenvalue in the RMT interval with a constant value C, chosen so that the trace of the matrix is unchanged. Thus, the filtering function $\phi^2_{RMT:repl}(\lambda_i)$ for eigenvalues is

$$\phi^2_{RMT:repl}(\lambda_i) = \begin{cases} 1, & \text{if } \lambda_i \geq \lambda_{\max} \text{ ,} \\ \frac{C}{\lambda_i}, & \text{otherwise.} \end{cases} \tag{3.5.12}$$

This approach does not require the application of Step 2 in Table 3.1 , since it replaces the smallest eigenvalues with a positive constant. The resulting covariance matrix does not preserve the original variances.

Estimator	Filtering function $\phi^2(\lambda_i)$
$\widetilde{\boldsymbol{\Sigma}}_{sample}$	$\phi_s^2(\lambda_i) = 1$
$\widetilde{\boldsymbol{\Sigma}}_{market}$[79]	$\phi_m^2(\lambda_i) \simeq \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{s\to m}$[55]	$\phi_{s\to m}^2(\lambda_i) \simeq \begin{cases} 1, & \text{if } i = 1, \\ 1-\gamma, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{trun}$[24]	$\phi_{trun}^2(\lambda_i) = \begin{cases} 1, & \text{if } i = 1, \ldots, \widehat{k}, \\ 0, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{s\to trun}$[5]	$\phi_{s\to trun}^2(\lambda_i) = \begin{cases} 1, & \text{if } i = 1, \ldots, \widehat{k}, \\ 1-\gamma, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{RMT:trun}$[69]	$\phi_{RMT:trun}^2(\lambda_i) = \begin{cases} 1, & \text{if } \lambda_i \geq \lambda_{\max}, \\ 0, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{RMT:repl}$[53]	$\phi_{RMT:repl}^2(\lambda_i) = \begin{cases} 1, & \text{if } \lambda_i \geq \lambda_{\max}, \\ \frac{C}{\lambda_i}, & \text{otherwise.} \end{cases}$
$\widetilde{\boldsymbol{\Sigma}}_{tikh}$	$\phi_{tikh}^2(\lambda_i) = \left(\dfrac{\lambda_i}{\lambda_i + \alpha^2}\right)^2$

Table 3.2: *Definition of the filter function $\phi^2(\lambda_i)$ for each covariance estimator where $i = 1, \ldots, \mathrm{rank}\left(\widetilde{\boldsymbol{\Sigma}}_{sample}\right)$.*

3.5.8 $\widetilde{\boldsymbol{\Sigma}}_{s\to I}$: Shrinkage toward *I* [56]

Ledoit et al. also introduced a shrinkage method from $\widetilde{\boldsymbol{\Sigma}}_{sample}$ to the identity matrix $\boldsymbol{I}_N$ as

$$\widetilde{\boldsymbol{\Sigma}}_{s\to I} = \gamma\left(m\boldsymbol{I}_N\right) + (1-\gamma)\widetilde{\boldsymbol{\Sigma}}_{sample}, \tag{3.5.13}$$

where $m = \dfrac{\mathrm{tr}\left(\widetilde{\boldsymbol{\Sigma}}_{sample}\right)}{N}$ and $0 \leq \gamma \leq 1$. They provide a method to estimate an optimal γ. (Refer to [56, Section 3] for a detailed description.) There is no simple expression for the filter factors. In addition, this method does not use Step 2 in Table 3.1 since its shrinkage target $\boldsymbol{I}_N$ has full rank.

3.5.9 Tikhonov Covariance Matrix $\widetilde{\mathbf{\Sigma}}_{tikh}$

As mentioned at Section 3.3.1, the Tikhonov covariance matrix $\widetilde{\mathbf{\Sigma}}_{tikh}$ has the filtering function $\phi^2_{tikh}(\lambda_i)$ for the eigenvalues λ_i of $\text{cov}_s[\mathbf{z}(t)]$, where

$$\phi^2_{tikh}(\lambda_i) = \left(\frac{\lambda_i}{\lambda_i + \alpha^2}\right)^2 , \tag{3.5.14}$$

where the parameter α is determined as described in Section 3.4.

3.5.10 Comparison

The derivations in Section 3.5 provide the proof of the following theorem.

Theorem 3.5.1 (Filtering functions)**.** *The eight covariance estimators are characterized by the choice of filtering functions specified in Table 3.2.*

Tikhonov filtering preserves potential information from less important principal components corresponding to small eigenvalues, rather than truncating them all like $\widetilde{\mathbf{\Sigma}}_{market}$, $\widetilde{\mathbf{\Sigma}}_{trun}$, and $\widetilde{\mathbf{\Sigma}}_{RMT:trun}$. In contrast to the single-rate attenuation of $\widetilde{\mathbf{\Sigma}}_{s\to m}$ and $\widetilde{\mathbf{\Sigma}}_{s\to trun}$ and the constant value replacement of $\widetilde{\mathbf{\Sigma}}_{RMT:repl}$, Tikhonov filtering reduces the effect of the smallest eigenvalues more intensely. This gradual down-weighting with respect to the magnitude of eigenvalues is the key difference between the Tikhonov method and other estimators.

In addition, all the estimators except $\widetilde{\mathbf{\Sigma}}_{s\to I}$ and $\widetilde{\mathbf{\Sigma}}_{RMT:repl}$ overcome the rank-deficiency of the covariance matrix by replacing the diagonal elements with the corresponding variances after filtering. This is what we did by preserving $\widetilde{\boldsymbol{D}}_\epsilon$ in Step 2 in

Table 3.1. However, most estimators have $\phi^2(\lambda_i) = 1$ for the largest eigenvalues as Table 3.2 shows, so the resulting covariance matrix can be still rank-deficient as we discussed in Section 3.3.3. During experiments in Section 3.6, we actually observed the rank-deficiency for some estimators even after preserving diagonal parts. This implies that an extra modification like (3.3.17) is necessary to overcome rank deficiency.

3.6 Experiments

In this section, we evaluate the covariance estimators using return data from the NYSE, AMEX, and NASDAQ. We collected the monthly data from January 1958 to December 2007 from the CRSP database (the Center for Research in Security Prices). There are 600 months over 50 years, and we randomly chose 100 stocks among those traded throughout this period.

Chopra and Ziemba [13] have noted that the MV problem is much more sensitive to errors in $\boldsymbol{\mu}$ than to errors in $\boldsymbol{\Sigma}$, and our experience confirms this observation. In fact, uncertainty in the estimates of $\boldsymbol{\mu}$ made the true return quite different from the target return. In addition, recently DeMiguel et al. [20] showed that some common portfolio strategies do not yield consistently better Sharpe ratios, certainty-equivalent returns, or turnovers, compared to a naive $1/N$ portfolio. The instability of the MV portfolio tends to increase turnover costs, so recent studies strengthen the stability by formulating new optimization problems [21]. However, since our study focuses on estimating the covariance matrix $\boldsymbol{\Sigma}$, we evaluated the estimators based on how well they minimize the risk variances in the

MV and GMV portfolios.

First, in Section 3.6.1, we evaluate the risk of GMV portfolio using the covariance estimators of Table 3.2 with various *in-sample* periods. We then compare the stability and performance of the Tikhonov estimator to that of the shrinkage estimate $\widetilde{\mathbf{\Sigma}}_{s \to m}$. Next, in Section 3.6.2, we perform similar experiments for the MV portfolio, varying the *in-sample* and *out-of-sample* periods as well as the required portfolio returns. We bypass the difficulties of estimating μ by assuming that it is known so that we can focus just on the effects of the different covariance estimators. Finally, in Section 3.6.3 we compare the GMV and MV portfolio returns, and in Section 3.6.4 we compare their predictions of risk.

3.6.1 GMV Portfolio

We simulate portfolio construction under the following scenario. We solve the GMV problem to construct a portfolio to hold for 1 month, the *out-of-sample* period T_o. We repeat this process for every month until we reach December 2007. Finally, we evaluate the variance of the *out-of-sample* returns from the GMV portfolio for each covariance estimator.

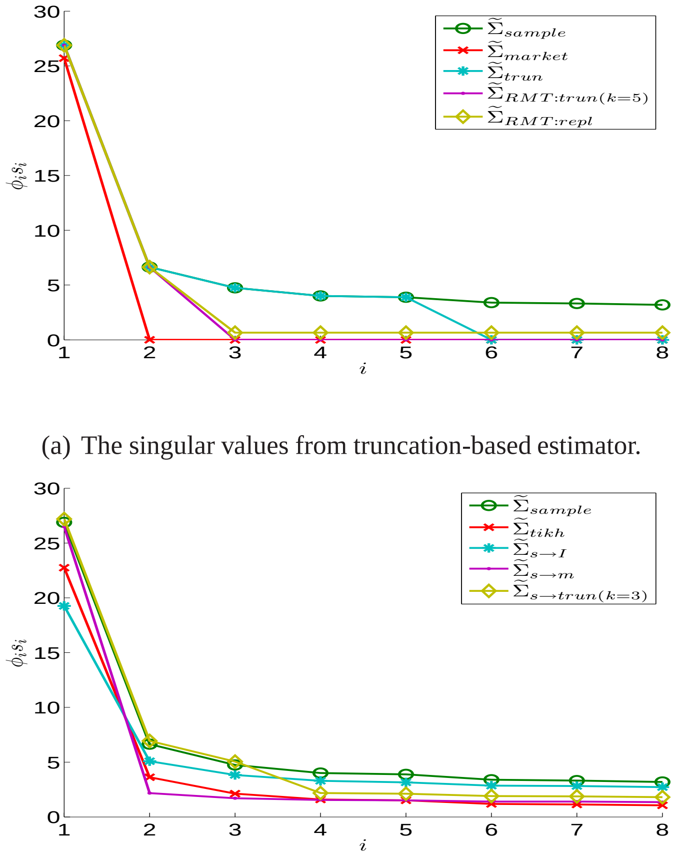
When performing this experiment, the choice of *in-sample* window size T_w is important. If T_w is too long, the data may include out-of-date information. On the other hand, if T_w is too short, the resulting covariance estimate could suffer from lack of information. We vary T_w from 1 year to 10 years. Later in Section 3.6.2, we will consider the change of the *out-of-sample* period T_o as well. We start each experiment at January

1968, giving 480 rebalancing steps for all values of T_w. For each covariance estimator, we perform the simulation for 20 different choices of 100 stocks.

Covariance Estimators in Experiments

We perform the experiment above for all the covariance estimators from Section 3.5.1 to Section 3.5.9 plus two diagonal matrices, $\widetilde{\mathbf{\Sigma}}_V$ and $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{I}}$, for a total of 11 estimators. $\widetilde{\mathbf{\Sigma}}_V$ has diagonal elements equal to $\text{var}_s[\boldsymbol{r}(t)]$, and any correlations between stocks are neglected. $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{I}}$ is an $N \times N$ identity matrix, which would yield an evenly distributed portfolio as the solution for the GMV problem (3.1.2); thus it is a good benchmark for a well-distributed portfolio. Since $\widetilde{\mathbf{\Sigma}}_{sample}$ is rank deficient, we modify it by adding small positive constants δ_i to its diagonal elements, as in (3.3.17). To compute $\widetilde{\mathbf{\Sigma}}_{market}$ and $\widetilde{\mathbf{\Sigma}}_{s\to m}$, we need the monthly market return data $r_m(t)$ in (3.5.1). In this experiment, we adopt equally-weighted market portfolio returns including distributions from CRSP database as $r_m(t)$. According to Ledoit et al. [55, p.607], an equally-weighted market portfolio is better than a value-weighted market portfolio for explaining stock market variances.

The parameter $\widehat{k}$ for $\widetilde{\mathbf{\Sigma}}_{trun}$ and $\widetilde{\mathbf{\Sigma}}_{s\to trun}$ is static, constant over all time periods. In our experiment, we perform the experiments with $\widehat{k} = 1, 5, 9$ for $\widetilde{\mathbf{\Sigma}}_{trun}$ and $\widehat{k} = 1, 2, 3$ for $\widetilde{\mathbf{\Sigma}}_{s\to trun}$. In contrast, the parameters of γ for $\widetilde{\mathbf{\Sigma}}_{s\to m}$ and $\widetilde{\mathbf{\Sigma}}_{s\to trun}$, $\widehat{k}$ for $\widetilde{\mathbf{\Sigma}}_{RMT:trun}$ and $\widetilde{\mathbf{\Sigma}}_{RMT:repl}$, and α for $\widetilde{\mathbf{\Sigma}}_{tikh}$ have their own parameter choice methods as described in Section 3.5, so we dynamically determine these parameters each time the portfolio is re-balanced.



(a) The singular values from truncation-based estimator.

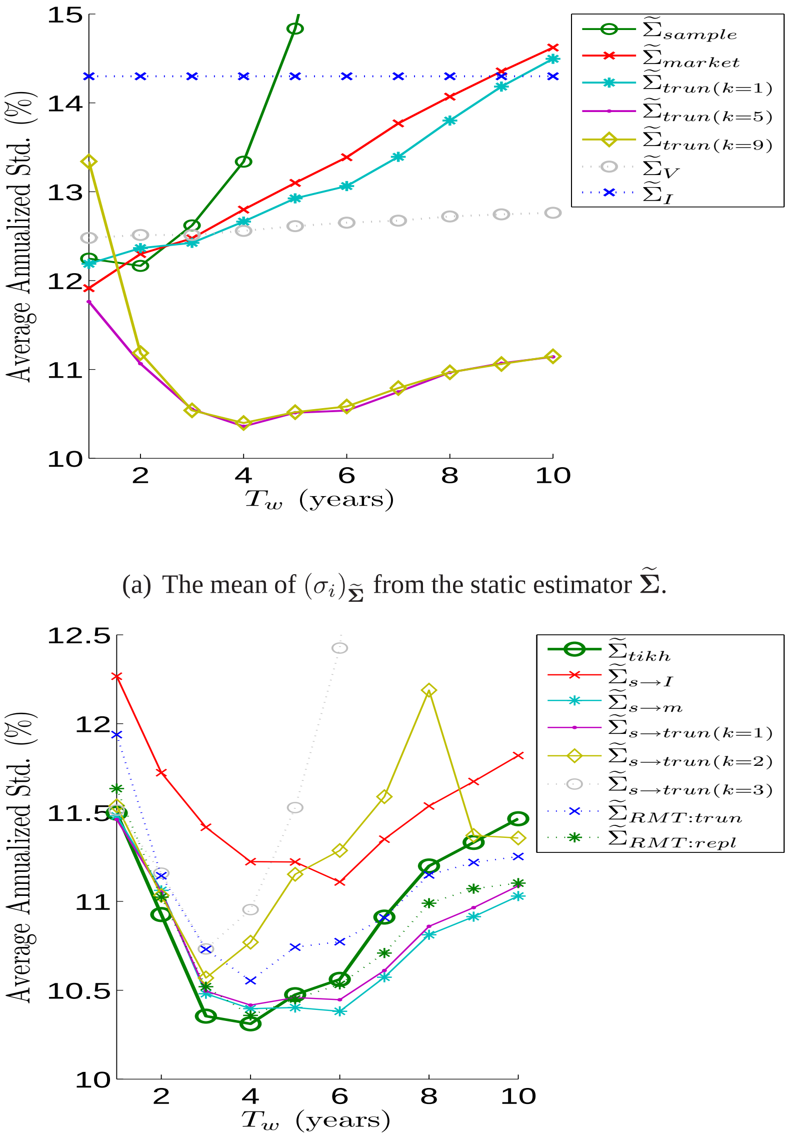
(b) The singular values from shrinkage-based estimator.

Figure 3.3: *GMV portfolios: The singular values from each estimator when* $T_w = 4$ *years.*

Figure 3.3 shows singular value plots from each estimator, which illustrates the filtering characteristics for the first *in-sample* period of $T_w = 4$ years with a particular set of 100 stocks.

Effect of *in-sample* Period T_w

For each randomly chosen data set ($i = 1, \ldots, 20$), we calculate $(\sigma_i)_{\widetilde{\boldsymbol{\Sigma}}}$, the annualized standard deviation of the sample portfolio return, by multiplying the monthly standard



(a) The mean of $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ from the static estimator $\widetilde{\mathbf{\Sigma}}$.

(b) The mean of $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ from the dynamic estimator $\widetilde{\mathbf{\Sigma}}$.

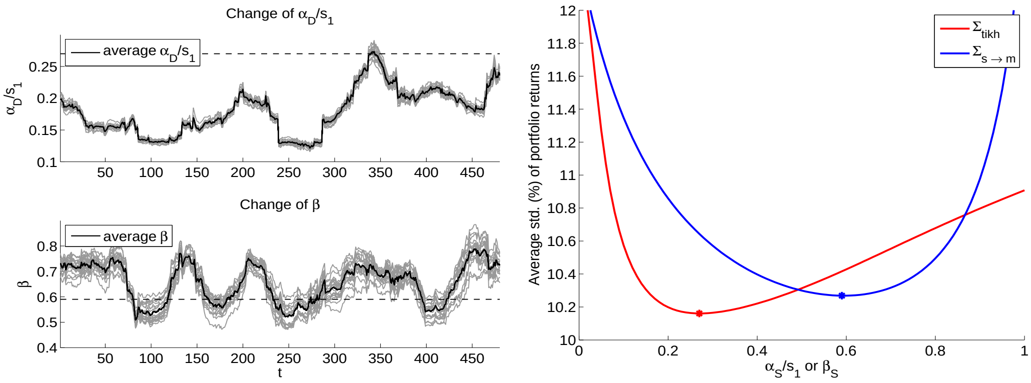
Figure 3.4: *GMV portfolios: The mean of* $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ *over different choice of* T_w.

deviation by $\sqrt{12}$. The subscript $\widetilde{\mathbf{\Sigma}}$ denotes the specific choice of covariance estimator. Figure 3.4(a) and Figure 3.4(b) show the means of $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ for the static estimators and the dynamic estimators. The standard deviations of the $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ from each estimator were at most 0.56 for all time periods, except for the occurrence of values up to 3.38 for $\widetilde{\mathbf{\Sigma}}_{sample}$

and up to 6.50 for $\widetilde{\mathbf{\Sigma}}_{s \to trun(\widehat{k}=3)}$, so the results did not seem sensitive to the particular choice of 100 stocks.

For most estimators, the $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$ decrease until a particular T_w and increase after that point, showing the advantage of using a sufficient amount of history but not too much out-of-date information. This is particularly evident for $\widetilde{\mathbf{\Sigma}}_{sample}$, since it assumes that all of its data are reliable. At the opposite extreme, $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}_{market}}$ from $\widetilde{\mathbf{\Sigma}}_{market}$ increases with T_w, which implies that the correlation among stocks cannot be fully explained by a single market index. For small values of k, $\widetilde{\mathbf{\Sigma}}_{trun}$ behaves like $\widetilde{\mathbf{\Sigma}}_{market}$, but performance can be improved by taking $k \approx 5$, making the estimator less sensitive to out-of-date information. The diagonal $\widetilde{\mathbf{\Sigma}}_V$ shows a better tolerance to out-of-date information than $\widetilde{\mathbf{\Sigma}}_{sample}$, which may imply that the sample variance estimation is less sensitive to the choice of T_w than the sample covariance estimation. The estimators that dynamically determine the filtering parameters ($\widetilde{\mathbf{\Sigma}}_{tikh}$, $\widetilde{\mathbf{\Sigma}}_{s \to m}$, $\widetilde{\mathbf{\Sigma}}_{s \to I}$, $\widetilde{\mathbf{\Sigma}}_{s \to trun(\widehat{k}=1)}$, $\widetilde{\mathbf{\Sigma}}_{RMT:repl}$, and $\widetilde{\mathbf{\Sigma}}_{RMT:trun}$) also show good tolerance. Therefore, modestly filtered factor structures are better at filtering the out-of-date information than a single factor or full factor structure, but all estimators benefit from an appropriate choice of window size.

Compared to the truncation-based estimators like $\widetilde{\mathbf{\Sigma}}_{RMT:trun}$ and $\widetilde{\mathbf{\Sigma}}_{trun}$, Tikhonov generally performs better when the *in-sample* period is shorter than its own optimal size, which is $T_w = 4$. This result can be explained by the characteristics of their filtering functions. While $\phi^2_{tikh}(\lambda_i)$ preserves the relative magnitudes of eigenvalues by gradual attenuation, $\phi^2_{RMT:trun}(\lambda_i)$ or $\phi^2_{trun}(\lambda_i)$ discard them all. Thus, when the smallest eigenvalues are still important, the Tikhonov filter empirically shows superiority. However, as



(a) Variation in the dynamic α_D and γ_D over the course of 20 experiments.

(b) Standard deviation of portfolio returns for various choices of α_S and γ_S.

Figure 3.5: *GMV portfolios: The performance of static and dynamic choice of* α *and* γ *in 20 experiments.*

noise level increases with longer T_w, the performance reverses.

Compared to the other shrinkage-based estimators, Tikhonov filtering $\phi^2_{tikh}(\lambda_i)$ preserves the smallest but still informative factors better than a single rate reduction by $\phi^2_{s\to m}(\lambda_i)$ and $\phi^2_{s\to trun}(\lambda_i)$ or a replacement with a constant value by $\phi^2_{RMT:repl}(\lambda_i)$ when T_w is relatively short ($T_w < 4$). On the other hand, for $T_w > 7$, it becomes evident that $\widetilde{\Sigma}_{s\to m}$, $\widetilde{\Sigma}_{s\to trun(\widehat{k}=1)}$, and $\widetilde{\Sigma}_{RMT:repl}$ show better performance than $\widetilde{\Sigma}_{tikh}$. This is because $\widetilde{\Sigma}_{tikh}$ has relatively weaker tolerance to the contamination by out-of-date information.

Stability of Tikhonov Parameter Choice

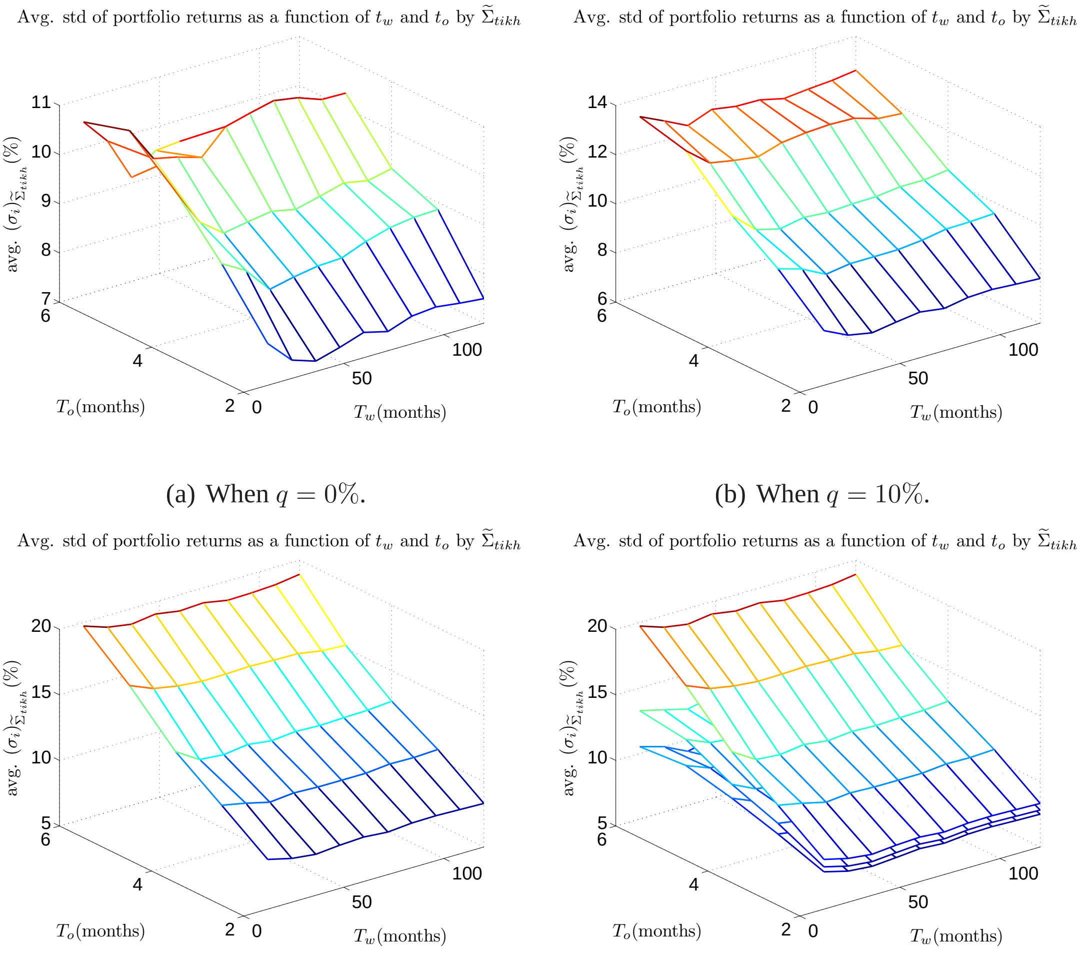
In this section, we evaluate the stability of our parameter choice methods from Section 3.4. For a particular choice of 100 stocks, we observe the change of the dynamic parameters α for $\widetilde{\Sigma}_{tikh}$ and γ for $\widetilde{\Sigma}_{s\to m}$. In this experiment, we set the window size as $T_w = 48$ because both estimators have the smallest mean value of $(\sigma_i)_{\widetilde{\boldsymbol{\Sigma}}}$ for that window size.

Figure 3.5(a) illustrates the change of the ratio of the dynamically chosen Tikhonov parameter α_D to the largest singular value s_1 of corr$_s[\boldsymbol{r}(t)]$, and the change of γ_D for $\widetilde{\boldsymbol{\Sigma}}_{s\to m}$. The results for 20 choices of the 100 stocks are shown, showing that both parameter choice methods for α_D and γ_D are quite stable during the whole experiment. The resulting annualized standard deviations of $(\sigma_i)_{\widetilde{\boldsymbol{\Sigma}}}$ range from 10.16% to 10.30% for $\widetilde{\boldsymbol{\Sigma}}_{tikh}$ and $\widetilde{\boldsymbol{\Sigma}}_{s\to m}$, for both the static and dynamically-determined parameters.

We repeated this numerical experiment keeping the ratio α/s_1 and the parameter γ constant over all time periods. (We use the notation α_S and γ_S for this statically determined parameter.) This static parameter choice may not be practical in real market trading, since we cannot access the future return information when we construct a portfolio. However, we can find a statically optimal ratio from this experiment for a comparison to α_D/s_1 and γ_D. Figure 3.5(b) shows how the standard deviation of portfolio returns changes as α_S/s_1 and γ_S increase. The optimal ratio α_S^*/s_1 was 0.27 with resulting standard deviation of portfolio returns 10.16%, and the optimal γ_S^* was 0.59 with resulting standard deviation 10.27%. These statically optimal values are represented by dashed lines in Figure 3.5(a). Therefore, we can see that both α_D/s_1 and γ_D remain near their statically optimal values α_S^*/s_1 and γ_S^*. Moreover, the static and varying α values produce similar risk variance.

3.6.2 MV Portfolio

Now, we observe the behavior of the MV portfolio resulting from each covariance esti-



(a) When $q = 0\%$.

(b) When $q = 10\%$.

(c) When $q = 20\%$.

(d) When $q = 0\%, 10\%$, and 20%.

Figure 3.6: *MV portfolios: The average annualized standard deviations* $(\sigma_i)_{\widetilde{\boldsymbol{\Sigma}}_{tikh}}$ *of portfolio returns as* in-sample *period* T_w *and* out-of-sample *period* T_o *changes with different settings of required portfolio return* q.

mator. In this experiment, we vary the *out-of-sample* period T_o and the required portfolio return q as well as the *in-sample* period T_w. We change T_o from 2 months to 6 months,[2]

[2] We omit the case of $T_o = 1$ month, since it gives us a trivial result that the portfolio returns are equal to the required portfolio return q making $(\sigma_i)_{\widetilde{\boldsymbol{\Sigma}}}$ zero for any covariance $\widetilde{\boldsymbol{\Sigma}}$ and any window size T_w. This is because $\boldsymbol{\mu}$ equals the realized stock returns $\mathbf{r}(t)$ in the *out-of-sample* period.

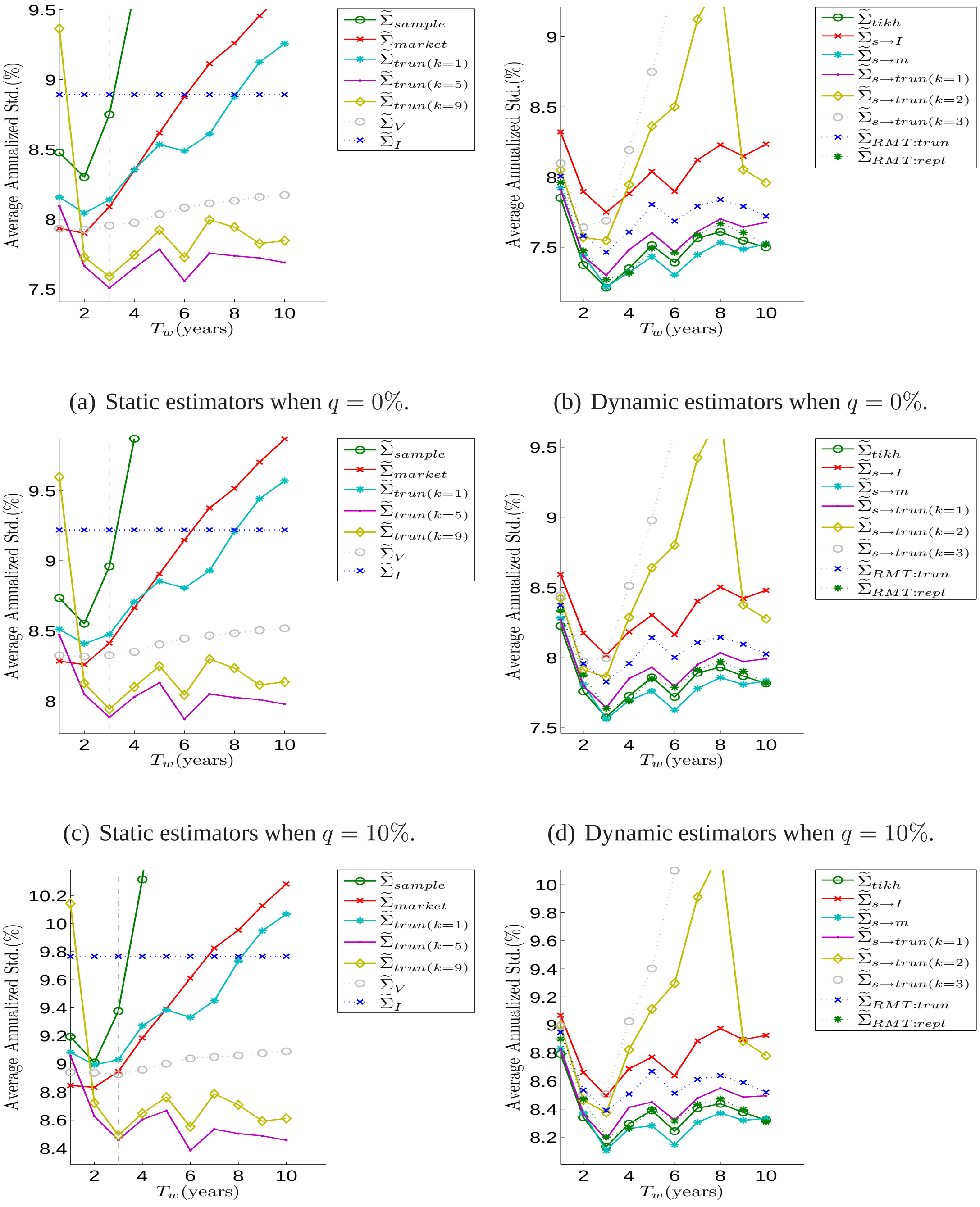
T_w from 1 year to 10 years, and q from 0 $\%$ to 20%. As we mentioned before, the performance of the MV portfolio is quite sensitive to the estimation of stock returns $\boldsymbol{\mu}$. In order to evaluate covariance estimation with no influence of mean estimation, we assume a perfect prediction of stock returns $\boldsymbol{\mu}$, which means we estimate $\boldsymbol{\mu}$ by the average $\boldsymbol{r}(t)$ during the *out-of-sample* period.

Effect of *out-of-sample* Period T_o

The *out-of-sample* period T_o determines how fast we react to the changes in the market. Figure 3.6 shows how the average $(\sigma_i)_{\tilde{\boldsymbol{\Sigma}}_{tikh}}$ changes as T_o and T_w vary, for $q = 0\%,\ 10\%,$ and 20%. We can see that $(\sigma_i)_{\tilde{\boldsymbol{\Sigma}}_{tikh}}$ has a tendency to increase as we hold the portfolio for longer T_o. Similar results were obtained for all other covariance estimators.

Effect of *in-sample* Period T_w

Similar to Figure 3.4 for the GMV experiment, we compared the mean of $(\sigma_i)_{\tilde{\boldsymbol{\Sigma}}}$ for different covariance estimators, varying T_w and q in Figure 3.7. Based on the result of Section 3.6.2, we fixed T_o as 2 months in order to compare the smallest standard deviations from the estimators. The behaviors of MV portfolios with respect to the change of T_w are very similar to the GMV portfolio for most covariance estimators. For example, as we observed for the previous GMV experiments, the MV portfolios in Figure 3.7 also suffered from lack of information when T_w was too short and suffered from out-of-date information when T_w was too long. This implies that the choice of window size T_w is very important for the MV portfolio as well as the GMV portfolio. Moreover, each estimator



(a) Static estimators when $q = 0\%$.

(b) Dynamic estimators when $q = 0\%$.

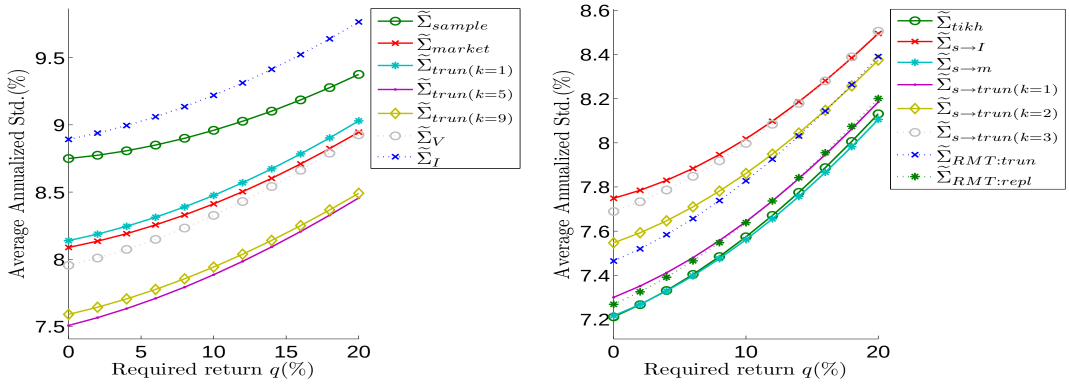
(c) Static estimators when $q = 10\%$.

(d) Dynamic estimators when $q = 10\%$.

(e) Static estimators when $q = 20\%$.

(f) Dynamic estimators when $q = 20\%$.

Figure 3.7: *MV portfolios: The mean of* $(\sigma_i)_{\widetilde{\Sigma}}$ *over different choice of* T_w *and* q *when* $T_o = 2$ *months.*



(a) Static estimators.

(b) Dynamic estimators.

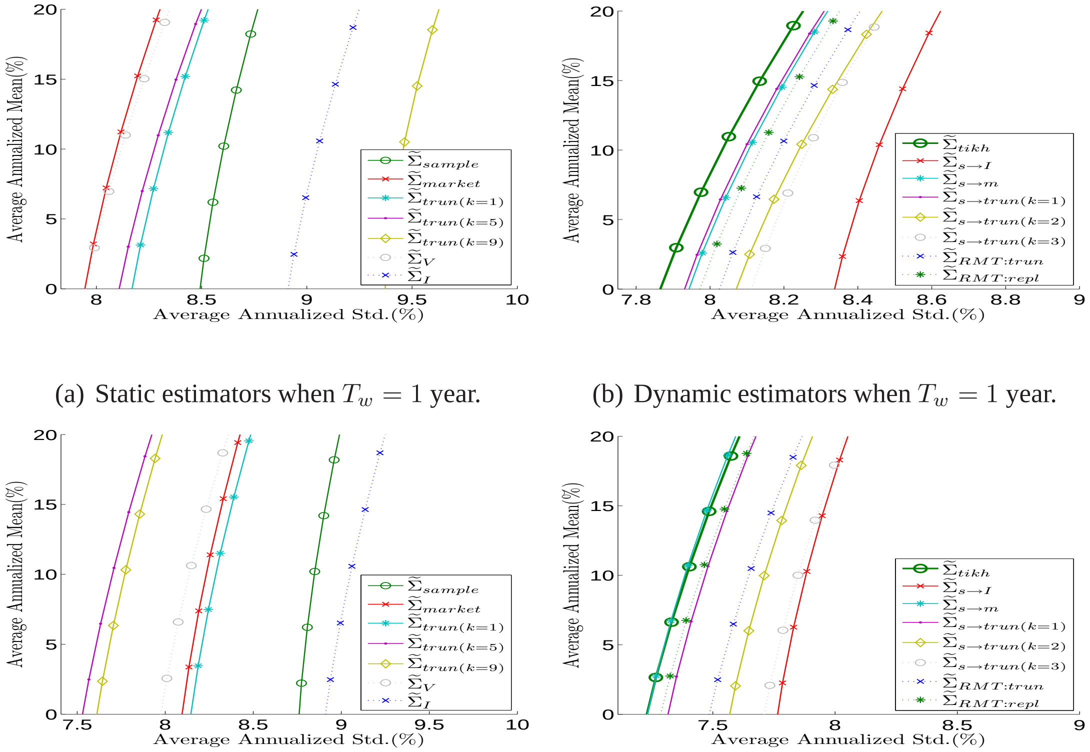
Figure 3.8: *MV portfolios: The average annualized $(\sigma_i)_{\widetilde{\Sigma}}$ versus required return q for each estimator when $T_o = 2$ months and $T_w = 3$ years.*

shows very similar shapes of curves for the GMV and the MV problems, except that the curves for the MV problems tend to shift upward as q increases.

However, in contrast to the GMV problem where most of competitive estimators have optimal T_w around 4 years, the optimal T_w for most estimators was around 3 years for the MV problem (Gray-colored vertical dot-dash lines indicate $T_w = 3$ years in Figure 3.7). This may be because they have different *out-of-sample* periods: $T_o = 1$ month for the GMV problem in Figure 3.4 and $T_o = 2$ months for the MV problem in Figure 3.7.

Effect of Required Portfolio Return q

Figure 3.6(d) summarizes the results from Figure 3.6(a) to Figure 3.6(c). As we can expect, the surfaces of $(\sigma_i)_{\widetilde{\Sigma}_{tikh}}$ move upward as q increases. For all the estimators $\widetilde{\Sigma}$ with particular choices of $T_o = 2$ months and $T_w = 3$ years, Figure 3.8 also shows that $(\sigma_i)_{\widetilde{\Sigma}}$ gradually increase as q increases from 0% to 20%, which explains a trade-off



(a) Static estimators when $T_w = 1$ year.

(b) Dynamic estimators when $T_w = 1$ year.

(c) Static estimators when $T_w = 3$ years.

(d) Dynamic estimators when $T_w = 3$ years.

Figure 3.9: *MV portfolios: The average annualized* $(\mu_i)_{\widetilde{\Sigma}}$ *versus average annualized* $(\sigma_i)_{\widetilde{\Sigma}}$.

between risk and return from the MV portfolio.

Efficiency of Portfolio

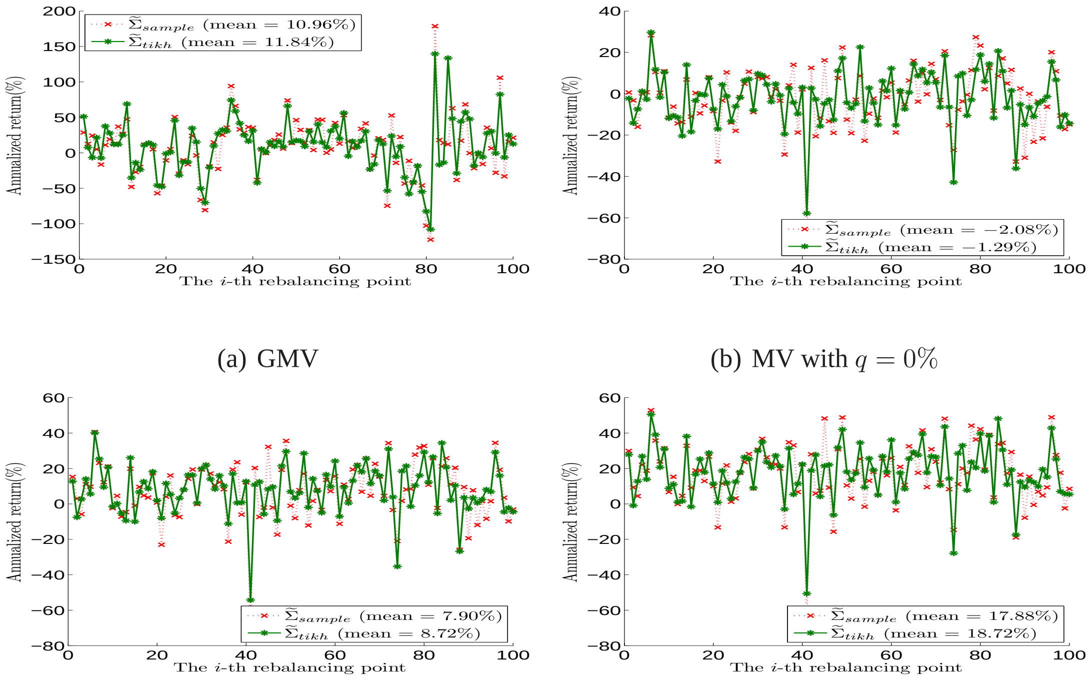
The mean-variance plot shows the efficiency of the MV portfolios. Let $(\mu_i)_{\widetilde{\Sigma}}$ denote the annualized mean of the realized portfolio returns in the i-th random choice of 100 stocks ($i = 1, \ldots, 20$). In order to evaluate the portfolio efficiency by each estimator, we compare the change of average $(\mu_i)_{\widetilde{\Sigma}}$ versus the change of average $(\sigma_i)_{\widetilde{\Sigma}}$, varying the required return q from 0% to 20%. Figure 3.9 presents the average of realized means and standard deviations of all the estimators for the cases of $T_o = 2$ months and $T_w = 1$ year or 3 years. Curves to the left of and above the others correspond to the more efficient portfolios.

When $T_w = 1$ year, where we have insufficient historical data, $\widetilde{\mathbf{\Sigma}}_{tikh}$ generates the most efficient portfolios (See Figure 3.9(b)). The shrinkage estimators with a target of a single factor like $\widetilde{\mathbf{\Sigma}}_{s\to m}$ and $\widetilde{\mathbf{\Sigma}}_{s\to trun(k=1)}$ are also efficient compared to other dynamic estimators. When $T_w = 3$ years, where we have near optimal historical data, $\widetilde{\mathbf{\Sigma}}_{tikh}$, $\widetilde{\mathbf{\Sigma}}_{s\to m}$, $\widetilde{\mathbf{\Sigma}}_{RMT:repl}$, and $\widetilde{\mathbf{\Sigma}}_{s\to m}$ generate relatively efficient portfolios (See Figure 3.9(d)).

3.6.3 Comparison of GMV and MV Portfolio

Now we observe how the covariance estimators affect the realized portfolio returns at every re-balancing point for the GMV and the MV problems. For instance, Figure 3.10 shows the fluctuations of the portfolio returns by $\widetilde{\mathbf{\Sigma}}_{sample}$ and $\widetilde{\mathbf{\Sigma}}_{tikh}$ at the first 100 re-balancing points when $T_w = 3$ years and $T_o = 2$ months. While the annualized returns of the GMV portfolios fluctuate around 11%, and the annualized returns of the MV portfolio fluctuate around their required return q. Note that the GMV mean return is greater than that for the MV portfolio with $q = 0\%$. Similarly, the standard deviations in Figure 3.4 are greater than the corresponding ones in Figure 3.7(a) and Figure 3.7(b).

On the other hand, for both GMV and MV, the $\widetilde{\mathbf{\Sigma}}_{tikh}$ portfolios have greater mean return and smaller variance than those from $\widetilde{\mathbf{\Sigma}}_{sample}$, which implies more efficient portfolios. This result is consistent with the plots of means versus standard deviations in Figure 3.9.



(a) GMV

(b) MV with $q = 0\%$

(c) MV with $q = 10\%$

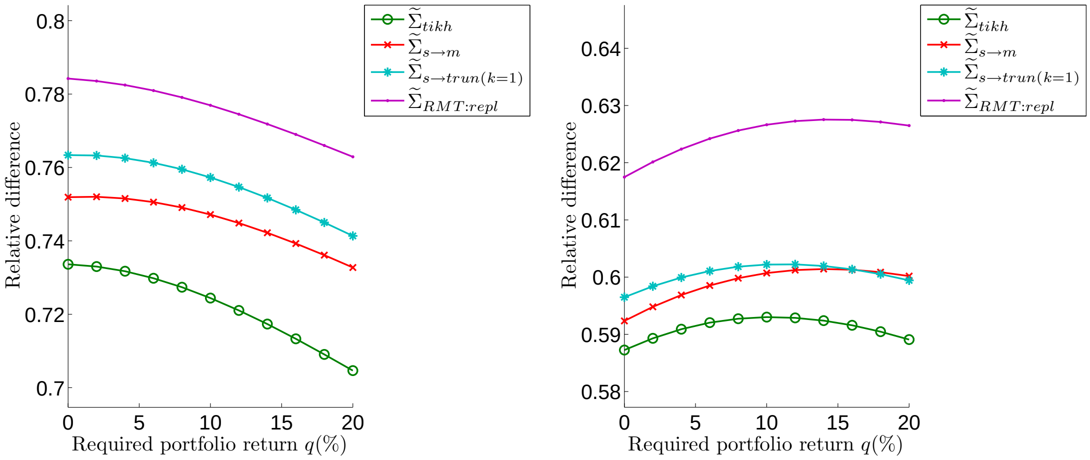
(d) MV with $q = 20\%$

Figure 3.10: *GMV and MV portfolios: The annualized portfolio returns at the rebalancing points for the GMV and the MV problem with different required returns* q.

3.6.4 Risk Prediction

Laloux et al. [54] showed empirically that their estimator $\widetilde{\Sigma}_{RMT:repl}$ predicts the risk more accurately than $\widetilde{\Sigma}_{sample}$. They simply divided the dataset into two equal time periods for *in-sample* and *out-of-sample* periods, and compared the estimated standard deviation $(\mathbf{w}^T\widetilde{\Sigma}\mathbf{w})^{\frac{1}{2}}$ from (3.1.1) to the realized standard deviation $(\sigma_i)_{\widetilde{\Sigma}}$ for the *out-of-sample* period. They assumed perfect prediction for means of stock returns as we did in Section 3.6.2.

We evaluate the accuracy of the risk prediction of each covariance estimator in a



(a) When $T_w = 1$ year.

(b) When $T_w = 3$ years.

Figure 3.11: *MV portfolios: The relative differences between average estimated risks and average realized risks by each covariance matrix, varying different required returns* q.

similar way. However, rather than following their equal division of *in-sample* and *out-of-sample* periods, we varied T_w with $T_o = 2$ months, and we simulated the re-balancing scenario as in Section 3.6.2. Finally, we compute the relative difference between the average estimated standard deviations from (3.1.1) and the average realized standard deviations for the most competitive estimators.

Figure 3.11 shows the relative difference for the case of $T_w = 1$ and 3 years which correspond to the case of insufficient historical data and the minimizer of average $(\sigma_i)_{\widetilde{\mathbf{\Sigma}}}$. The realized standard deviations were greater than the estimated standard deviations for all estimators. However, it turns out that $\widetilde{\mathbf{\Sigma}}_{tikh}$ has the smallest difference for both cases, giving us the best risk prediction.

3.7 Conclusion

In this study, we applied Tikhonov regularization to improve the covariance matrix estimate used in the Markowitz portfolio selection problem. We put the previous covariance estimators in a common framework based on the filtering function $\phi^2(\lambda_i)$ for the eigenvalues of $\text{corr}_s[\boldsymbol{r}(t)]$. The Tikhonov estimator $\widetilde{\boldsymbol{\Sigma}}_{tikh}$ attenuates smaller eigenvalues more intensely, which is a key difference between it and the other filter functions.

In order to choose an appropriate Tikhonov parameter α that determines the intensity of attenuation, we formulated an optimization problem minimizing the difference between $\text{corr}_s[\boldsymbol{\epsilon}_z(t)]$ and $\boldsymbol{I}_N$ based on the assumption that the unsystematic factors are uncorrelated.

We performed empirical experiments to evaluate covariance estimators. For the GMV portfolio selection problem, the Tikhonov choice gave the smallest average standard deviation of the return when the *in-sample* period was 3 or 4 years, and was not much worse than competitors for other periods. The choice of parameter was relatively stable. For the MV portfolio selection problem, the Tikhonov choice was among the most efficient portfolios and the best estimates of risk. Moreover, the Tikhonov estimator performs relatively well in the circumstance of insufficient historical data. We believe that this parameter selection method is quite promising relative to previously proposed methods.

Chapter 4

Constraint Reduction in Semidefinite Programming

In this chapter, we study matrix reduction in semidefinite programming (SDP). In interior point methods for constrained convex optimization, we can use the Schur complement matrix to solve a reduced linear system for each iteration.

Matrix reduction is applied to the Schur complement matrix. In contrast to the problems introduced in the previous chapters, the reduced parts of the matrix are neither error nor noise, but unnecessary constraints. These unnecessary constraints are inactive and do not make an important contribution to following the path toward the optimal solution, but still increase the computational load.

We present an infeasible primal-dual *predictor-corrector* interior point method for SDP with constraint reduction. Through experiments, we see the effect of matrix reduction and make important observations used in the next chapter to construct an algorithm with global convergence.

4.1 Introduction

Constraint reduction in interior point methods (IPMs) has been deeply studied especially for linear programming (LP) problems. That is because IPMs require many computations per iteration compared to the simplex method, but tend to require fewer iterations.

Prior work on constraint reduction in LP begins with Dantzig and Ye [16]. They developed a *build-up* variant of a dual *affine-scaling* algorithm. In their method, starting with a small working set, they add more constraints to the working set until the current step becomes feasible with respect to the full set of constraints. Tone [86] proposed an *active set* version of the dual potential-reduction algorithm by Ye [95]. This algorithm also starts with a small working set and adds constraints if the current working set does not sufficiently decrease the potential function. Kaliski and Ye [48] modified Tone's algorithm to exploit the structure of a large-scale transportation problems. Later, den Hertog, Roos, and Terlaky [22] proposed a *build-up-and-down* path following method with a logarithmic barrier function, which follows a central path defined by a small working set as long as it is feasible with respect to the full set of constraints. Once it becomes infeasible, the working set is updated appropriately, and it restarts from the previous iterate.

Tits, Absil, and Woessner [84] developed a new constraint reduced version of a *primal-dual affine-scaling* method (rPDAS) and Mehrotra's *predictor-corrector* method (rMPC). While previous constraint reduction schemes test the feasibility of the current working set with respect to the full set of constraints, their method adaptively updates the working set without any acceptability test. They proved global convergence and quadratic local convergence of rPDAS under a nondegeneracy assumption, but polynomial complexity was not proved. Later, Winternitz et al. [93] proved the global convergence of a new version of rMPC relaxing the assumptions of [84].

Adaptive constraint reduction has been applied to a series of optimization problems. Jung, O'Leary, and Tits [46] proposed a constrained reduction for training support vector machines (SVM), and Williams [92] applied preconditioning to SVM training to improve its efficiency. Later, Jung, O'Leary, and Tits [47] developed a constraint-reduced *affine-scaling* method for convex quadratic programming (QP), and verified its global convergence and quadratic local convergence.

In this study, we extend constraint reduction to a *predictor-corrector* method for diagonal block-structured SDP problems. The most computationally intensive step in an IPM for SDP is the construction of the Schur complement matrix. By ignoring unnecessary constraints, we can reduce the computational load for computing the Schur complement matrix, so that each iteration can finish with less cost.

We summarize the organization of this chapter: In Section 4.2, we present an IPM for SDP and discuss the main computational step. In Section 4.3, we see how block diagonal structure simplifies the computation, and present a constraint-reduced *predictor-corrector* algorithm. In Section 4.4, we demonstrate how well the proposed algorithm solves SDP problems. In Section 4.5, we summarize important observations from the experiments to guide a new algorithm introduced in Chapter 5. Before proceeding, we

$\mathcal{S}^n$	the set of $n \times n$ symmetric matrices
$\widetilde{\mathcal{S}}^n$	the set of $n \times n$ skew-symmetric matrices
$\mathcal{S}^n_+$	the set of $n \times n$ symmetric positive semidefinite matrices
$\mathcal{S}^n_{++}$	the set of $n \times n$ symmetric positive definite matrices
$\boldsymbol{X} \succ \boldsymbol{0}$	a positive definite matrix
$\boldsymbol{X} \succeq \boldsymbol{0}$	a positive semidefinite matrix
$\boldsymbol{A} \bullet \boldsymbol{B} = \text{tr}\left(\boldsymbol{A}\boldsymbol{B}^T\right)$	the dot-product of matrices
$\mu = (\boldsymbol{X} \bullet \boldsymbol{Z})/n$	the duality gap
$\text{vec}\,(\boldsymbol{X})$	the vectorization of a given matrix $\boldsymbol{X}$
$\text{mat}\,(\boldsymbol{x})$	the inverse of $\text{vec}\,(\boldsymbol{X})$
$\text{symm}\,(\boldsymbol{X}) = \frac{1}{2}(\boldsymbol{X} + \boldsymbol{X}^T)$	the symmetric part of $\boldsymbol{X}$
$x^{\overline{2}} = \frac{x(x+1)}{2}$	symmetric square
$\sqrt[\overline{2}]{y}$	symmetric square root: an inverse of $y = x^{\overline{2}}$

Table 4.1: *Notation for the SDP.*

highlight some special cases of SDP.

4.1.1 Special cases of SDP

We briefly explain the relation between SDP and other optimization problems [1]. We make use of the definitions in Table 4.1. The primal and dual SDP problems are as follows:

$$\text{Primal SDP:} \quad \min_{\boldsymbol{X}} \boldsymbol{C} \bullet \boldsymbol{X} \quad \text{s.t.} \;\; \boldsymbol{A}_i \bullet \boldsymbol{X} = b_i \text{ for } i = 1, \ldots, m, \;\; \boldsymbol{X} \succeq \boldsymbol{0}, \tag{4.1.1}$$

$$\text{Dual SDP:} \quad \max_{\boldsymbol{y}} \boldsymbol{b}^T \boldsymbol{y} \quad \text{s.t.} \;\; \sum_{i=1}^{m} y_i \boldsymbol{A}_i + \boldsymbol{Z} = \boldsymbol{C}, \;\; \boldsymbol{Z} \succeq \boldsymbol{0}, \tag{4.1.2}$$

where $\boldsymbol{C} \in \mathcal{S}^n$, $\boldsymbol{A}_i \in \mathcal{S}^n$, $\boldsymbol{X} \in \mathcal{S}^n$, and $\boldsymbol{Z} \in \mathcal{S}^n$.

To explain some special cases, the following property of a Schur complement matrix is useful.

[1] The book by Boyd and Vandenberghe [7] is a good reference for detailed explanation.

Lemma 4.1.1 (Schur Complement)**.** *When*

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12}^T & \mathbf{H}_{22} \end{bmatrix},$$

where $\mathbf{H}_{11} \succ \mathbf{0}$ *and* $\mathbf{H}_{22}$ *is symmetric, then* $\mathbf{H}$ *is positive (semi)definite if and only if* $(\mathbf{H}_{22} - \mathbf{H}_{12}^T\mathbf{H}_{11}^{-1}\mathbf{H}_{12})$ *is positive (semi)definite.*

Proof. See Theorem A.9 in [17, p.239]. □

First, LP and QP have a linear inequality constraint,

$$\boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c},$$

where $\boldsymbol{A}$ has m columns. It is easy to see the linear inequality constraint is a special case of (4.1.2) in which all the $\boldsymbol{A}_i$ and $\boldsymbol{C}$ are diagonal matrices.

Second, quadratically constrained quadratic programming (QCQP) has quadratic inequality constraints,

$$\boldsymbol{y}^T\boldsymbol{Q}_j\boldsymbol{y} + \boldsymbol{q}_j^T\boldsymbol{y} + c_j \leq 0 \quad \text{for } j = 1, \ldots, p,$$

where $\boldsymbol{Q}_j \in \mathcal{S}_+^m$. By Lemma 4.1.1, this is equivalent to

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{M}_j\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{M}_j^T & -c_j - \boldsymbol{q}_j^T\boldsymbol{y} \end{bmatrix} \succeq \mathbf{0},$$

where $\boldsymbol{Q}_j = \boldsymbol{M}_j^T\boldsymbol{M}_j$. We can rewrite this as

$$\sum_{i=1}^{m} y_i \begin{bmatrix} 0 & -\boldsymbol{m}_{ji} \\ -\boldsymbol{m}_{ji}^T & q_{ji} \end{bmatrix} + \boldsymbol{Z}_j = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & -c_j \end{bmatrix}, \quad \boldsymbol{Z}_j \succeq \mathbf{0},$$

where $\boldsymbol{m}_{ji}$ is the i-th column of $\boldsymbol{M}_j$, and q_{ji} is the i-th entry of $\boldsymbol{q}_j$. We can see that the quadratic constraint is the special case of (4.1.2) in which $\boldsymbol{A}_i$ contains the diagonal block whose elements in the last row and the last column are non-zeros.

Third, second order cone programming (SOCP) has inequality constraints,

$$\|\boldsymbol{M}_j\boldsymbol{y} + \boldsymbol{d}_j\| \leq \boldsymbol{q}_j^T\boldsymbol{y} + c_j, \quad \text{for } j = 1, \dots, p,$$

which is equivalent to

$$\begin{bmatrix} (\boldsymbol{q}_j^T\boldsymbol{y} + c_j)\boldsymbol{I} & \boldsymbol{M}_j\boldsymbol{y} + \boldsymbol{d}_j \\ (\boldsymbol{M}_j\boldsymbol{y} + \boldsymbol{d}_j)^T & \boldsymbol{q}_j^T\boldsymbol{y} + c_j \end{bmatrix} \succeq \boldsymbol{0},$$

by Lemma 4.1.1. We can rewrite the inequality above as

$$\sum_{i=1}^{m} y_i \begin{bmatrix} -q_{ji}\boldsymbol{I} & -\boldsymbol{m}_{ji} \\ -\boldsymbol{m}_{ji}^T & -q_{ji} \end{bmatrix} + \boldsymbol{Z}_j = \begin{bmatrix} c_j\boldsymbol{I} & \boldsymbol{d}_j \\ \boldsymbol{d}_j^T & c_j \end{bmatrix}, \quad \boldsymbol{Z}_j \succeq \boldsymbol{0}.$$

Hence, the second order inequality constraint is the special case of (4.1.2) in which $\boldsymbol{A}_i$ contains the diagonal block whose elements in the diagonal, the last row, and the last column are non-zeros (arrow-shaped).

Therefore, diagonal block-structured SDP includes LP, QP, QCQP, and SOCP as special cases. From this point of view, this study is a generalized version of [47, 84, 93].

4.2 Interior Point Methods for SDP

We discuss how standard IPMs find an optimal solution of SDP. For more details, see, for example, [17, 45].

4.2.1 Interior Point Methods for SDP with symmetrization

We assume that all the constraint matrices $\boldsymbol{A}_i$ for $i = 1, \ldots, m$ are independent. This assumption guarantees a unique direction which will be introduced now. In addition, we assume that the primal and dual SDP problems (4.1.1) and (4.1.2) have finite optimal solutions with equal optimal values. Under this assumption, $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$ is an optimal solution of (4.1.1) and (4.1.2) if and only if it satisfies

$$\boldsymbol{A}_i \bullet \boldsymbol{X} = b_i \quad \text{for } i = 1, \ldots, m, \tag{4.2.1}$$

$$(\sum_{i=1}^{m} y_i \boldsymbol{A}_i) + \boldsymbol{Z} = \boldsymbol{C}, \tag{4.2.2}$$

$$\boldsymbol{X} \bullet \boldsymbol{Z} = 0, \tag{4.2.3}$$

$$\boldsymbol{X} \succeq \boldsymbol{0}, \ \ \boldsymbol{Z} \succeq \boldsymbol{0}. \tag{4.2.4}$$

A duality gap is the difference between the primal and dual objective values for a given point $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$. For simplicity of notation, we measure the duality gap by μ defined as

$$\mu := (\boldsymbol{C} \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y})/n.$$

For a feasible solution satisfying (4.2.1) and (4.2.2), the duality gap μ can be computed as

$$\begin{aligned}
\mu &= \frac{1}{n}(\boldsymbol{C} \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y}) = \frac{1}{n}\left((\sum_{i=1}^{m} y_i \boldsymbol{A}_i + \boldsymbol{Z}) \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{m} y_i (\boldsymbol{A}_i \bullet \boldsymbol{X}) + \boldsymbol{Z} \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{m} y_i b_i + \boldsymbol{Z} \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y}\right) = \frac{1}{n}(\boldsymbol{b}^T \boldsymbol{y} + \boldsymbol{Z} \bullet \boldsymbol{X} - \boldsymbol{b}^T \boldsymbol{y}) = \frac{1}{n}(\boldsymbol{X} \bullet \boldsymbol{Z}).
\end{aligned}$$

So, (4.2.3) implies that the optimal values for the primal and dual problems are equal, as we assumed.

Primal-dual IPMs for SDPs make use of the following system of equations to define the Newton step and to measure closeness to optimality:

$$\boldsymbol{A}_i \bullet \Delta\boldsymbol{X} = r_{pi} \quad \text{for } i = 1, \ldots, m, \tag{4.2.5}$$

$$(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i) + \Delta\boldsymbol{Z} = \boldsymbol{R}_d, \tag{4.2.6}$$

$$\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z} = \boldsymbol{R}_c, \tag{4.2.7}$$

where the primal residual, dual residual, and complementarity residual are defined by

$$r_{pi} = b_i - \boldsymbol{A}_i \bullet \boldsymbol{X} \quad \text{for } i = 1, \ldots, m, \tag{4.2.8}$$

$$\boldsymbol{R}_d = \boldsymbol{C} - \boldsymbol{Z} - \sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i, \tag{4.2.9}$$

$$\boldsymbol{R}_c = \overline{\mu}\boldsymbol{I} - \boldsymbol{X}\boldsymbol{Z}, \tag{4.2.10}$$

and $\overline{\mu}$ defines the current target duality gap on the central path. The equation (4.2.7) is motivated by the goal of computing $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Z}$ such that

$$(\boldsymbol{X} + \Delta\boldsymbol{X})(\boldsymbol{Z} + \Delta\boldsymbol{Z}) = \overline{\mu}\boldsymbol{I}.$$

When this equation is satisfied, the duality gap becomes

$$\frac{1}{n}(\boldsymbol{X} + \Delta\boldsymbol{X}) \bullet (\boldsymbol{Z} + \Delta\boldsymbol{Z}) = \frac{1}{n}\text{tr}\,((\boldsymbol{X} + \Delta\boldsymbol{X})(\boldsymbol{Z} + \Delta\boldsymbol{Z})) = \frac{1}{n}\text{tr}\,(\overline{\mu}\boldsymbol{I}) = \overline{\mu}.$$

That is why we call $\overline{\mu}$ the target duality gap. Note that the term $\Delta\boldsymbol{X}\Delta\boldsymbol{Z}$ is ignored by linearization.

By solving (4.2.5)-(4.2.7) setting $\overline{\mu} = 0$, we can find a direction $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ for an updated point $(\boldsymbol{X} + \Delta\boldsymbol{X}, \boldsymbol{y} + \Delta\boldsymbol{y}, \boldsymbol{Z} + \Delta\boldsymbol{Z})$ to satisfy (4.2.1)-(4.2.3) ignoring the linearization error $\Delta\boldsymbol{X}\Delta\boldsymbol{Z}$. However, we may not be able take a full step in this direction due to the semidefinite inequality constraints $\boldsymbol{X} \succeq \boldsymbol{0}$ and $\boldsymbol{Z} \succeq \boldsymbol{0}$ in (4.2.4). So, we find the longest step length $\theta \in [0, 1]$ for which the inequality constraints are still satisfied, so that the point is updated as

$$\boldsymbol{X}^{+} = \boldsymbol{X} + \theta\Delta\boldsymbol{X}, \;\; \boldsymbol{y}^{+} = \boldsymbol{y} + \theta\Delta\boldsymbol{y}, \;\; \boldsymbol{Z}^{+} = \boldsymbol{Z} + \theta\Delta\boldsymbol{Z}.$$

We repeat this process until a given tolerance is satisfied. This algorithm is called as an *affine-scaling* method. Alternatively, we can solve (4.2.5)-(4.2.7), decreasing the target duality gap $\overline{\mu}$. This method is a *path-following* method since the iterates follow a central path, defined as the set of points satisfying $\boldsymbol{XZ} = \overline{\mu}\,\boldsymbol{I}$. Practically, most effective methods are *predictor-corrector methods*, in which a predictor step solves (4.2.5)-(4.2.7) setting $\overline{\mu} = 0$ to estimate a target duality gap $\overline{\mu}$, and a corrector step solves the equations again using the estimated duality gap. All of these methods are categorized as IPMs. In this work, we apply constraint reduction to a *predictor-corrector* method.

Specially in SDP, IPMs require a symmetrization process. Since the solution of (4.2.7) is not necessarily symmetric, we replace $\Delta\boldsymbol{X}$ with its symmetric part

$$\Delta\boldsymbol{X} \leftarrow \frac{1}{2}(\Delta\boldsymbol{X} + \Delta\boldsymbol{X}^{T})$$

after solving (4.2.5)-(4.2.7). The solution with this symmetrization is called the HKM direction, named after Helmberg, Kojima, and Monteiro [38, 51, 61]. Note that $\Delta\boldsymbol{Z}$ is always symmetric by (4.2.6). Thus, we effectively solve the equations

$$\boldsymbol{A}_i \bullet \Delta\dot{\boldsymbol{X}} = r_{pi} \quad \text{for } i = 1, \dots, m, \tag{4.2.11}$$

$$(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i) + \Delta\boldsymbol{Z} = \boldsymbol{R}_d, \tag{4.2.12}$$

$$\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\dot{\boldsymbol{X}}\boldsymbol{Z} = \boldsymbol{R}_c, \tag{4.2.13}$$

$$\Delta\dot{\boldsymbol{X}} = \Delta\boldsymbol{X} + \boldsymbol{W}, \tag{4.2.14}$$

where $\Delta\boldsymbol{X} \in \mathcal{S}^n$ and $\boldsymbol{W} \in \widetilde{\mathcal{S}}^n$. By (4.2.14), $\Delta\boldsymbol{X}$ is the symmetric part of $\Delta\dot{\boldsymbol{X}}$, so

$$\Delta\boldsymbol{X} = \text{ symm}\left(\Delta\dot{\boldsymbol{X}}\right).$$

Since $\boldsymbol{A}_i \in \mathcal{S}^n$ and $\boldsymbol{W} \in \widetilde{\mathcal{S}}^n$, $\boldsymbol{A}_i \bullet \boldsymbol{W} = 0$. By this property, the symmetrized direction $\Delta\boldsymbol{X}$ from (4.2.11)-(4.2.14) also satisfies (4.2.5), so the primal residual is the same with or without symmetrization.

4.2.2 Predictor-Corrector Algorithm

To solve (4.2.11)-(4.2.14), we vectorize the equations and reduce the equations to an equation involving the Schur complement matrix. For further discussion, let us briefly introduce a vectorization operation and Kronecker product. A vectorization, $\text{vec}\,(\boldsymbol{X}) \in \mathbb{R}^{n^2}$ for a matrix $\boldsymbol{X} \in \mathbb{X}^{n\times n}$ is defined as

$$\text{vec}\,(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix},$$

where $\boldsymbol{x}_i$ is the i-th column of $\boldsymbol{X}$. The vectorized variables will be denoted by lower-case letters: for example, $\boldsymbol{x} = \text{ vec}\,(\boldsymbol{X})$.

For $\boldsymbol{G} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{H} \in \mathbb{R}^{s\times t}$, the Kronecker product ($\otimes$) is defined as

$$\boldsymbol{G} \otimes \boldsymbol{H} = \begin{bmatrix} g_{11}\boldsymbol{H} & \dots & g_{1q}\boldsymbol{H} \\ \vdots & \ddots & \vdots \\ g_{p1}\boldsymbol{H} & \dots & g_{pq}\boldsymbol{H} \end{bmatrix},$$

where g_{ij} is the (i,j) entry of $\boldsymbol{G}$. Along with the vectorization, we will frequently use the following properties of the Kronecker product. For appropriate size of matrices,

$$\begin{aligned} (\boldsymbol{E} \otimes \boldsymbol{F})(\boldsymbol{G} \otimes \boldsymbol{H}) &= (\boldsymbol{EG}) \otimes (\boldsymbol{FH}), \\ (\boldsymbol{E} \otimes \boldsymbol{F})^{-1} &= \boldsymbol{E}^{-1} \otimes \boldsymbol{F}^{-1}, \\ (\boldsymbol{E} \otimes \boldsymbol{F})^{T} &= \boldsymbol{E}^{T} \otimes \boldsymbol{F}^{T}, \\ (\boldsymbol{E} \otimes \boldsymbol{F})\,\text{vec}\,(\boldsymbol{X}) &= \text{vec}\left(\boldsymbol{EXF}^{T}\right). \end{aligned}$$

Using the vectorization, we define $\mathcal{A} \in \mathbb{R}^{m\times n^2}$, containing all $\text{vec}\,(\boldsymbol{A}_i)$, as

$$\mathcal{A} = \begin{bmatrix} \text{vec}\,(\boldsymbol{A}_1)^T \\ \vdots \\ \text{vec}\,(\boldsymbol{A}_m)^T \end{bmatrix}.$$

With the matrix $\mathcal{A}$, by using the vectorization and the Kronecker product, we vectorize the equations (4.2.11)-(4.2.13) as

$$\mathcal{A}\Delta\dot{\boldsymbol{x}} = \boldsymbol{r}_p, \tag{4.2.15}$$

$$\mathcal{A}^T\Delta\boldsymbol{y} + \Delta\boldsymbol{z} = \boldsymbol{r}_d, \tag{4.2.16}$$

$$(\boldsymbol{X} \otimes \boldsymbol{I})\Delta\boldsymbol{z} + (\boldsymbol{I} \otimes \boldsymbol{Z})\Delta\dot{\boldsymbol{x}} = \boldsymbol{r}_c, \tag{4.2.17}$$

where

$$\boldsymbol{r}_p = \boldsymbol{b} - \mathcal{A}\boldsymbol{x}, \tag{4.2.18}$$

$$\boldsymbol{r}_d = \boldsymbol{c} - \boldsymbol{z} - \mathcal{A}^T\boldsymbol{y}, \tag{4.2.19}$$

$$\boldsymbol{r}_c = \text{ vec}\left(\overline{\mu}\boldsymbol{I} - \boldsymbol{X}\boldsymbol{Z}\right), \tag{4.2.20}$$

where $\boldsymbol{r}_p \in \mathbb{R}^m$ contains primal residuals r_{pi} for $i = 1, \ldots, m$.

Using Gauss elimination, we can reduce the equations. First, we rewrite (4.2.16) as

$$\Delta\boldsymbol{z} = \boldsymbol{r}_d - \mathcal{A}^T\Delta\boldsymbol{y} \tag{4.2.21}$$

By substituting $\Delta\boldsymbol{z}$ from (4.2.21) into (4.2.17), we have

$$(\boldsymbol{X} \otimes \boldsymbol{I})(\boldsymbol{r}_d - \mathcal{A}^T\Delta\boldsymbol{y}) + (\boldsymbol{I} \otimes \boldsymbol{Z})\Delta\dot{\boldsymbol{x}} = \boldsymbol{r}_c,$$

$$(\boldsymbol{I} \otimes \boldsymbol{Z})\Delta\dot{\boldsymbol{x}} = (\boldsymbol{X} \otimes \boldsymbol{I})(\mathcal{A}^T\Delta\boldsymbol{y} - \boldsymbol{r}_d) + \boldsymbol{r}_c.$$

By multiplying $(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})$ to the left of both sides, we have

$$\Delta\dot{\boldsymbol{x}} = (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})(\mathcal{A}^T\Delta\boldsymbol{y} - \boldsymbol{r}_d) + (\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c. \tag{4.2.22}$$

Finally, by substituting $\Delta\dot{\boldsymbol{x}}$ from the equation above to (4.2.15), we have

$$\mathcal{A}\Delta\dot{\boldsymbol{x}} = \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c - \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})(\boldsymbol{r}_d - \mathcal{A}^T\Delta\boldsymbol{y}) = \boldsymbol{r}_p,$$

$$\mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y} = \boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c.$$

Thus, with Schur complement matrix $\boldsymbol{M}$, we have a reduced linear equation,

$$\boldsymbol{M}\Delta\boldsymbol{y} = \boldsymbol{g}, \tag{4.2.23}$$

1. Input : $(\boldsymbol{X}^0, \boldsymbol{y}^0, \boldsymbol{Z}^0)$ (initial value)
2. Repeat until convergence criteria are satisfied: For $k = 0, 1, \ldots,$
   - (a) $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \leftarrow (\boldsymbol{X}^k, \boldsymbol{y}^k, \boldsymbol{Z}^k)$
   - (b) Setting $\overline{\mu} = 0$, compute $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ by (4.2.21)-(4.2.23).
   - (c) Find the longest step length $\overline{\theta}$ such that $\overline{\boldsymbol{X}} \succeq \mathbf{0}$ and $\overline{\boldsymbol{Z}} \succeq \mathbf{0}$ where $\overline{\boldsymbol{X}} = \boldsymbol{X} + \overline{\theta}\Delta\boldsymbol{X}, \;\; \overline{\boldsymbol{Z}} = \boldsymbol{Z} + \overline{\theta}\Delta\boldsymbol{Z}.$
   - (d) Compute a target duality gap $\overline{\mu} \leftarrow (\overline{\boldsymbol{X}} \bullet \overline{\boldsymbol{Z}})/n$.
   - (e) Using the updated target duality gap $\overline{\mu}$, compute $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ by (4.2.21)-(4.2.23).
   - (f) Find the longest step length θ such that $\boldsymbol{X}^+ \succeq \mathbf{0}$ and $\boldsymbol{Z}^+ \succeq \mathbf{0}$ where $\boldsymbol{X}^+ = \boldsymbol{X} + \theta\Delta\boldsymbol{X}, \;\; \boldsymbol{y}^+ = \boldsymbol{y} + \theta\Delta\boldsymbol{y}, \;\; \boldsymbol{Z}^+ = \boldsymbol{Z} + \theta\Delta\boldsymbol{Z}.$
   - (g) $(\boldsymbol{X}^{(k+1)}, \boldsymbol{y}^{(k+1)}, \boldsymbol{Z}^{(k+1)}) \leftarrow (\boldsymbol{X}^+, \boldsymbol{y}^+, \boldsymbol{Z}^+)$.
   - (h) Update $\boldsymbol{r}_p$ and $\boldsymbol{r}_d$ by (4.2.18) - (4.2.19).

Table 4.2: *Constraint-reduced Predictor-corrector method.*

where

$$\boldsymbol{M} = \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T,$$

$$\boldsymbol{g} = \boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c.$$

We can then compute $\Delta\dot{\boldsymbol{x}}$ and $\Delta\boldsymbol{z}$ by (4.2.22) and (4.2.21).

Using equations (4.2.21)-(4.2.23), we establish the *predictor-corrector* algorithm for SDP as Table 4.2 similar to [17, Section 7.6]. In the predictor step, we solve the equations setting $\overline{\mu} = 0$. With the predictor direction $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$, we determine the longest step length $\overline{\theta}$ which makes $\boldsymbol{X} + \overline{\theta}\Delta\boldsymbol{X} \succeq 0$ and $\boldsymbol{Z} + \overline{\theta}\Delta\boldsymbol{Z} \succeq 0$. Then, we compute the duality gap for $(\boldsymbol{X} + \overline{\theta}\Delta\boldsymbol{X})$ and $(\boldsymbol{Z} + \overline{\theta}\Delta\boldsymbol{Z})$, and we use this estimate as a target duality gap $\overline{\mu}$ for the corrector step. In the corrector step, with the estimated target duality gap, we solve the system again and take the longest step θ in the resulting correction direction which makes $\boldsymbol{X} + \theta\Delta\boldsymbol{X} \succeq \mathbf{0}$ and $\boldsymbol{Z} + \theta\Delta\boldsymbol{Z} \succeq \mathbf{0}$.

Note that we compute the Schur complement matrix $\widehat{\boldsymbol{M}}$ only once for each iteration, and use it twice for the predictor step and the corrector step. This is because we use the predictor step only to estimate the target duality gap $\overline{\mu}$ without updating $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$.

4.3 Constraint-Reduced Predictor-Corrector Method for Block-Diagonal-Structured SDP

4.3.1 Block Structure

In this work, we focus on problems in which the matrices $\boldsymbol{A}_i$ and $\boldsymbol{C}$ are block diagonal:

$$\boldsymbol{A}_i = \begin{bmatrix} \boldsymbol{A}_{i1} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{A}_{ip} \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{C}_p \end{bmatrix},$$

where $\boldsymbol{A}_{ij}, \boldsymbol{C}_j \in \mathcal{S}^{n_j}$ for $i = 1, \dots, m$ and $j = 1, \dots, p$. Then, we define a matrix $\mathcal{A}_j \in \mathbb{R}^{m \times n_j^2}$ containing all $\operatorname{vec}(\boldsymbol{A}_{ij})$ as

$$\mathcal{A}_j = \begin{bmatrix} \operatorname{vec}(\boldsymbol{A}_{ij})^T \\ \vdots \\ \operatorname{vec}(\boldsymbol{A}_{mj})^T \end{bmatrix}.$$

For such problems, there is a block diagonal optimal solution $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$. This is because any nonzero elements outside of the diagonal block of $\boldsymbol{Z}$ immediately violate the dual constraint of (4.1.2), and nonzero elements outside of the diagonal blocks in $\boldsymbol{X}$ do not make any contribution to minimize the primal objective value $\boldsymbol{C} \bullet \boldsymbol{X}$. So we will require our iterates to have the form

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{X}_p \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{Z}_p \end{bmatrix}.$$

Using this block structure, the Schur complement matrix $\boldsymbol{M}$ in (4.2.23) can be computed as

$$\boldsymbol{M} = \sum_{j=1}^{p} \boldsymbol{M}_j,$$

where

$$\boldsymbol{M}_j = \mathcal{A}_j(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})\mathcal{A}_j^T.$$

Hence, each element $(\boldsymbol{M}_j)_{lh}$ of $\boldsymbol{M}_j$ can be computed as

$$(\boldsymbol{M}_j)_{lh} = (\boldsymbol{X}_j \, \boldsymbol{A}_{lj} \, \boldsymbol{Z}_j^{-1}) \bullet \boldsymbol{A}_{hj} \tag{4.3.1}$$

where $1 \leq l \leq m,\ l \leq h \leq m,\ 1 \leq j \leq p$.

Suppose that $\boldsymbol{A}_{ij}$ is dense.[2] Then the cost of computing the entire Schur complement matrix $\boldsymbol{M}$, including Cholesky factorization of $\boldsymbol{Z}_j$, is

$$\sum_{j=1}^{p}(4m + 1/3)n_j^3 + 2m^2n_j^2 \text{ operations.} \tag{4.3.2}$$

The computation of the Schur complement matrix is the most expensive part of IPMs for SDP and is $O(mn^3 + m^2n^2)$. It is our goal to drop the matrices $\boldsymbol{M}_j$ which do not play important roles in the Schur complement matrix $\boldsymbol{M}$, so that we reduce the computational

[2] Refer to Fujisawa, Kojima, and Nakata [26] to see how to exploit the sparsity of $\boldsymbol{A}_{ij}$

cost. In the next section, we classify the blocks into active and inactive blocks, and discuss why the latter can be dropped.

4.3.2 Active and Inactive Blocks

From the optimality condition (4.2.3), we can see that

$$r_x + r_z \leq n,$$

where r_x and r_z are the ranks[3] of an optimal solution $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$. This implies that there may exist blocks $\boldsymbol{X}_j^*$ and $\boldsymbol{Z}_j^*$ such that $\boldsymbol{X}_j^* = \mathbf{0}$ and $\boldsymbol{Z}_j^*$ has full rank, so $\boldsymbol{Z}_j \succ \mathbf{0}$ and $\boldsymbol{Z}_j$ is in the interior of the semidefinite cone. We will say that such sub-blocks are *inactive* and the other blocks are *active*.

For an inactive block, $(\boldsymbol{X}_j^* \otimes \boldsymbol{Z}_j^{*-1}) = \mathbf{0}$. We use this fact to guide our algorithm: we try to find blocks $(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})$ having norms small enough to ignore in forming $\boldsymbol{M}$.

Let us assume that we have a criterion to identify inactive and active blocks in a given $\boldsymbol{X}$ and $\boldsymbol{Z}$. Without loss of generality, we assume that the first $\widehat{p}$ blocks are active and the remaining of $\tilde{p}$ blocks are inactive. We let $\widehat{\boldsymbol{A}}_i$ and $\widetilde{\boldsymbol{A}}_i$ denote the active and inactive

[3] According to Alizadeh, Haeberly, and Overton [1, Theorem 6 in p.9], for a nondegenerate optimal solution,

$$n - \sqrt[2]{n^2 - m} \leq r_x \leq \sqrt[2]{m},$$

$$n - \sqrt[2]{m} \leq r_z \leq \sqrt[2]{n^2 - m}.$$

blocks of $\boldsymbol{A}_i$, so

$$\widehat{\boldsymbol{A}}_i = \begin{bmatrix} \boldsymbol{A}_{i1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{A}_{i\widehat{p}} \end{bmatrix}, \quad \widetilde{\boldsymbol{A}}_i = \begin{bmatrix} \boldsymbol{A}_{i(\widehat{p}+1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{A}_{ip} \end{bmatrix},$$

where $\widehat{\boldsymbol{A}}_i \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $\widetilde{\boldsymbol{A}}_i \in \mathbb{R}^{\widetilde{n}\times\widetilde{n}}$, and $n = \widehat{n} + \widetilde{n}$. Furthermore, let denote $\widehat{n}_j$ and $\widetilde{n}_j$ denote the size of active and inactive blocks, so that

$$\widehat{n} = \sum_{j=1}^{\widehat{p}} \widehat{n}_j, \qquad \widetilde{n} = \sum_{j=1}^{\tilde{p}} \widetilde{n}_j.$$

In a similar way, block matrices $(\widehat{\boldsymbol{X}}, \widetilde{\boldsymbol{X}})$, $(\widehat{\boldsymbol{Z}}, \widetilde{\boldsymbol{Z}})$, $(\widehat{\boldsymbol{R}}_d, \widetilde{\boldsymbol{R}}_d)$, and $(\widehat{\boldsymbol{R}}_c, \widetilde{\boldsymbol{R}}_c)$ are also defined. We also define $\widehat{\mathcal{A}} \in \mathbb{R}^{m\times\widehat{n}^2}$ and $\widetilde{\mathcal{A}} \in \mathbb{R}^{m\times\widetilde{n}^2}$ as

$$\widehat{\mathcal{A}} = \begin{bmatrix} \operatorname{vec}\left(\widehat{\boldsymbol{A}}_1\right)^T \\ \vdots \\ \operatorname{vec}\left(\widehat{\boldsymbol{A}}_m\right)^T \end{bmatrix}, \quad \widetilde{\mathcal{A}} = \begin{bmatrix} \operatorname{vec}\left(\widetilde{\boldsymbol{A}}_1\right)^T \\ \vdots \\ \operatorname{vec}\left(\widetilde{\boldsymbol{A}}_m\right)^T \end{bmatrix}.$$

Then we can expand $\boldsymbol{M}$ into active and inactive parts as

$$\boldsymbol{M} = \widehat{\boldsymbol{M}} + \widetilde{\boldsymbol{M}},$$

where

$$\widehat{\boldsymbol{M}} = \widehat{\mathcal{A}}(\widehat{\boldsymbol{X}} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T,$$
$$\widetilde{\boldsymbol{M}} = \widetilde{\mathcal{A}}(\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T.$$

If $\|(\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\|$ is small, we expect $\widetilde{\boldsymbol{M}}$ is also negligible and we can omit it when we solve the linear system.

4.3.3 Constraint-Reduced Predictor-Corrector Method

Now, we consider the constraint-reduced linear system

$$\widehat{\boldsymbol{M}}\Delta\boldsymbol{y} = \boldsymbol{g}, \tag{4.3.3}$$

where we replace $\boldsymbol{M}$ in (4.2.23) with $\widehat{\boldsymbol{M}}$. So, we solve

$$\left(\widehat{\mathcal{A}}(\widehat{\boldsymbol{X}} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T\right)\Delta\boldsymbol{y} = \boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c.$$

In addition, $\Delta\dot{\boldsymbol{x}}$ and $\Delta\boldsymbol{z}$ are computed by

$$\Delta\dot{\boldsymbol{x}} = (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y} - (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d + (\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c, \tag{4.3.4}$$

$$\Delta\boldsymbol{z} = \boldsymbol{r}_d - \mathcal{A}^T\Delta\boldsymbol{y}. \tag{4.3.5}$$

After solving these equations, we can obtain the HKM direction by computing $\Delta\boldsymbol{X}$ as

$$\Delta\boldsymbol{X} = \text{ symm}\left(\Delta\dot{\boldsymbol{X}}\right). \tag{4.3.6}$$

Using the equations (4.3.3)-(4.3.6), we can develop a *predictor-corrector* method. Our new algorithm takes an additional input parameter, the threshold κ, by which active and inactive constraint blocks are classified: If $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\| > \kappa$, then we assume the block is active. Otherwise, it is assumed inactive.

Thus, we modify step 2.(b) and 2.(e) of the algorithm in Table 4.2:

2.(b)' Initially, $\widehat{\boldsymbol{M}} \leftarrow \boldsymbol{0}$. For the j-th block where $j = 1, \ldots, p$,
$\widehat{\boldsymbol{M}} \leftarrow \widehat{\boldsymbol{M}} + \mathcal{A}_j(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})\mathcal{A}_j{}^T$ if $\|\boldsymbol{X}_i \otimes \boldsymbol{Z}_i^{-1}\| \geq \kappa$.
Setting $\overline{\mu} = 0$, compute $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ using (4.3.3) - (4.3.6) with $\widehat{\boldsymbol{M}}$ in place of $\boldsymbol{M}$.

2.(e)' Using the updated target duality gap $\overline{\mu}$, compute $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ using (4.3.3) - (4.3.6) with $\widehat{\boldsymbol{M}}$ in place of $\boldsymbol{M}$.

Problem	Data file	m	n	$\#$ of blocks	max block size
Binary Code	Schrijver_$A(19,6)$	156	632	432	20
	Schrijver_$A(26,10)$	227	999	635	27
	Schrijver_$A(28,8)$	466	1746	1326	29
	Schrijver_$A(37,15)$	468	2049	1327	38
	Schrijver_$A(40,15)$	720	2900	2060	41
	Schrijver_$A(48,15)$	1728	6198	4998	49
	Schrijver_$A(50,15)$	2056	7278	5978	51
TSP	TSPbay29	6090	13862	15	29
	TSPeil51	33150	71502	26	51
Kissing Number	kissing_3_5_5 ($K(3)$)	297	220	15	56
	kissing_4_7_7 ($K(4)$)	695	488	17	120
	kissing_6_10_10 ($K(6)$)	1792	1210	20	286
QAP	QAP_Esc64a_red	517	976	8	65
	QAP_Esc16e_red	90	179	6	17

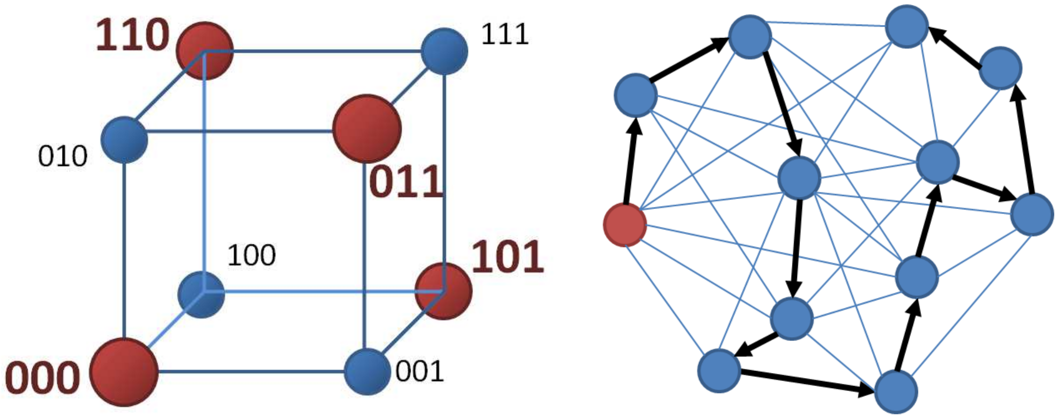
Table 4.3: *Structure of SDP problems.*

In this algorithm, we assume that the resulting Schur complement matrix $\widehat{\boldsymbol{M}}$ has full rank, so the equation (4.3.3) has a unique solution $\Delta \boldsymbol{y}$. This assumption will be dealt with below in Chapter 5.

Next we discuss some problems for which this algorithm is appropriate and results of some numerical experiments.

4.4 Problems and Experiments

In this section, we demonstrate how well the constraint-reduced version of the algorithm in Table 4.2 solves block diagonal semidefinite programming problems.



(a) Binary code when $n = 3$ and $d = 2$. (b) Example of TSP.

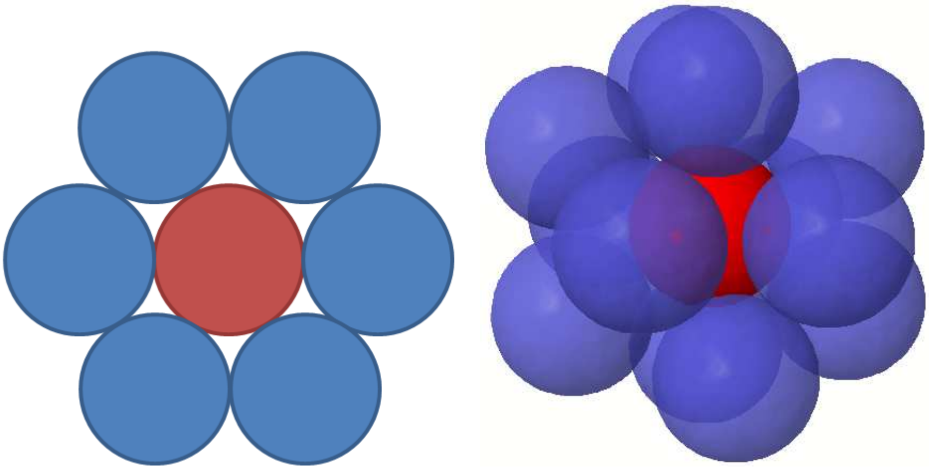
Figure 4.1: *Example of Binary code and TSP.*

4.4.1 Applications

We introduce problems to which the constraint-reduced SDP algorithm can be applied. All of these problems have diagonal block structures, and we summarize their structures in Table 4.3. All of these examples result from relaxing a problem with integer variables to one involving continuous variables.

Maximum Size of Binary Code

For a given word length n, we want to know the maximum number $A(n, d)$ of words in a binary code with Hamming distance at least d between each pair of words. For $n = 3$ and $d = 2$, $A(3, 2) = 4$ achieved by the binary code $\{(0, 0, 0), (1, 1, 0), (0, 1, 1)(0, 1, 0)\}$ (See Figure 4.1(a)). In 1979, Schrijver [77] relaxed the maximum binary code problem to SDP.



(a) Case of $n = 2$. (b) Case of $n = 3$.

Figure 4.2: *Example of kissing numbers.*

Traveling Salesman's Problem

The traveling salesman's problem (TSP) is a very well-known NP-complete problem. We are given a weighted graph $G(V, E)$ which has a set of vertices V and a set of edges E with pairwise weights (distances) w_{ij}. For a given starting point v_1, the TSP finds a path visiting all vertices in V with minimum sum of distances. (See Figure 4.1(b)). In 2008, de Klerk, Pasechnik, and Sotirov [18] relaxed TSP to SDP.

Kissing Number

The kissing number $K(n)$ is the maximum number of identical hyperspheres in n dimensions which touch a hypersphere of the same radius with no intersection. It is obvious that $K(1) = 2$ since two identical balls can be placed on the left and right side of a given ball. In the two dimensional case, a circle can be surrounded by 6 identical circles, so

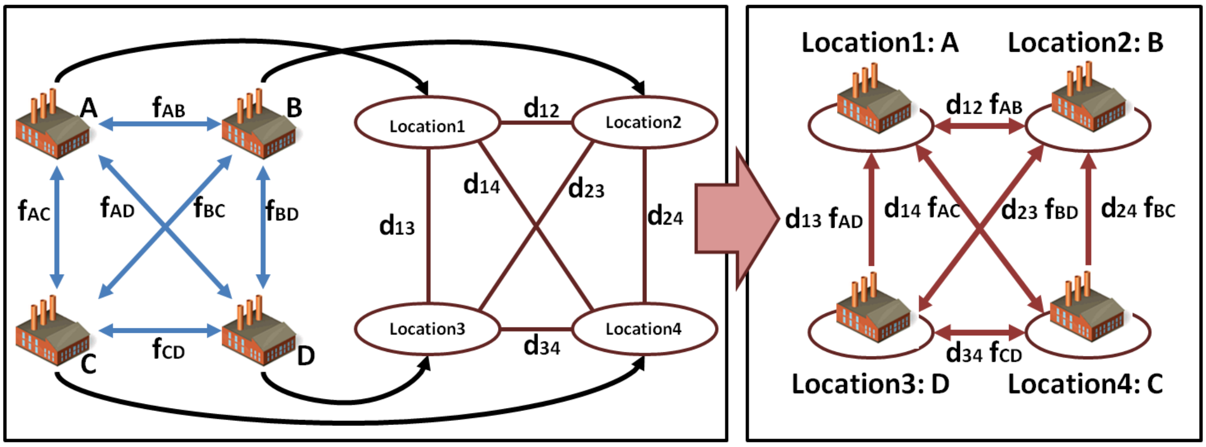


Figure 4.3: *Example of QAP when* $n = 4$.

$K(2) = 6$. (See Figure 4.2(a)). Newton believed that $K(3) = 12$, but it was first proved in 1874 by Bender [4] (See Figure 4.2(b)[4]). In 2007, Bachoc and Vallentin [2] relaxed the kissing problem to SDP.

Quadratic Assignment Problem

Suppose that, for given n facilities and n locations, we know pairwise flows $f(i, j)$ between facilities and pairwise distances $d(i, j)$ between locations. We want to assign each facility to one of the available locations in order to minimize the total flow load, defined to be the sum of flows times distances. Let g be a one-to-one correspondence function which specifies the location for each facility. If $g(1) = 2$, then the first facility is assigned to the second location. Using this assignment function $g(i)$, we can express the total flow

[4] This image is obtained from `http://en.wikipedia.org/wiki/Kissing_number_problem`

load L as

$$L(g) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} f(i,j) d(g(i), g(j)).$$

Thus, the quadratic assignment problem (QAP) determines an assignment function g minimizing $L(g)$. Figure 4.3 is an example with $n = 4$. In 1998, Zhao et al. [96] relaxed the QAP to SDP.

4.4.2 Implementation

We performed the experiments using a modified version of SDPT3 version 4.0[5] implemented by Toh, Todd, and Tütüncü [85]. Before starting an iteration, SDPT3 detects dependent rows in $\mathcal{A}$ to be removed. The iteration starts with an infeasible point on an exact central path by setting $\boldsymbol{y}^0 = \mathbf{0}$ and $\boldsymbol{X}_j^0 = \rho_x \boldsymbol{I}$ and $\boldsymbol{Z}_j^0 = \rho_z \boldsymbol{I}$ where

$$\begin{aligned}
\rho_x &= \max_{j=1,\dots,p} \left(1, \sqrt{n_j}, \max_{i=1,\dots,m} \left(\frac{1 + |b_i|}{1 + \|\boldsymbol{A}_{ij}\|_F} \right) \right), \\
\rho_z &= \max_{j=1,\dots,p} \left(1, \sqrt{n_j}, \max_{i=1,\dots,m} \left(1 + \|\boldsymbol{A}_{ij}\|_F \right), \max_{i=1,\dots,m} \left(1 + \|\boldsymbol{C}_{ij}\|_F \right) \right).
\end{aligned}$$

We modified SDPT3, as described in Section 4.3.3, to ignore the terms $\mathcal{A}_j(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})\mathcal{A}_j^T$ in the Schur complement matrix $\widehat{\boldsymbol{M}}$ for a HKM direction when $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\| < \kappa$ for a given κ. Then, the direction $\Delta\boldsymbol{y}$ is computed by solving (4.3.3), and the directions $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Z}$ are computed by (4.3.4) and (4.3.5). We vary the threshold κ from 0 to 10^7 so that we can see how constraint reduction affects the IPM. Note that constraint reduction does not occur when $\kappa = 0$.

[5] The MATLAB package is available in `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.

The original SDPT3 uses the *SYMQMR* (Symmetric Quasi-Minimal Residual algorithm) to solve the linear equation (4.2.23), using the Cholesky factor of $\boldsymbol{M}$ as a preconditioner. *SYMQMR* minimizes a quasi-residual norm from Lanczos biorthogonalization. We replaced *SYMQMR* with *SYMMLQ* (Symmetric LQ) [63] , which used fewer iterations to solve the linear systems.

We performed the experiment with the following SDP problems:[6]

1. Binary code problem: Schrijver_A(19,6), Schrijver_A(26,10), Schrijver_A(28,8), Schrijver_A(37,15), Schrijver_A(40,15),

2. Kissing number problem: kissing_3_5_5, kissing_4_7_7,

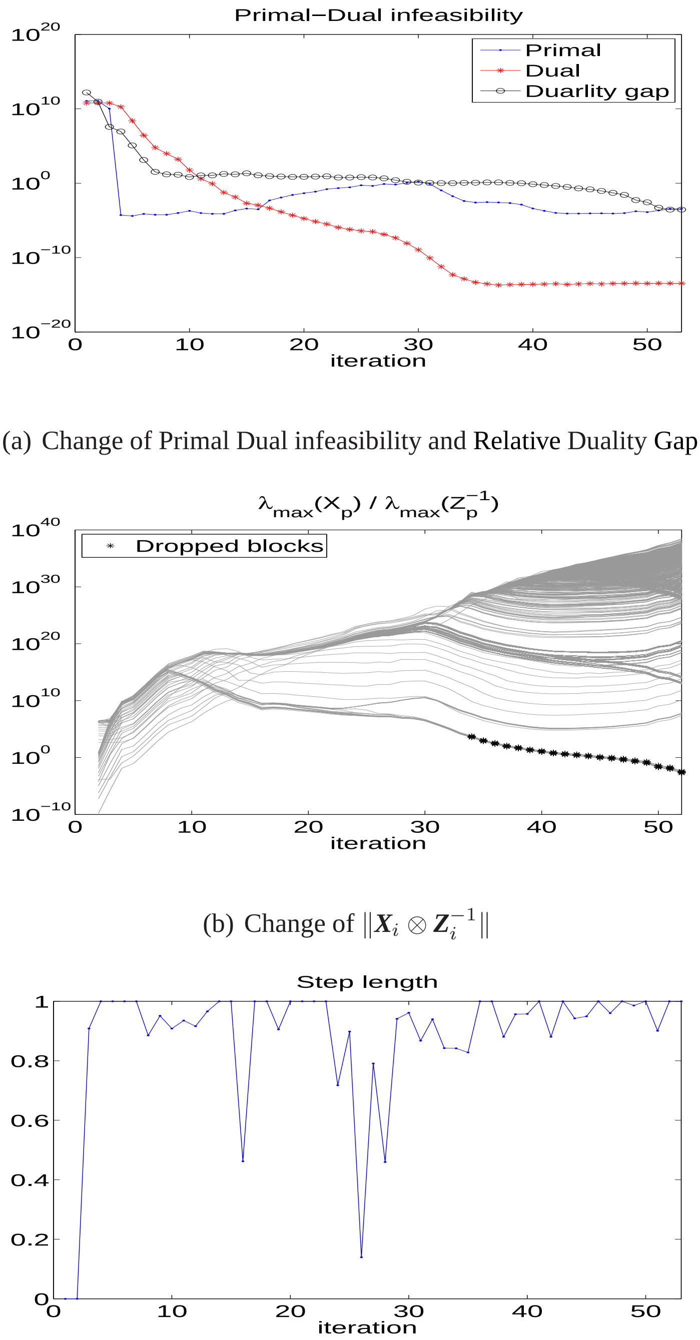
3. Quadratic assignment problem: QAP_Esc16e_red.

4.4.3 Results of Experiments

In Table 4.4, all the results of experiments are summarized. In addition, in Figure 4.4, Figure 4.5, and Figure 4.6, we trace the change in infeasibility and duality gaps, the change in $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\|$ for each block, and the change of step length, for Schrijver_A(40,15) when $\kappa = 10^4$, 10^6, and 10^7 .

We can observe that primal infeasibility, the dual infeasibility, and the duality gap gradually increase as the threshold κ increases. As expected, the computation saved by constraint reduction also tends to increase as the threshold increases.

[6] The data files are obtained from the webpage `http://lyrawww.uvt.nl/~sotirovr/library/` of E. de Klerk and R. Sotirov .

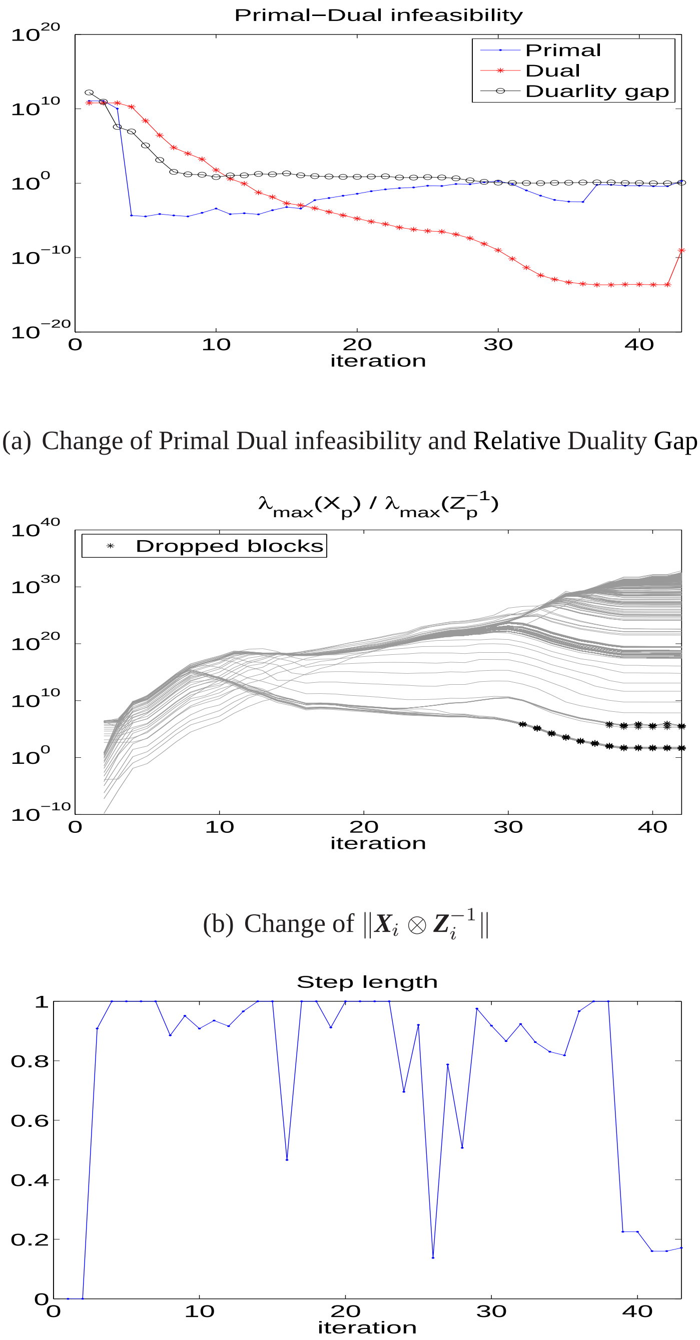


(a) Change of Primal Dual infeasibility and Relative Duality Gap

(b) Change of $\|\boldsymbol{X}_i \otimes \boldsymbol{Z}_i^{-1}\|$

(c) Change of step length θ

Figure 4.4: *Convergence measures, dropping criteria, and step lengths for Schrijver_A(40,15) when* $\kappa = 10^4$.

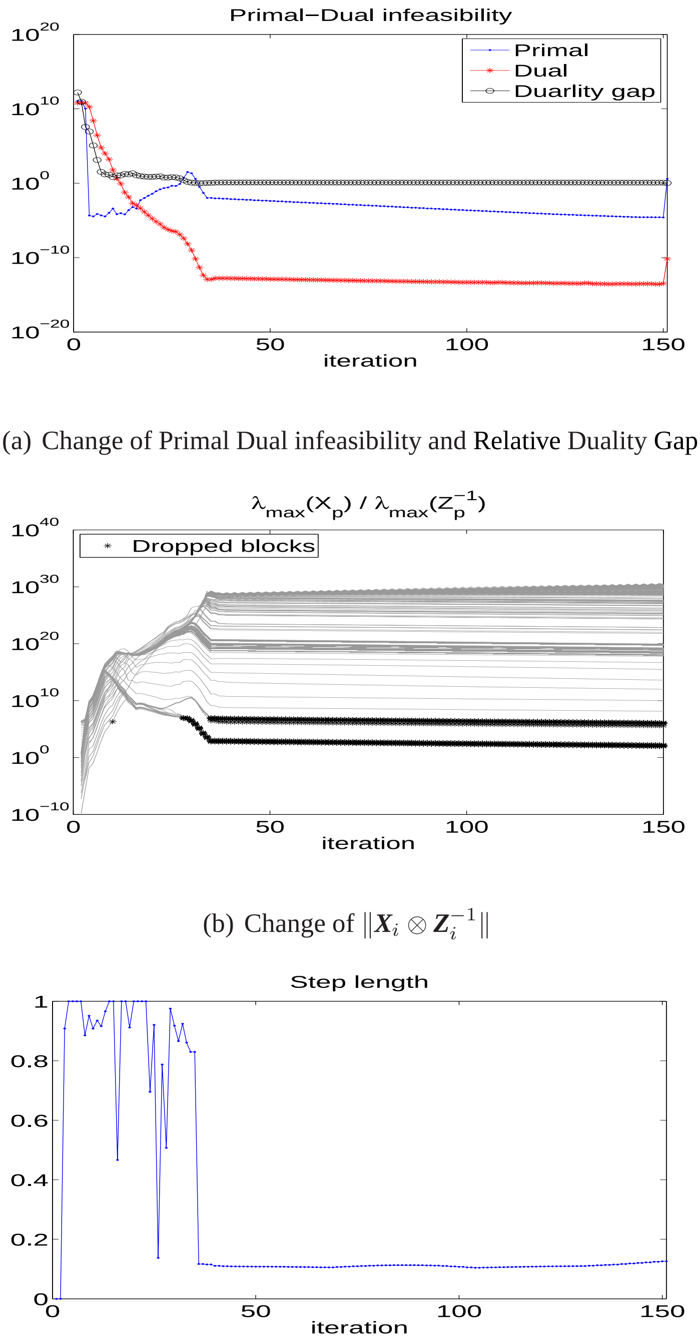


(a) Change of Primal Dual infeasibility and Relative Duality Gap

(b) Change of $\|\boldsymbol{X}_i \otimes \boldsymbol{Z}_i^{-1}\|$

(c) Change of step length θ

Figure 4.5: *Convergence measures, dropping criteria, and step lengths for Schrijver_A(40,15) when* $\kappa = 10^6$.



(a) Change of Primal Dual infeasibility and Relative Duality Gap

(b) Change of $\|\boldsymbol{X}_i \otimes \boldsymbol{Z}_i^{-1}\|$

(c) Change of step length θ

Figure 4.6: *Convergence measures, dropping criteria, and step lengths for Schrijver_A(40,15) when* $\kappa = 10^7$.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
problem	κ	primal	dual	relative	# of	# of red.	saved
		residual	residual	duality gap	iter.	blks/iter.	FLOP's/iter.
Schrijver _A(19,6)	*0.0	5.31×10^{-10}	3.32×10^{-13}	5.40×10^{-9}	29	0	0 (0.0%)
	$*1.0 \times 10^{2}$	9.98×10^{-7}	4.95×10^{-10}	4.73×10^{-6}	29	98.00	6711 (6.6%)
	1.0×10^{3}	2.10×10^{-4}	4.63×10^{-8}	2.70×10^{-4}	24	63.00	4732 (4.7%)
	1.0×10^{4}	1.05×10^{-11}	1.38×10^{-10}	1.69×10^{-1}	29	5.50	25916 (25.6%)
Schrijver _A(26,10)	*0.0	2.89×10^{-7}	1.45×10^{-14}	7.07×10^{-8}	52	0	0 (0.0%)
	$*1.0 \times 10^{2}$	3.35×10^{-8}	1.19×10^{-14}	4.13×10^{-7}	51	9.71	2620 (8.3%)
	$*1.0 \times 10^{3}$	2.60×10^{-8}	1.57×10^{-14}	2.55×10^{-7}	51	11.50	26778 (8.5%)
	1.0×10^{4}	4.85×10^{-7}	1.54×10^{-8}	1.74×10^{-2}	30	9.94	35342 (11.2%)
Schrijver _A(28,8)	*0.0	1.12×10^{-7}	3.99×10^{-13}	9.67×10^{-9}	34	0	0 (0.0%)
	$*1.0 \times 10^{2}$	8.39×10^{-8}	3.67×10^{-13}	7.52×10^{-9}	34	11.88	96 (0.0%)
	$*1.0 \times 10^{3}$	8.36×10^{-8}	3.46×10^{-13}	3.79×10^{-8}	34	22.16	195 (0.0%)
	$*1.0 \times 10^{4}$	8.45×10^{-8}	6.65×10^{-13}	2.48×10^{-6}	34	49.15	3237 (0.8%)
Schrijver _A(37,15)	*0.0	1.78×10^{-6}	9.35×10^{-15}	2.41×10^{-7}	57	0	0 (0.0%)
	$*1.0 \times 10^{0}$	2.28×10^{-6}	9.48×10^{-15}	2.41×10^{-7}	57	1.00	3.00 (0.0%)
	1.0×10^{1}	5.90×10^{-6}	2.04×10^{-10}	3.00×10^{-2}	57	7.70	57029 (4.8%)
	1.0×10^{2}	1.44×10^{-4}	1.77×10^{-9}	1.60×10^{-1}	57	10.80	91999 (7.8%)
Schrijver _A(40,15)	*0.0	2.18×10^{-4}	3.25×10^{-14}	1.53×10^{-4}	53	0	0 (0.0%)
	$* 1.0 \times 10^{4}$	3.73×10^{-4}	3.53×10^{-14}	2.80×10^{-4}	53	5.66	77940.38(4.9%)
	$*1.0 \times 10^{5}$	1.72×10^{-4}	3.41×10^{-14}	1.81×10^{-4}	53	15.29	209428 (13.2%)
	1.0×10^{6}	1.98×10^{0}	9.94×10^{-10}	1.20×10^{0}	43	13.23	296751 (18.7%)
	1.0×10^{7}	3.63×10^{-3}	1.22×10^{-13}	1.23×10^{0}	53	12.27	350146 (22.1%)
kissing _3_5_5	*0.0	4.20×10^{-11}	3.22×10^{-12}	3.24×10^{-9}	22	0	0 (0.0%)
	$*1.0 \times 10^{0}$	3.25×10^{-11}	1.39×10^{-10}	2.75×10^{-9}	22	3.17	11 (0.0%)
	$*1.0 \times 10^{1}$	5.23×10^{-6}	1.42×10^{-7}	1.23×10^{-6}	22	3.58	21 (0.0%)
	1.0×10^{2}	2.52×10^{-5}	1.19×10^{-4}	2.75×10^{-1}	16	6.67	84 (0.0%)
kissing _4_7_7	*0.0	8.41×10^{-9}	1.63×10^{-10}	4.17×10^{-8}	27	0	0 (0.0%)
	$*1.0 \times 10^{1}$	2.43×10^{-8}	5.41×10^{-11}	1.59×10^{-8}	27	6.94	46 (0.0%)
	1.0×10^{2}	8.86×10^{0}	3.37×10^{-8}	4.14×10^{-2}	27	8.59	99 (0.0%)
	1.0×10^{3}	7.97×10^{0}	3.05×10^{-7}	8.34×10^{-2}	27	10.82	264 (0.0%)
QAP _Esc16e_red	*0.0	4.03×10^{-9}	2.92×10^{-9}	5.14×10^{-8}	18	0	0 (0.0%)
	$*1.0 \times 10^{0}$	2.98×10^{-8}	1.05×10^{-8}	8.00×10^{-7}	18	31.80	95 (0.2%)
	$*1.0 \times 10^{1}$	1.37×10^{-7}	2.85×10^{-9}	5.46×10^{-7}	18	42.86	129 (0.3%)
	$*1.0 \times 10^{2}$	1.70×10^{-6}	3.61×10^{-8}	7.80×10^{-6}	18	45.12	135 (0.3%)

Table 4.4: *Result of constraint reduction. Starred entries (*) correspond to convergent iterations. Columns (6) and (7) display the number of reduced blocks and saved operations per iteration, averaged over iterations where constraint reduction is applied.*

With an excessively large κ, iterations fail to converge. For example, in Figure 4.6(a), we can see that the infeasibility and the duality gap do not decrease when too many constraints blocks are reduced. In Figure 4.6(c), the step length θ becomes very short after the 40-th iteration. This is because the directions $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ try to move away from the semidefinite cone since the corresponding constraint blocks are not included in the Schur complement matrix even though the current point is very close to its boundary.

On the other hand, with a moderate value of κ, the IPM converges with the same number of iterations as the case of no constraint reduction. In addition, infeasibility and duality gap are not so much sacrificed. (For instance, see the cases of Schrijver_A(26,10) when $\kappa = 10^2$, Schrijver_A(37,15) when $\kappa = 10^2$, Schrijver_A(40,15) when $\kappa = 10^5$, kissing_3_5_5 when $\kappa = 10^0$, and kissing_4_7_7 when $\kappa = 10^0$.) In Figure 4.4(b), which is the case of successful constraint reduction, the inactive constraint blocks start to be dropped only after the active blocks and the inactive blocks are clearly distinguishable. These results imply that we need to find an appropriate threshold κ by which the active and inactive blocks are classified correctly.

In this experiment, we kept the threshold κ static during the algorithm. Figure 4.5(b) indicates that this static threshold may cause incorrect classification. In this example, the threshold $\kappa = 10^6$ was a correct criterion at the 30-th iteration, but it turns out to be too high around the 40-th iteration. This implies that the threshold κ should be adjusted adaptively considering current values of $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\|$.

Constraint reduction shows its merit for problems in which inactive constraint blocks of moderate sizes occur such as Schrijver_A(26,10) and Schrijver_A(40,15). In particular, Schrijver_A(26,10) contains 9 inactive constraint blocks whose sizes $\widetilde{n}_j$ are 1, 3, 5, 7, 9, 11, 13, 15, and 17. Schrijver_A(40,15) contains 15 inactive constraint blocks whose sizes $\widetilde{n}_j$ are 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, and 29. We could save 8.3% of the computational cost for the Schur complement matrix when $\kappa = 10^2$ in case of A(26,10) and 13.2% (when $\kappa = 10^5$) in case of A(40,15).

In contrast, the effect of constraint reduction are not visible in Schrijver_A(28,8), kissing_3_5_5 and kissing_3_5_5. This is because those problems contain either few or no inactive constraint blocks. Schrijver_A(28,8) contains 77 inactive constraints of size $\widetilde{n}_j = 1$ and only one inactive constraint block of size $\widetilde{n}_j = 3$. kissing_3_5_5 contains only 4 inactive constraints of size $\widetilde{n}_j = 1$. However, our constraint reduction is effective for SDP problems that have a large number of large inactive dual constraints.

4.5 Conclusion

In this chapter, we showed how we can apply constraint reduction to block diagonal SDP using a *predictor-corrector* method.

In addition, we demonstrated how varying the threshold κ influences the iterations of the interior point method. From the experiments, we make three important observations.

1. For successful constraint reduction, the threshold κ must be able to distinguish the inactive constraint blocks from the active blocks.

2. The threshold κ needs to be adaptively adjusted because $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\|$ changes dynamically for each iteration.

3. Constraint reduction becomes effective when the SDP has a large number of inactive constraint blocks.

In the next chapter, we will resolve the issues arising from the first two observations by presenting adaptive criteria for constraint reduction and verifying validity by proving global convergence.

Chapter 5

Constraint-Reduced Predictor-Corrector Algorithm for Semidefinite Programming with Polynomial Complexity

The previous chapter introduced how constraint reduction can be applied to the *predictor-corrector* method for SDP. The experiments with test problems raised a few issues about the criteria to adaptively reduce constraint blocks.

In this chapter, we propose a new infeasible *predictor-corrector* algorithm with adaptive criteria for constraint reduction. We verify its validity by proving global convergence. We also prove its polynomial complexity, $O(n \ln(\epsilon_0/\epsilon))$, for a given convergence tolerance ϵ and an initial residual ϵ_0.

The algorithm is a modification of one with no constraint reduction, due to Potra and Sheng [71], and can be applied when the data matrices are block diagonal. The constraint reduction generates an extra term $\Delta \boldsymbol{X}_\epsilon$ in the primal direction which is not reflected in updating $\boldsymbol{X}$, but perturbs the complementarity equation. Due to this new $\Delta \boldsymbol{X}_\epsilon$, a series of lemmas for global convergence by Potra and Sheng [71] need to be modified. The proposed adaptive criteria restrain the magnitude of $\Delta \boldsymbol{X}_\epsilon$ so that we can guarantee the step length θ is long enough for iterations to converge.

5.1 Constraint-Reduced Predictor-Corrector Method for SDP

We use the notation defined in Chapter 4 with minor changes. We say that a point $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$ is feasible if it satisfies the primal and dual constraints in (4.1.1) and (4.1.2). Throughout this chapter, we assume the following.

Assumption 5.1 (Slater condition)**.** *There exists a primal and dual feasible point* $(\mathbf{X}, \mathbf{y}, \mathbf{Z})$ *such that* $\mathbf{X} \succ \mathbf{0}$ *and* $\mathbf{Z} \succ \mathbf{0}$.

Under Assumption 5.1 the primal and dual SDP problems have optimal solutions with equal optimal values[1].

5.1.1 HKM Direction for Symmetrization

[1] See, for example, de Klerk [17, Theorem 2.6 in p.33]

In this section, we briefly review the equations introduced in Chapter 4 and introduce new equations having a symmetric solution $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Z}$ with no extra symmetrization step. The equations introduced in this section are very useful when we prove the global convergence of a new *predictor-corrector* algorithm in Section 5.2.

Under Assumption 5.1, $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$ is an optimal solution if and only if

$$\boldsymbol{A}_i \bullet \boldsymbol{X} = \boldsymbol{b} \quad \text{for } i = 1, \ldots, m, \tag{5.1.1}$$

$$(\sum_{i=1}^{m} y_i \boldsymbol{A}_i) + \boldsymbol{Z} = \boldsymbol{C}, \tag{5.1.2}$$

$$\boldsymbol{X} \bullet \boldsymbol{Z} = 0, \tag{5.1.3}$$

$$\boldsymbol{X} \succeq 0, \ \ \boldsymbol{Z} \succeq 0. \tag{5.1.4}$$

So, we solve the following Newton equations to find a direction toward the optimal solution.

$$\boldsymbol{A}_i \bullet \Delta\boldsymbol{X} = r_{pi} \quad \text{for } i = 1, \ldots, m, \tag{5.1.5}$$

$$(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i) + \Delta\boldsymbol{Z} = \boldsymbol{R}_d, \tag{5.1.6}$$

$$\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z} = \boldsymbol{R}_c, \tag{5.1.7}$$

where the primal residual, dual residual, and complementarity residuals are defined by

$$\begin{aligned} r_{pi} &= b_i - \boldsymbol{A}_i \bullet \boldsymbol{X} \quad \text{for } i = 1, \ldots, m, && (5.1.8) \\ \boldsymbol{R}_d &= \boldsymbol{C} - \boldsymbol{Z} - \sum_{i=1}^{m} y_i \boldsymbol{A}_i, && (5.1.9) \\ \boldsymbol{R}_c &= \overline{\mu}\boldsymbol{I} - \boldsymbol{X}\boldsymbol{Z}, && (5.1.10) \end{aligned}$$

where $\overline{\mu}$ defines the current target point on the central path. In SDP, IPMs require symmetrization since $\Delta\boldsymbol{X}$ from (5.1.7) is not necessarily symmetric (See Section 4.2.1). Thus, we effectively solve the equations

$$\boldsymbol{A}_i \bullet \Delta\dot{\boldsymbol{X}} = r_{pi} \quad \text{for } i = 1, \ldots, m, \tag{5.1.11}$$

$$(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i) + \Delta\boldsymbol{Z} = \boldsymbol{R}_d, \tag{5.1.12}$$

$$\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\dot{\boldsymbol{X}}\boldsymbol{Z} = \boldsymbol{R}_c, \tag{5.1.13}$$

$$\Delta\dot{\boldsymbol{X}} = \Delta\boldsymbol{X} + \boldsymbol{W}, \tag{5.1.14}$$

where $\Delta\boldsymbol{X} \in \mathcal{S}^n$ and $\boldsymbol{W} \in \widetilde{\mathcal{S}^n}$. By (5.1.14), $\Delta\boldsymbol{X}$ is the symmetric part of $\Delta\dot{\boldsymbol{X}}$,

$$\Delta\boldsymbol{X} = \text{ symm}\left(\Delta\dot{\boldsymbol{X}}\right).$$

The direction $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ is called the HKM direction; named after Helmberg, Kojima, and Monteiro [38, 51, 61]. Since $\boldsymbol{A}_i \in \mathcal{S}^n$ and $\boldsymbol{W} \in \widetilde{\mathcal{S}^n}$, $\boldsymbol{A}_i \bullet \boldsymbol{W} = 0$. By this property, the symmetrized direction $\Delta\boldsymbol{X}$ from (5.1.11)-(5.1.14) also satisfies (5.1.5), so the primal residual is the same with or without symmetrization.

For a fixed weighting parameter $d \in [0, 1]$, Kojima, Shindoh, and Hara [51, Theorem 4.2 on p.100] showed that the equations (5.1.11) and (5.1.12) with

$$\boldsymbol{X}(\Delta\boldsymbol{Z} + d\boldsymbol{W}) + (\Delta\boldsymbol{X} + (1-d)\boldsymbol{W})\boldsymbol{Z} = \boldsymbol{R}_c, \tag{5.1.15}$$

have a unique solution $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z}, \boldsymbol{W}) \in \mathcal{S}^n \times \mathbb{R}^m \times \mathcal{S}^n \times \widetilde{\mathcal{S}^n}$. From this point of view, (5.1.13) with (5.1.14) is the case of $d = 0$ in (5.1.15), and the equations (5.1.11)-(5.1.14) have a unique solution. Later, Monteiro [61] showed that we can obtain the same direction without the extra symmetrization step by solving

$$\boldsymbol{A}_i \bullet \Delta \boldsymbol{X} = r_{pi} \quad \text{for } i = 1, \ldots, m, \tag{5.1.16}$$

$$(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i) + \Delta \boldsymbol{Z} = \boldsymbol{R}_d, \tag{5.1.17}$$

$$\text{symm}\left(\boldsymbol{Z}^{1/2}(\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z})\boldsymbol{Z}^{-1/2}\right) = \overline{\mu}\boldsymbol{I} - \boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}. \tag{5.1.18}$$

Specifically, Monteiro [61, Lemma 2.1 and following discussion] proved that the solution of (5.1.11)-(5.1.14) is the unique solution of (5.1.16)-(5.1.18). So, we will frequently refer to (5.1.18) for convergence analysis later.

5.1.2 Constraint-Reduced Linear System

As discussed in Section 4.2.2, the equations (5.1.11)-(5.1.14) can be reduced to

$$\boldsymbol{M}\Delta\boldsymbol{y} = \boldsymbol{g},$$

where $\boldsymbol{M} = \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T$ and $\boldsymbol{g} = \boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c$.

In Section 4.2.2, we discussed how we can apply constraint reduction to the linear equation, so we have the constraint-reduced equation,

$$\widehat{\boldsymbol{M}}\Delta\boldsymbol{y} = \boldsymbol{g}, \tag{5.1.19}$$

by replacing $\boldsymbol{M}$ with $\widehat{\boldsymbol{M}}$. So, we solve

$$\left(\widehat{\mathcal{A}}(\widehat{\boldsymbol{X}} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T\right)\Delta\boldsymbol{y} = \boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c. \tag{5.1.20}$$

For uniqueness of the solution $\Delta\boldsymbol{y}$ of (5.1.20), we assume independent rows of $\widehat{\mathcal{A}}$ as follows.

Assumption 5.2. *For any re-ordering of the blocks of the $\widehat{\mathbf{A}}_i$'s and any $\widehat{p}$ such that $\sum_{j=1}^{\widehat{p}} \hat{n}_j^{\overline{2}} \geq m$, the matrices $\widehat{\mathbf{A}}_i$, $i = 1, \ldots, m$ are linearly independent.*

If $\widehat{\boldsymbol{X}} \succ \mathbf{0}$, $\widehat{\boldsymbol{Z}} \succ \mathbf{0}$, and $\sum_{j=1}^{\widehat{p}} \hat{n}_j^{\overline{2}} \geq m$ where $x^{\overline{2}} = x(x+1)/2$, the reduced Schur complement matrix $\widehat{\boldsymbol{M}}$ has full rank by Assumption 5.2, so the equations (5.1.19) and (5.1.20) have a unique solution $\Delta\boldsymbol{y}$.

So far, we follow the equations in Chapter 4. However, in contrast to Section 4.2.2, we now compute $\Delta\dot{\boldsymbol{x}}$ and $\Delta\boldsymbol{z}$ by

$$\Delta\dot{\boldsymbol{x}} = \operatorname{vec}\left(\begin{bmatrix} \Delta\dot{\widehat{\boldsymbol{X}}} & \mathbf{0} \\ \mathbf{0} & \Delta\dot{\widetilde{\boldsymbol{X}}} \end{bmatrix}\right), \tag{5.1.21}$$

$$\Delta\boldsymbol{z} = \boldsymbol{r}_d - \mathcal{A}^T\Delta\boldsymbol{y}, \tag{5.1.22}$$

where

$$\Delta\dot{\widehat{\boldsymbol{X}}} = \operatorname{mat}\left((\widehat{\boldsymbol{X}} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T\Delta\boldsymbol{y} - (\widehat{\boldsymbol{X}} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\boldsymbol{r}}_d + (\boldsymbol{I} \otimes \widehat{\boldsymbol{Z}}^{-1})\widehat{\boldsymbol{r}}_c\right), \tag{5.1.23}$$

$$\Delta\dot{\widetilde{\boldsymbol{X}}} = \operatorname{mat}\left(-(\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\boldsymbol{r}}_d + (\boldsymbol{I} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\boldsymbol{r}}_c\right). \tag{5.1.24}$$

The residuals $(\widehat{\boldsymbol{r}}_d, \widetilde{\boldsymbol{r}}_d)$ and $(\widehat{\boldsymbol{r}}_c, \widetilde{\boldsymbol{r}}_c)$ are vectorizations of $(\widehat{\boldsymbol{R}}_d, \widetilde{\boldsymbol{R}}_d)$ and $(\widehat{\boldsymbol{R}}_c, \widetilde{\boldsymbol{R}}_c)$ defined in Section 4.3.2. Note that while (4.2.22) contains $(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y}$ as its first term, $\Delta\dot{\widetilde{\boldsymbol{X}}}$ in (5.1.24) does not have the corresponding term $(\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\boldsymbol{y}$, which will cause a perturbation $\Delta\dot{\boldsymbol{X}}_\epsilon$ in the primal direction as we derive next.

In the constraint-reduced linear system, we replaced the Schur complement matrix $\boldsymbol{M}$ with $\widehat{\boldsymbol{M}}$. How does this influence the solution? In the following lemma, we show that $\Delta\dot{\boldsymbol{x}}$, $\Delta\boldsymbol{z}$, and $\Delta\boldsymbol{y}$ from equations (5.1.19), (5.1.21) , and (5.1.22), are a solution of the following perturbed equations

$$\mathcal{A}\Delta\dot{\boldsymbol{x}} = \boldsymbol{r}_p, \tag{5.1.25}$$

$$\mathcal{A}^T\Delta\boldsymbol{y} + \Delta\boldsymbol{z} = \boldsymbol{r}_d, \tag{5.1.26}$$

$$(\boldsymbol{X}\otimes\boldsymbol{I})\Delta\boldsymbol{z} + (\boldsymbol{I}\otimes\boldsymbol{Z})(\Delta\dot{\boldsymbol{x}} + \Delta\dot{\boldsymbol{x}}_\epsilon) = \boldsymbol{r}_c, \tag{5.1.27}$$

where

$$\Delta\dot{\boldsymbol{X}}_\epsilon = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \operatorname{mat}\left((\widetilde{\boldsymbol{X}}\otimes\widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\boldsymbol{y}\right) \end{bmatrix}. \tag{5.1.28}$$

Note the new vector $\Delta\dot{\boldsymbol{x}}_\epsilon$ in the second term of (5.1.27).

Lemma 5.1.1 (Perturbed Newton equations)**.** *The solution* $(\Delta\dot{\mathbf{x}}, \Delta\mathbf{y}, \Delta\mathbf{z})$ *of (5.1.19), (5.1.21), and (5.1.22) satisfies equations (5.1.25)-(5.1.27).*

Proof. First, we show the primal equation (5.1.25) is satisfied. By (5.1.21),

$$\begin{aligned}
\mathcal{A}\Delta\dot{\boldsymbol{x}} &= \widehat{\mathcal{A}}\Delta\dot{\widehat{\boldsymbol{x}}} + \widetilde{\mathcal{A}}\Delta\dot{\widetilde{\boldsymbol{x}}} \\
&= \widehat{\mathcal{A}}(\widehat{\boldsymbol{X}}\otimes\widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T\Delta\boldsymbol{y} - \widehat{\mathcal{A}}(\widehat{\boldsymbol{X}}\otimes\widehat{\boldsymbol{Z}}^{-1})\widehat{\boldsymbol{r}}_d + \widehat{\mathcal{A}}(\boldsymbol{I}\otimes\widehat{\boldsymbol{Z}}^{-1})\widehat{\boldsymbol{r}}_c \\
&\quad - \widetilde{\mathcal{A}}(\widetilde{\boldsymbol{X}}\otimes\widetilde{\boldsymbol{Z}}^{-1})\widetilde{\boldsymbol{r}}_d + \widetilde{\mathcal{A}}(\boldsymbol{I}\otimes\widetilde{\boldsymbol{Z}}^{-1})\widetilde{\boldsymbol{r}}_c \text{ (by (5.1.23) and (5.1.24))} \\
&= \widehat{\mathcal{A}}(\widehat{\boldsymbol{X}}\otimes\widehat{\boldsymbol{Z}}^{-1})\widehat{\mathcal{A}}^T\Delta\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_d + \mathcal{A}(\boldsymbol{I}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_c \\
&= \left(\boldsymbol{r}_p + \mathcal{A}(\boldsymbol{X}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_d - \mathcal{A}(\boldsymbol{I}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_c.\right) \\
&\quad - \mathcal{A}(\boldsymbol{X}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_d + \mathcal{A}(\boldsymbol{I}\otimes\boldsymbol{Z}^{-1})\boldsymbol{r}_c \text{ (by (5.1.20))} \\
&= \boldsymbol{r}_p,
\end{aligned}$$

so (5.1.25) is satisfied.

In addition, (5.1.26) is immediately satisfied by (5.1.22).

To see (5.1.27) is satisfied, we first calculate $(\Delta\dot{\boldsymbol{x}} + \Delta\boldsymbol{x}_\epsilon)$. By (5.1.21), (5.1.23), (5.1.24), and (5.1.28),

$$\begin{aligned}
\Delta\dot{\boldsymbol{X}} + \Delta\dot{\boldsymbol{X}}_\epsilon &= \begin{bmatrix} \Delta\dot{\hat{\boldsymbol{X}}} & \mathbf{0} \\ \mathbf{0} & \Delta\dot{\tilde{\boldsymbol{X}}} + \text{ mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\boldsymbol{y}\right) \end{bmatrix} \\
&= \text{mat}\left((\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y} - (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d + (\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c\right),
\end{aligned}$$

so

$$\Delta\dot{\boldsymbol{x}} + \Delta\dot{\boldsymbol{x}}_\epsilon = (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y} - (\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d + (\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c.$$

Thus,

$$\begin{aligned}
(\boldsymbol{I} \otimes \boldsymbol{Z})(\Delta\dot{\boldsymbol{x}} + \Delta\dot{\boldsymbol{x}}_\epsilon) &= (\boldsymbol{I} \otimes \boldsymbol{Z})(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\mathcal{A}^T\Delta\boldsymbol{y} - (\boldsymbol{I} \otimes \boldsymbol{Z})(\boldsymbol{X} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_d \\
&\quad + (\boldsymbol{I} \otimes \boldsymbol{Z})(\boldsymbol{I} \otimes \boldsymbol{Z}^{-1})\boldsymbol{r}_c \\
&= (\boldsymbol{X} \otimes \boldsymbol{I})\mathcal{A}^T\Delta\boldsymbol{y} - (\boldsymbol{X} \otimes \boldsymbol{I})\boldsymbol{r}_d + (\boldsymbol{I} \otimes \boldsymbol{I})\boldsymbol{r}_c \\
&= (\boldsymbol{X} \otimes \boldsymbol{I})(\mathcal{A}^T\Delta\boldsymbol{y} - \boldsymbol{r}_d) + \boldsymbol{r}_c \\
&= -(\boldsymbol{X} \otimes \boldsymbol{I})\Delta\boldsymbol{z} + \boldsymbol{r}_c \;\; \text{(by (5.1.22))}.
\end{aligned}$$

Therefore,

$$(\boldsymbol{X} \otimes \boldsymbol{I})\Delta\boldsymbol{z} + (\boldsymbol{I} \otimes \boldsymbol{Z})(\Delta\dot{\boldsymbol{x}} + \Delta\dot{\boldsymbol{x}}_\epsilon) = \boldsymbol{r}_c$$

□

From the equations (5.1.25)-(5.1.28) and Lemma 5.1.1, we can see that constraint reduction does not affect the primal and dual equations (5.1.5) and (5.1.6), but solely the complementarity equation (5.1.7). Furthermore, considering the relations between

(5.1.11)-(5.1.14) and (5.1.16)-(5.1.18), the solution $(\Delta\dot{\boldsymbol{x}}, \Delta\boldsymbol{y}, \Delta\boldsymbol{z})$ of (5.1.25)-(5.1.27) also satisfies the following equations by the symmetrization of $\Delta\boldsymbol{X} = \text{ symm }\big(\Delta\dot{\boldsymbol{X}}\big)$.

$$\mathcal{A}\Delta\boldsymbol{x} = \boldsymbol{r}_p, \tag{5.1.29}$$

$$\mathcal{A}^T\Delta\boldsymbol{y} + \Delta\boldsymbol{z} = \boldsymbol{r}_d, \tag{5.1.30}$$

$$\boldsymbol{Z}^{1/2}(\boldsymbol{X} + \Delta\boldsymbol{X}_\epsilon)\boldsymbol{Z}^{1/2} \ + \ \text{symm}\left(\boldsymbol{Z}^{1/2}(\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z})\boldsymbol{Z}^{-1/2}\right) = \overline{\mu}\boldsymbol{I}, \tag{5.1.31}$$

where

$$\Delta\boldsymbol{X}_\epsilon = \text{ symm }\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\boldsymbol{y}\right) \end{bmatrix}\right). \tag{5.1.32}$$

5.1.3 Algorithm

In this section, we introduce an interior point method, similar to that of Potra and Sheng [71], but including constraint reduction. It is a *predictor-corrector* algorithm, but, like Potra and Sheng's algorithm, it is somewhat unusual in that it does not reuse the predictor matrix in the corrector step.

We define a set $\mathcal{F}$ of feasible solutions and a set $\mathcal{F}^*$ of optimal solutions as

$$\begin{aligned} \mathcal{F} &= \{(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \in \mathcal{S}^n_+ \times \mathbb{R}^m \times \mathcal{S}^n_+ : \ (\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \text{ satisfies (5.1.1) and (5.1.2). }\}, \\ \mathcal{F}^* &= \{(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \in \mathcal{F} : \boldsymbol{X} \bullet \boldsymbol{Z} = 0\}. \end{aligned}$$

We also define the neighborhood $\mathcal{N}(\gamma, \tau)$ of the central path as

$$\mathcal{N}(\gamma, \tau) = \{(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{S}^n_{++} \times \mathcal{S}^n_{++} : \|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\|_F \le \gamma\tau\}.$$

In the predictor step, given the current iterate $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$ and "inactive blocks" $(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Z}})$ of

$(\boldsymbol{X}, \boldsymbol{Z})$, we find a solution $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ of (5.1.29)-(5.1.32), setting $\overline{\mu} = 0$, so

$$\mathcal{A}\Delta\boldsymbol{x} = \boldsymbol{r}_p, \tag{5.1.33}$$

$$\mathcal{A}^T\Delta\boldsymbol{y} + \Delta\boldsymbol{z} = \boldsymbol{r}_d, \tag{5.1.34}$$

$$\boldsymbol{Z}^{1/2}(\boldsymbol{X} + \Delta\boldsymbol{X}_\epsilon)\boldsymbol{Z}^{1/2} \;+\; \text{symm}\left(\boldsymbol{Z}^{1/2}(\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z})\boldsymbol{Z}^{-1/2}\right) = \boldsymbol{0}, \tag{5.1.35}$$

$$\Delta\boldsymbol{X}_\epsilon = \text{symm}\left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \text{mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\boldsymbol{y}\right) \end{bmatrix}\right). \tag{5.1.36}$$

We then compute an updated point $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{Z}})$ by taking a step of length $\overline{\theta} < 1$ in this direction.

In the corrector step, we set the target duality gap $\overline{\mu} = (1-\overline{\theta})\tau$, where the parameter τ decreases at each iteration. Then, with inactive blocks $(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Z}})$ of $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}})$, we find a solution $(\Delta\overline{\boldsymbol{X}}, \Delta\overline{\boldsymbol{y}}, \Delta\overline{\boldsymbol{Z}})$ of (5.1.29)-(5.1.32) with $\boldsymbol{r}_p = \boldsymbol{0}$ and $\boldsymbol{r}_d = \boldsymbol{0}$, so

$$\mathcal{A}\Delta\overline{\boldsymbol{x}} = \boldsymbol{0}, \tag{5.1.37}$$

$$\mathcal{A}^T\Delta\overline{\boldsymbol{y}} + \Delta\overline{\boldsymbol{z}} = \boldsymbol{0}, \tag{5.1.38}$$

$$\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}} + \Delta\overline{\boldsymbol{X}}_\epsilon)\overline{\boldsymbol{Z}}^{1/2} \;+\; \text{symm}\left(\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{X}\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{(-1/2)}\right) = (1-\overline{\theta})\tau\boldsymbol{I}, \tag{5.1.39}$$

$$\Delta\overline{\boldsymbol{X}}_\epsilon = \text{symm}\left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \text{mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T\Delta\overline{\boldsymbol{y}}\right) \end{bmatrix}\right). \tag{5.1.40}$$

We define a few variables to denote the magnitude of directions as

$$\delta := \frac{1}{\tau}\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\|_F, \tag{5.1.41}$$

$$\delta_x := \|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_F, \tag{5.1.42}$$

$$\delta_\epsilon := \frac{1}{\tau}\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}_\epsilon\boldsymbol{Z}^{1/2}\|_F, \tag{5.1.43}$$

$$\overline{\delta}_\epsilon := \frac{1}{\tau}\|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_\epsilon\overline{\boldsymbol{Z}}^{1/2}\|_F. \tag{5.1.44}$$

We use two fixed positive parameters α and β with the property

$$\frac{\beta^2}{2(1-\beta)^2} < \alpha < \beta \leq \frac{\beta}{1-\beta} < 1. \tag{5.1.45}$$

This inequality restrains the ranges of α and β as $0 < \alpha < \beta < 0.5$. For example, we can choose $(\alpha, \beta) = (0.17, 0.3)$. Based on these parameters, we define $\widehat{\theta}$ and $\breve{\theta}$ (which change at each iteration) as

$$\begin{aligned}\widehat{\theta} &= \frac{(\alpha-\beta-\delta_\epsilon) + \sqrt{(\alpha-\beta-\delta_\epsilon)^2 + 4\delta(\beta-\alpha)}}{2\delta} \\ &= \frac{2(\beta-\alpha)}{\sqrt{(\beta-\alpha+\delta_\epsilon)^2 + 4\delta(\beta-\alpha)} - (\beta-\alpha+\delta_\epsilon)},\end{aligned} \tag{5.1.46}$$

$$\breve{\theta} = \max\{\tilde{\theta} \in [0,1] : (\boldsymbol{X}+\theta\Delta\boldsymbol{X}, \boldsymbol{y}+\theta\Delta\boldsymbol{y}, \boldsymbol{Z}+\theta\Delta\boldsymbol{Z}) \in \mathcal{N}(\beta, (1-\theta)\tau),\ \forall\theta \in [0,\tilde{\theta}]\}. \tag{5.1.47}$$

The following two conditions are used in the *predictor-corrector* algorithm. The first one applies to the predictor step, and the second one applies to the corrector step.

Condition 5.1.

$$\delta_\epsilon \leq \frac{q}{\tau}\delta_x, \tag{5.1.48}$$

or equivalently

$$\|\mathbf{Z}^{1/2}\Delta\mathbf{X}_\epsilon\mathbf{Z}^{1/2}\|_F \leq q\|\mathbf{Z}^{1/2}\Delta\mathbf{X}\mathbf{Z}^{1/2}\|_F, \tag{5.1.49}$$

where the input parameter q of the algorithm has a range

$$0 \leq q < 1 - \alpha. \tag{5.1.50}$$

Condition 5.2.

$$\overline{\delta}_\epsilon < (1 - \overline{\theta})(\sqrt{s^2 + t} - s), \tag{5.1.51}$$

where

$$s = \beta^2 - \beta + 1, \;\; t = 2\alpha(1 - \beta)^2 - \beta^2. \tag{5.1.52}$$

Condition 5.1 ensures that the ratio of the perturbation term $\Delta\boldsymbol{X}_\epsilon$ to the primal direction $\Delta\boldsymbol{X}$ is bounded by the given ratio q. Condition 5.2 plays a role for the corrector step to move the iterate into $\mathcal{N}(\alpha, (1 - \overline{\theta})\tau)$, the neighborhood of the central path. Condition 5.1 and Condition 5.2 can be checked at low cost compared to the cost of solving the full (unreduced) system.

Based on these parameters and conditions, we now define our *predictor-corrector* algorithm in Table 5.1. In step 3.(d), the choice of step length in the predictor step is valid only when $\widehat{\theta} \leq \breve{\theta}$, which will be proved in Lemma 5.2.3. Since $\breve{\theta}$ is a theoretical upper bound for $\overline{\theta}$, it may be not practical to compute $\breve{\theta}$. For practical implementation, $\overline{\theta}$ can be chosen to be defined by (5.1.46). In step 3.(e), the algorithm terminates since $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{Z}})$ is an optimal solution, which will be shown in Lemma 5.2.3.

Before starting analysis, the following overview is useful.

1. Since $\boldsymbol{r}_p = 0$ and $\boldsymbol{r}_d = 0$ in the corrector step, the corrector step makes no contribution to reducing primal and dual residuals. Its only purpose is to move the point toward the central path.

1. Input : $\mathcal{A}$, $\boldsymbol{b}$, $\boldsymbol{C}$; α and β satisfying (5.1.45); convergence tolerance τ^*; ρ such that $\rho \geq \max(\|\boldsymbol{X}^*\|, \|\boldsymbol{Z}^*\|)$ for $(\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{Z}^*) \in \mathcal{F}^*$; and q, the perturbation bound for the primal direction in the predictor step, satisfying (5.1.50).

2. Set $(\boldsymbol{X}^0, \boldsymbol{y}^0, \boldsymbol{Z}^0) = (\rho \boldsymbol{I}, \boldsymbol{0}, \rho \boldsymbol{I})$. Set $\tau = \tau_0 = \mu_0 = (\boldsymbol{X}^0 \bullet \boldsymbol{Z}^0)/n = \rho^2$.

3. Repeat until $\tau < \tau^*$: For $k = 0, 1, \dots,$

   (a) Set $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) = (\boldsymbol{X}^k, \boldsymbol{y}^k, \boldsymbol{Z}^k)$ and $\tau = \tau_k$.

   (b) Sort the constraint blocks in decreasing order of $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\|$.

   (c) Initially, $\widehat{\boldsymbol{M}}_p = \boldsymbol{0}$. For $j = 1, \dots, p$, until $\sum_{l=1}^{j} \hat{n}_l^{\overline{2}} \geq m$ and Condition 5.1 (above) is satisfied, replace $\widehat{\boldsymbol{M}}_p$ by $\widehat{\boldsymbol{M}}_p + \mathcal{A}_j(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})\mathcal{A}_j^T$. Set $\widehat{p} = j$.

   (d) By solving (5.1.20) with $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}_p$ and $\boldsymbol{r}_c = \operatorname{vec}(-\boldsymbol{X}\boldsymbol{Z})$ find $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$ satisfying (5.1.33) - (5.1.36). Choose a step length $\overline{\theta} \in [\widehat{\theta}, \breve{\theta}]$ defined by (5.1.46) and (5.1.47),
   $\overline{\boldsymbol{X}} = \boldsymbol{X} + \overline{\theta}\Delta\boldsymbol{X}, \;\; \overline{\boldsymbol{y}} = \boldsymbol{y} + \overline{\theta}\Delta\boldsymbol{y}, \;\; \overline{\boldsymbol{Z}} = \boldsymbol{Z} + \overline{\theta}\Delta\boldsymbol{Z}.$

   (e) If $\overline{\theta} = 1$, terminate the iteration with optimal solution $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{Z}})$.

   (f) Sort the constraint blocks in decreasing order of $\|\overline{\boldsymbol{X}}_j \otimes \overline{\boldsymbol{Z}}_j^{-1}\|$.

   (g) Initially, $\widehat{\boldsymbol{M}}_c = \boldsymbol{0}$. For $j = 1, \dots, p$, until $\sum_{l=1}^{j} \hat{n}_l^{\overline{2}} \geq m$ and Condition 5.2 (above) is satisfied, replace $\widehat{\boldsymbol{M}}_c$ by $\widehat{\boldsymbol{M}}_c + \mathcal{A}_j(\overline{\boldsymbol{X}}_j \otimes \overline{\boldsymbol{Z}}_j^{-1})\mathcal{A}_j^T$. Set $\widehat{p} = j$.

   (h) By solving (5.1.20) with $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}_c$, $\boldsymbol{r}_p = \boldsymbol{0}$, $\boldsymbol{r}_d = \boldsymbol{0}$, and $\boldsymbol{r}_c = \operatorname{vec}\big((1-\overline{\theta})\tau\boldsymbol{I} - \overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}\big)$, find $(\Delta\overline{\boldsymbol{X}}, \Delta\overline{\boldsymbol{y}}, \Delta\overline{\boldsymbol{Z}})$ satisfying (5.1.37) - (5.1.40).
   Take a full step as
   $\boldsymbol{X}^{(k+1)} = \boldsymbol{X}^+ = \overline{\boldsymbol{X}} + \Delta\overline{\boldsymbol{X}}, \;\; \boldsymbol{y}^{(k+1)} = \boldsymbol{y}^+ = \overline{\boldsymbol{y}} + \Delta\overline{\boldsymbol{y}}, \;\; \boldsymbol{Z}^{(k+1)} = \boldsymbol{Z}^+ = \overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{Z}}.$

   (i) Set $\tau_{k+1} = (1-\overline{\theta})\tau$.

   (j) Update $\boldsymbol{r}_p = \boldsymbol{b} - \mathcal{A}\boldsymbol{x}$ and $\boldsymbol{r}_d = \boldsymbol{c} - \boldsymbol{z} - \mathcal{A}^T\boldsymbol{y}$.

Table 5.1: *Predictor-corrector algorithm.*

2. By the definition of $\widehat{\theta}$ in (5.1.46), $\widehat{\theta}$ is a decreasing function of δ_ϵ. Thus, there is a trade-off between the allowance for the constraint reduction and the step length in the predictor step.

3. $(\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1}) = \boldsymbol{0}$ when we use the full Schur complement matrix, so by (5.1.36) and (5.1.40), Condition 5.1 and Condition 5.2 can always be satisfied by taking enough

"active blocks" .

4. We will prove that the predictor step moves the point from $\mathcal{N}(\alpha, \tau)$ into $\mathcal{N}(\beta, (1 - \overline{\theta})\tau)$, and the corrector step moves the point into $\mathcal{N}(\alpha, (1 - \overline{\theta})\tau)$.

5. Condition 5.1 and Condition 5.2 restrict the magnitude of $\Delta\boldsymbol{X}_\epsilon$ and $\Delta\overline{\boldsymbol{X}}_\epsilon$, which are the perturbations caused by constraint reduction. Considering that the matrices contain $\operatorname{mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T \Delta\boldsymbol{y}\right)$ and $\operatorname{mat}\left((\widetilde{\boldsymbol{X}} \otimes \widetilde{\boldsymbol{Z}}^{-1})\widetilde{\mathcal{A}}^T \Delta\overline{\boldsymbol{y}}\right)$ before symmetrization, these conditions judge the activeness of the j-th constraint block by the magnitude of $\|\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1}\|$, just as the algorithm in Chapter 4 uses the threshold κ. However, the thresholds are updated dynamically for every iteration, so they are adaptive criteria in contrast to the static κ.

In order to check that the conditions are satisfied, we can solve for $\Delta\boldsymbol{y}$ and $\Delta\bar{\boldsymbol{y}}$ and calculate $\Delta\boldsymbol{X}_\epsilon$ and $\Delta\overline{\boldsymbol{X}}_\epsilon$, which may require the Cholesky factor of $\widehat{\boldsymbol{M}}$ to compute $\Delta\boldsymbol{y} = \widehat{\boldsymbol{M}}^{-1}\boldsymbol{g}$. For practical implementation, we can use rank-1 updating of the Cholesky factor,[2] depending on the size of m and n_j. We now discuss this updating.

Let $\boldsymbol{R}_{X_j}$ and $\boldsymbol{R}_{Z_j}$ be Cholesky factors of $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$. Note that the factor $\boldsymbol{R}_{Z_j}$ is required to compute $\boldsymbol{M}_j$ by (4.3.1), regardless of constraint reduction, unless $\boldsymbol{Z}_j^{-1}$ is computed explicitly. Then, the partial Schur complement $\boldsymbol{M}_j$ can be written as

$$\begin{aligned}\boldsymbol{M}_j &= \mathcal{A}_j(\boldsymbol{X}_j \otimes \boldsymbol{Z}_j^{-1})\mathcal{A}_j^T = \mathcal{A}_j\left((\boldsymbol{R}_{X_j}^T\boldsymbol{R}_{X_j}) \otimes (\boldsymbol{R}_{Z_j}^T\boldsymbol{R}_{Z_j})^{-1}\right)\mathcal{A}_j^T \\ &= \mathcal{A}_j\left((\boldsymbol{R}_{X_j}^T \otimes \boldsymbol{R}_{Z_j}^{-1})(\boldsymbol{R}_{X_j} \otimes \boldsymbol{R}_{Z_j}^{-T})\right)\mathcal{A}_j^T = \boldsymbol{H}_j\boldsymbol{H}_j^T, \qquad (5.1.53)\end{aligned}$$

[2] Rank-1 modification of Cholesky factor is implemented by *"schud.f"* and *"dchud.f"* in LINPACK. See Gill et al. [27] and LINPACK documentation [23].

where

$$\boldsymbol{H}_j = \mathcal{A}_j(\boldsymbol{R}_{X_j}^T \otimes \boldsymbol{R}_{Z_j}^{-1}) \in \mathbb{R}^{m \times n_j^2}.$$

Thus, $\boldsymbol{h}_l^T$, the l-th row of $\boldsymbol{H}_j$, can be computed as

$$\boldsymbol{h}_l = \text{ vec}\left(\boldsymbol{R}_{X_j}\,\boldsymbol{A}_{lj}\,\boldsymbol{R}_{Z_j}^{-1}\right).$$

Furthermore, we can rewrite $(\boldsymbol{M}_j)_{lh}$ in (4.3.1) as

$$\begin{aligned}(\boldsymbol{M}_j)_{lh} &= (\boldsymbol{X}\boldsymbol{A}_{lj}\boldsymbol{Z}_j^{-1}) \bullet \boldsymbol{A}_{hj} = \left((\boldsymbol{R}_{X_j}^T\boldsymbol{R}_{X_j})\boldsymbol{A}_{lj}(\boldsymbol{R}_{Z_j}^{-1}\boldsymbol{R}_{Z_j}^{-T})\right) \bullet \boldsymbol{A}_{hj} \\ &= \left(\boldsymbol{R}_{X_j}^T(\boldsymbol{R}_{X_j}\boldsymbol{A}_{lj}\boldsymbol{R}_{Z_j}^{-1})\boldsymbol{R}_{Z_j}^{-T}\right) \bullet \boldsymbol{A}_{hj} \\ &= \left(\boldsymbol{R}_{X_j}^T\,\text{mat}\,(\boldsymbol{h}_l)\,\boldsymbol{R}_{Z_j}^{-T}\right) \bullet \boldsymbol{A}_{hj}.\end{aligned}$$

Therefore, $\boldsymbol{H}_j$ can be obtained as a byproduct of computing $\boldsymbol{M}_j$ with additional computation for the factor $\boldsymbol{R}_{X_j}$ of $\boldsymbol{X}_j$.

From (5.1.53), we can write the j-th update of $\widehat{\boldsymbol{M}}$ in step 3.(c) and 3.(f) in the algorithm as

$$\widehat{\boldsymbol{M}}^{(j)} = \widehat{\boldsymbol{M}}^{(j-1)} + \boldsymbol{M}_j = \widehat{\boldsymbol{M}}^{(j-1)} + \boldsymbol{H}_j\boldsymbol{H}_j^T.$$

If we already have the Cholesky factor $\boldsymbol{R}_{\widehat{\boldsymbol{M}}}^{(j-1)}$ of $\widehat{\boldsymbol{M}}^{(j-1)}$, the Cholesky factor $\boldsymbol{R}_{\widehat{\boldsymbol{M}}}^{(j)}$ of $\widehat{\boldsymbol{M}}^{(j)}$ can be computed by n_j^2 the rank-1 Cholesky updates. According to Gill et al. [27], the rank-1 update of Cholesky factor requires $2m^2 + O(m)$ flops. Using the updated factor of $\widehat{\boldsymbol{M}}$, we can compute $\Delta\boldsymbol{y} = \widehat{\boldsymbol{M}}^{-1}\boldsymbol{g} = \boldsymbol{R}_{\widehat{\boldsymbol{M}}}^{-1}(\boldsymbol{R}_{\widehat{\boldsymbol{M}}}^{-T}\boldsymbol{g})$ in $2m^2$ flops. Since we do not need a very accurate $\Delta\boldsymbol{y}$ for determining constraint reduction, iterative refinement may not be necessary. Once we finish updating $\widehat{\boldsymbol{M}}$, the factor $\boldsymbol{R}_{\widehat{\boldsymbol{M}}}$ can be reused as a preconditioner for an iterative method like *SYMMLQ* to compute $\Delta\boldsymbol{y}$ to a high accuracy. In summary, for each update of $\widehat{\boldsymbol{M}}$, it takes extra cost for

1. Cholesky factorization of $\boldsymbol{X}_j$: $n_j^3/3$ flops,
2. Update of Cholesky factor of $\widehat{\boldsymbol{M}}$: $n_j^2(2m^2 + O(m))$ flops,
3. Compute $\Delta \boldsymbol{y} = \widehat{\boldsymbol{M}}^{-1}\boldsymbol{g}$: $2m^2$ flops,

so, in total,

$$\frac{1}{3}{n_j}^3 + 2m^2(n_j^2 + 1) + O(mn_j^2).$$

This is a reasonable cost for the constraint reduction decision, considering that it takes $(4m + 1/3)n_j^3 + 2m^2n_j^2$ to compute $\boldsymbol{M}_i$ by (4.3.1) and (4.3.2).

If $m^3/3 < (n_j^3/3 + 2m^2n_j^2)$, then we can compute the Cholesky factor $\boldsymbol{R}_{\widehat{\boldsymbol{M}}}$ of $\widehat{\boldsymbol{M}}$ explicitly with no Cholesky factorization of $\boldsymbol{X}_j$ and no updating of the factor $\boldsymbol{R}_{\widehat{\boldsymbol{M}}}$. In that case, it costs

$$\frac{1}{3}m^3 + 2m^2.$$

5.2 Global Convergence of the Constraint-Reduced SDP Algorithm

5.2.1 Primal and Dual Residuals

The primal and dual residual norms decrease at each iteration, bringing us closer to feasibility.

Lemma 5.2.1. *In the Constraint-Reduced SDP Algorithm,* $\mathbf{r}_d^+ = (1 - \overline{\theta})\mathbf{r}_d$ *and* $\mathbf{r}_p^+ = (1 - \overline{\theta})\mathbf{r}_p$.

Proof. First, let us see how the dual residual changes. By (5.1.34) and (5.1.38),

$$\Delta \boldsymbol{z} = \boldsymbol{r}_d - \mathcal{A}^T \Delta \boldsymbol{y}, \quad \Delta \bar{\boldsymbol{z}} = -\mathcal{A}^T \Delta \bar{\boldsymbol{y}}.$$

So,

$$\begin{aligned}
\boldsymbol{r}_d^+ &= \boldsymbol{c} - \boldsymbol{z}^+ - \mathcal{A}^T \boldsymbol{y}^+ \\
&= \boldsymbol{c} - (\boldsymbol{z} + \overline{\theta}\Delta \boldsymbol{z} + \Delta \bar{\boldsymbol{z}}) - \mathcal{A}^T(\boldsymbol{y} + \overline{\theta}\Delta \boldsymbol{y} + \Delta \bar{\boldsymbol{y}}) \\
&= (\boldsymbol{c} - \boldsymbol{z} - \mathcal{A}^T \boldsymbol{y}) - \overline{\theta}(\Delta \boldsymbol{z} + \mathcal{A}^T \Delta \boldsymbol{y}) - (\Delta \bar{\boldsymbol{z}} + \mathcal{A}^T \Delta \bar{\boldsymbol{y}}) \\
&= \boldsymbol{r}_d - \overline{\theta}\, \boldsymbol{r}_d = (1 - \overline{\theta})\boldsymbol{r}_d.
\end{aligned}$$

Next, we consider the primal residual. By (5.1.33) and (5.1.37),

$$\mathcal{A}\Delta \boldsymbol{x} = \boldsymbol{r}_p, \quad \mathcal{A}\Delta \bar{\boldsymbol{x}} = 0.$$

So,

$$\begin{aligned}
\boldsymbol{r}_p^+ &= \boldsymbol{b} - \mathcal{A}(\boldsymbol{x}^+) \\
&= \boldsymbol{b} - \mathcal{A}(\boldsymbol{x} + \overline{\theta}\Delta \boldsymbol{x} + \Delta \bar{\boldsymbol{x}}) \\
&= \boldsymbol{r}_p - \overline{\theta}\mathcal{A}\Delta \boldsymbol{x} - \mathcal{A}\Delta \bar{\boldsymbol{x}} = \boldsymbol{r}_p - \overline{\theta}\boldsymbol{r}_p \\
&= (1 - \overline{\theta})\boldsymbol{r}_p.
\end{aligned}$$

□

5.2.2 Closeness to Central Path

We analyze how the iterate moves, relative to the central path, during the predictor and corrector steps.

Assume that the current point $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{N}(\alpha, \tau)$. The initial point $(\boldsymbol{X}^0, \boldsymbol{y}^0, \boldsymbol{Z}^0)$ in the algorithm is perfectly placed on the central path, so this assumption is satisfied. With this assumption, we show in Lemma 5.2.3 that

$$(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}) \in \mathcal{N}(\beta, (1-\overline{\theta})\tau), \tag{5.2.1}$$

after the predictor step, and in Lemma 5.2.6 that

$$(\boldsymbol{X}^+, \boldsymbol{Z}^+) \in \mathcal{N}(\alpha, (1-\overline{\theta})\tau), \tag{5.2.2}$$

after the corrector step. In the proofs, we frequently use the relation between Frobenius norm and the eigenvalues of symmetric matrix. For a matrix $\boldsymbol{E} \in \mathbb{R}^{n\times n}$,

$$\|\boldsymbol{E}\|_F^2 = \sum_{i=1}^{n} \sigma_i^2(\boldsymbol{E}),$$

where $\sigma_i(\boldsymbol{E})$ is the i-th singular value of $\boldsymbol{E}$. For a matrix $\boldsymbol{E} \in \mathcal{S}^n$,

$$|\lambda_i(\boldsymbol{E})| \leq \sigma_{\max}(\boldsymbol{E}) = \sqrt{\sigma_{\max}^2(\boldsymbol{E})} \leq \sqrt{\sum_{i=1}^{n} (\sigma_i^2(\boldsymbol{E}))} = \|\boldsymbol{E}\|_F,$$

so

$$-\|\boldsymbol{E}\|_F \leq \lambda_i(\boldsymbol{E}) \leq \|\boldsymbol{E}\|_F. \tag{5.2.3}$$

In addition, the following lemma gives us a bound for a symmetrized matrix.

Lemma 5.2.2. *Suppose that* $\mathbf{M} \in \mathbb{R}^{p\times p}$ *is nonsingular and* $\mathbf{E} \in \mathbb{R}^{p\times p}$ *has only real eigenvalues. Then,*

$$\lambda_{\max}(\mathbf{E}) \quad \leq \quad \lambda_{\max}\left(\text{symm}\left(\mathbf{MEM}^{-1}\right)\right), \tag{5.2.4}$$

$$\lambda_{\min}(\mathbf{E}) \quad \geq \quad \lambda_{\min}\left(\text{symm}\left(\mathbf{MEM}^{-1}\right)\right). \tag{5.2.5}$$

If $\mathbf{E} \in \mathcal{S}^p$*, then*

$$\|\mathbf{E}\|_F \leq \| \operatorname{symm}\left(\mathbf{M}\mathbf{E}\mathbf{M}^{-1}\right) \|_F. \tag{5.2.6}$$

Proof. See [61, Lemma 3.3 in pp.668-669] and [71, Lemma 2.2 in pp.1011-1012]. □

By the definition of $\breve{\theta}$ in (5.1.47), we can prove (5.2.1), by proving that $\widehat{\theta} \leq \breve{\theta}$. The following lemma is a modification of Potra and Sheng [71, Lemma 2.5 in pp.1012-1013].

Lemma 5.2.3. *If* $(\mathbf{X}, \mathbf{Z}) \in \mathcal{N}(\alpha, \tau)$ *then*

$$\widehat{\theta} \leq \breve{\theta}.$$

In particular,

1. *if* $\overline{\theta} < 1$, *then* $(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \in \mathcal{N}(\beta, (1-\overline{\theta})\tau)$, *so* $\overline{\mathbf{X}} \succ \mathbf{0}$ *and* $\overline{\mathbf{Z}} \succ \mathbf{0}$.
2. *if* $\overline{\theta} = 1$, *then* $\overline{\mathbf{X}}\,\overline{\mathbf{Z}} = \mathbf{0}$.

Proof. Let $\boldsymbol{X}(\theta) = \boldsymbol{X} + \theta\Delta\boldsymbol{X}$ and $\boldsymbol{Z}(\theta) = \boldsymbol{Z} + \theta\Delta\boldsymbol{Z}$, then

$$\begin{aligned} \boldsymbol{X}(\theta)\boldsymbol{Z}(\theta) - (1-\theta)\tau\boldsymbol{I} &= (\boldsymbol{X} + \theta\Delta\boldsymbol{X})(\boldsymbol{Z} + \theta\Delta\boldsymbol{Z}) - (1-\theta)\tau\boldsymbol{I} \\ &= (1-\theta)(\boldsymbol{X}\boldsymbol{Z} - \tau\boldsymbol{I}) + \theta(\boldsymbol{X}\boldsymbol{Z} + \boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z}) + \theta^2\Delta\boldsymbol{X}\Delta\boldsymbol{Z}. \end{aligned}$$

Define

$$\begin{aligned} \boldsymbol{P}(\theta) &= \boldsymbol{Z}^{1/2}(\boldsymbol{X}(\theta)\boldsymbol{Z}(\theta) - (1-\theta)\tau\boldsymbol{I})\boldsymbol{Z}^{-1/2} \\ &= \boldsymbol{Z}^{1/2}((\boldsymbol{X} + \theta\Delta\boldsymbol{X})(\boldsymbol{Z} + \theta\Delta\boldsymbol{Z}) - (1-\theta)\tau\boldsymbol{I})\boldsymbol{Z}^{-1/2} \\ &= \boldsymbol{Z}^{1/2}(\boldsymbol{X}\boldsymbol{Z} + \theta(\Delta\boldsymbol{X}\boldsymbol{Z} + \boldsymbol{X}\Delta\boldsymbol{Z}) + \theta^2\Delta\boldsymbol{X}\Delta\boldsymbol{Z} - (1-\theta)\tau\boldsymbol{I})\boldsymbol{Z}^{-1/2} \\ &= (1-\theta)(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}) + \theta^2\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2} \\ &\quad +\theta\left[\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} + \boldsymbol{Z}^{1/2}(\boldsymbol{X}\Delta\boldsymbol{Z} + \Delta\boldsymbol{X}\boldsymbol{Z})\boldsymbol{Z}^{-1/2}\right]. \end{aligned}$$

Then, by (5.1.35),

$$\begin{aligned}\text{symm}\,(\boldsymbol{P}(\theta)) = (1-\theta)(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}) + \theta^2\,\text{symm}\left(\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\right)\\ +\,\theta\left[\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} +\ \text{symm}\left(\boldsymbol{Z}^{1/2}(\boldsymbol{X}\Delta\boldsymbol{Z}+\Delta\boldsymbol{X}\boldsymbol{Z})\boldsymbol{Z}^{-1/2}\right)\right]\\ =(1-\theta)(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}) + \theta^2\,\text{symm}\left(\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\right) - \theta(\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}_\epsilon\,\boldsymbol{Z}^{1/2}).\end{aligned}$$

Thus, since $(\boldsymbol{X},\boldsymbol{Z})\in\mathcal{N}(\alpha,\tau)$, and using (5.1.41), (5.1.43), and (5.2.6), we have

$$\begin{aligned}&\|\,\text{symm}\,(\boldsymbol{P}(\theta))\,\|_F\\ &\leq (1-\theta)\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}-\tau\boldsymbol{I}\|_F + \theta^2\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\|_F + \theta\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}_\epsilon\boldsymbol{Z}^{1/2}\|_F\\ &= \alpha\tau(1-\theta)+\theta^2\delta\tau+\theta\delta_\epsilon\tau. \qquad (5.2.7)\end{aligned}$$

Furthermore, adding and subtracting $\beta(1-\theta)\tau$, then

$$\begin{aligned}\alpha\tau(1-\theta)+\theta^2\delta\tau+\theta\delta_\epsilon\tau &= \tau\left(\delta\theta^2+(\delta_\epsilon-\alpha+\beta)\theta+(\alpha-\beta)\right)+\beta(1-\theta)\tau\\ &= \delta\tau(\theta-\theta_1)(\theta-\theta_2)+\beta(1-\theta)\tau,\end{aligned}$$

where

$$\begin{aligned}\theta_1 &= \frac{(\alpha-\beta-\delta_\epsilon)+\sqrt{(\alpha-\beta-\delta_\epsilon)^2+4\delta(\beta-\alpha)}}{2\delta},\\ \theta_2 &= \frac{(\alpha-\beta-\delta_\epsilon)-\sqrt{(\alpha-\beta-\delta_\epsilon)^2+4\delta(\beta-\alpha)}}{2\delta}.\end{aligned}$$

Since $\widehat{\theta}=\theta_1$ by definition (5.1.46) of $\widehat{\theta}$ and $\theta_2\leq\theta_1$, the first term in the equation above becomes negative when $0\leq\theta\leq\widehat{\theta}$, so writing (5.2.7),

$$\|\,\text{symm}\,(\boldsymbol{P}(\theta))\,\|_F \leq \beta(1-\theta)\tau,\ \ \forall\theta\in[0,\widehat{\theta}].$$

By (5.2.6), with $\boldsymbol{M}=\boldsymbol{Z}^{1/2}$ and $\boldsymbol{E}=\boldsymbol{X}(\theta)\boldsymbol{Z}(\theta)-(1-\theta)\tau\boldsymbol{I}$,

$$\|\boldsymbol{X}(\theta)\boldsymbol{Z}(\theta)-(1-\theta)\tau\boldsymbol{I}\|_F\leq\beta(1-\theta)\tau,\ \forall\theta\in[0,\widehat{\theta}]. \qquad (5.2.8)$$

Note that this implies that $\boldsymbol{X}(1)\boldsymbol{Z}(1) = \mathbf{0}$ when $\widehat{\theta} = 1$. From this result, if $(\boldsymbol{Z}(\theta))^{-1/2}$ exists for $\forall\theta \in [0, \widehat{\theta}]$, then, since the Frobenius norm is invariant under similarity transformation, (5.2.8) implies

$$\|\boldsymbol{Z}(\theta)^{1/2}\boldsymbol{X}(\theta)\boldsymbol{Z}(\theta)^{1/2} - (1-\theta)\tau\boldsymbol{I}\|_F \leq \beta(1-\theta)\tau, \ \forall\theta \in [0, \widehat{\theta}]. \tag{5.2.9}$$

To conclude, we show $\boldsymbol{X}(\theta) \succ 0$ and $\boldsymbol{Z}(\theta) \succ 0$ for $\forall\theta \in [0, \widehat{\theta}]$ when $\widehat{\theta} < 1$; (Claim (5.2.9) holds by continuity for $\widehat{\theta} = 1$ as well.) Otherwise, there must exist $\theta' \in [0, \widehat{\theta}]$ such that $\boldsymbol{X}(\theta')\boldsymbol{Z}(\theta')$ is singular, which implies that

$$\lambda_{\min}(\boldsymbol{X}(\theta')\boldsymbol{Z}(\theta') - (1-\theta')\tau\boldsymbol{I}) \leq -(1-\theta')\tau. \tag{5.2.10}$$

However, by (5.2.5) with $\boldsymbol{M} = \boldsymbol{Z}^{1/2}$ and $\boldsymbol{E} = \boldsymbol{X}(\theta')\boldsymbol{Z}(\theta') - (1-\theta')\tau\boldsymbol{I}$, and by the relation of Frobenius norm and eigenvalues of symmetric matrix in (5.2.3),

$$\begin{aligned} \lambda_{\min}(\boldsymbol{X}(\theta')\boldsymbol{Z}(\theta') - (1-\theta')\tau\boldsymbol{I}) &\geq \lambda_{\min}\left(\text{ symm}\left(\boldsymbol{P}(\theta')\right)\right) \\ &\geq -\|\text{ symm}\left(\boldsymbol{P}(\theta')\right)\|_F \\ &\geq -\beta(1-\theta')\tau, \end{aligned}$$

which contradicts (5.2.10) since $\beta \in (0,1)$. Hence, $\boldsymbol{X}(\theta) \succ 0$ and $\boldsymbol{Z}(\theta) \succ 0$ for $\forall\theta \in [0, \widehat{\theta}]$.

□

Next, we prove that condition (5.2.2) is satisfied after the corrector step. To prepare for this, we need a preliminary lemma, a modification of Monteiro [61, Lemma 4.4 in p.671].

Lemma 5.2.4. *For* $(\mathbf{X}', \mathbf{Z}') \in \mathcal{N}(\gamma, \tau')$ *and* $(\Delta\mathbf{X}', \Delta\mathbf{y}', \Delta\mathbf{Z}')$ *such that*

$$\begin{aligned} \mathbf{A}_i \bullet \Delta\mathbf{X}' = 0 \quad \textit{for } i &= 1, \dots, m, && (5.2.11) \\ \sum_{i=1}^{m} \Delta y_i' \mathbf{A}_i + \Delta\mathbf{Z}' &= \mathbf{0}, && (5.2.12) \end{aligned}$$

define

$$\begin{aligned} \mathbf{H} &= \operatorname{symm}\left(\mathbf{Z}'^{1/2}(\mathbf{X}'\Delta\mathbf{Z}' + \Delta\mathbf{X}'\mathbf{Z}')\mathbf{Z}'^{-1/2}\right), && (5.2.13) \\ \delta_x' &= \|\mathbf{Z}'^{1/2}\Delta\mathbf{X}'\mathbf{Z}'^{1/2}\|_F, && (5.2.14) \\ \delta_z' &= \tau'\|\mathbf{Z}'^{-1/2}\Delta\mathbf{Z}'\mathbf{Z}'^{-1/2}\|_F. && (5.2.15) \end{aligned}$$

Then

$$\begin{aligned} \delta_x'\delta_z' &\le \frac{1}{2}(\delta_x'^{\,2} + \delta_z'^{\,2}) \le \frac{\|\mathbf{H}\|_F^2}{2(1-\gamma)^2}, && (5.2.16) \\ \delta_x' &\le \frac{\|\mathbf{H}\|_F}{1-\gamma}, && (5.2.17) \\ \delta_z' &\le \frac{\|\mathbf{H}\|_F}{1-\gamma}. && (5.2.18) \end{aligned}$$

Proof. Adding and subtracting $(\tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2})$, we compute

$$\begin{aligned} \boldsymbol{H} &= \frac{1}{2}\boldsymbol{Z}'^{1/2}(\boldsymbol{X}'\Delta\boldsymbol{Z}' + \Delta\boldsymbol{X}'\boldsymbol{Z}')\boldsymbol{Z}'^{-1/2} + \frac{1}{2}\boldsymbol{Z}'^{-1/2}(\Delta\boldsymbol{Z}'\boldsymbol{X}' + \boldsymbol{Z}'\Delta\boldsymbol{X}')\boldsymbol{Z}'^{1/2} \\ &= \boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2} + \tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2} + \operatorname{symm}\left((\boldsymbol{Z}'^{1/2}\boldsymbol{X}'\boldsymbol{Z}'^{1/2} - \tau'\boldsymbol{I})\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\right), \end{aligned}$$

so, by using the fact $(\boldsymbol{X}', \boldsymbol{Z}') \in \mathcal{N}(\gamma, \tau')$,

$$\begin{aligned} \|\boldsymbol{H}\|_F &\ge \|\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2} + \tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F - \|(\boldsymbol{Z}'^{1/2}\boldsymbol{X}'\boldsymbol{Z}'^{1/2} - \tau'\boldsymbol{I})\|_F\|\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F \\ &\ge \|\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2} + \tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F - (\gamma\tau')(\delta_z'/\tau') \\ &= \|\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2} + \tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F - \gamma\delta_z'. \end{aligned}$$

Now, the square of the first term in the equation above is

$$\begin{aligned}
&\|\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2} + \tau'\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F^2 \\
&\qquad = \|\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2}\|_F^2 + \tau'^2\|\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}\|_F^2 + 2\tau'(\boldsymbol{Z}'^{1/2}\Delta\boldsymbol{X}'\boldsymbol{Z}'^{1/2}) \bullet (\boldsymbol{Z}'^{-1/2}\Delta\boldsymbol{Z}'\boldsymbol{Z}'^{-1/2}) \\
&\qquad = {\delta'_x}^2 + {\delta'_z}^2 + 2\tau'(\Delta\boldsymbol{Z}' \bullet \Delta\boldsymbol{X}') = {\delta'_x}^2 + {\delta'_z}^2,
\end{aligned}$$

since

$$\Delta\boldsymbol{Z}' \bullet \Delta\boldsymbol{X}' = (-\sum_{i=1}^{m}(\Delta y'_i\boldsymbol{A}_i)) \bullet \Delta\boldsymbol{X}' = (-\sum_{i=1}^{m}(\Delta y'_i\boldsymbol{A}_i \bullet \Delta\boldsymbol{X}')) = 0,$$

by (5.2.11)-(5.2.12). Hence,

$$\|\boldsymbol{H}\|_F \geq \sqrt{{\delta'_x}^2 + {\delta'_z}^2} - \gamma\delta'_z \geq (1-\gamma)\sqrt{{\delta'_x}^2 + {\delta'_z}^2},$$

and the rest of the proof is straightforward. □

One other technical lemma prepares us to prove that condition (5.2.2) is satisfied after the corrector step.

Lemma 5.2.5. *Under Condition 5.2,*

$$\overline{\delta}_\epsilon < (1-\overline{\theta})(1-2\beta).$$

Proof. Recall that

$$s = \beta^2 - \beta + 1, \quad t = 2\alpha(1-\beta)^2 - \beta^2,$$

by their definitions in Condition 5.2. By Condition 5.2 , it suffices to show

$$\sqrt{s^2+t} - s < 1 - 2\beta,$$

or equivalently, since $0 < \beta < 1/2$ and $s > 0$,

$$(s + (1-2\beta))^2 - (\sqrt{s^2+t})^2 > 0.$$

By (5.1.45) and (5.1.52), we have

$$
\begin{aligned}
(s+(1-2\beta))^2-(\sqrt{s^2+t}\,)^2 &= (1-2\beta)^2+2s(1-2\beta)-t \\
&= (1-2\beta)^2+2(\beta^2-\beta+1)(1-2\beta)-2\alpha(1-\beta)^2+\beta^2 \\
&> (1-2\beta)^2+2(\beta^2-\beta+1)(1-2\beta)-2\beta(1-\beta)^2+\beta^2 \\
&= -6\beta^3+15\beta^2-12\beta+3 \\
&= 3(1-2\beta)(\beta-1)^2>0,\ \ \forall\beta\in(0,1/2).
\end{aligned}
$$

So,

$$\overline{\delta}_\epsilon<(1-\overline{\theta})(1-2\beta).$$

□

Now, we are ready to show (5.2.2), which says that $(\boldsymbol{X}^+,\boldsymbol{Z}^+)\in\mathcal{N}(\alpha,(1-\overline{\theta})\tau)$. The following lemma is a modification of Potra and Sheng [71, Theorem 2.6 in pp.1013-1015].

Lemma 5.2.6. *Suppose that* $(\overline{\mathbf{X}},\overline{\mathbf{Z}})\in\mathcal{N}(\beta,(1-\overline{\theta})\tau)$ *in the* predictor-corrector *algorithm. Then, after the corrector step,*

$$(\mathbf{X}^+,\mathbf{Z}^+)\in\mathcal{N}(\alpha,(1-\overline{\theta})\tau).$$

Proof.

$$
\begin{aligned}
\boldsymbol{X}^+\boldsymbol{Z}^+-(1-\overline{\theta})\tau\boldsymbol{I} &= (\overline{\boldsymbol{X}}+\Delta\overline{\boldsymbol{X}})(\overline{\boldsymbol{Z}}+\Delta\overline{\boldsymbol{Z}})-(1-\overline{\theta})\tau\boldsymbol{I} \\
&= \overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}-(1-\overline{\theta})\tau\boldsymbol{I}+\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}+\Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}+\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}.
\end{aligned}
$$

Since $\overline{\theta} < 1$ due to step 3.(e) in Table 5.1, we know that $\overline{\boldsymbol{X}} \succ 0$ and $\overline{\boldsymbol{Z}} \succ 0$ by Lemma 5.2.3. Thus, we can define

$$\begin{aligned}
\boldsymbol{P} &= \overline{\boldsymbol{Z}}^{1/2}(\boldsymbol{X}^{+}\boldsymbol{Z}^{+} - (1-\overline{\theta})\tau\boldsymbol{I})\,\overline{\boldsymbol{Z}}^{(-1/2)} \\
&= \overline{\boldsymbol{Z}}^{1/2}((\overline{\boldsymbol{X}} + \Delta\overline{\boldsymbol{X}})(\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{Z}}) - (1-\overline{\theta})\tau\boldsymbol{I})\,\overline{\boldsymbol{Z}}^{(-1/2)} \\
&= [\overline{\boldsymbol{Z}}^{1/2}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}] + \overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{(-1/2)} + \overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}
\end{aligned}$$

By (5.1.39), we have

$$\begin{aligned}
\text{symm}\,(\boldsymbol{P}) &= [\overline{\boldsymbol{Z}}^{1/2}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}] + \text{ symm}\left(\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{(-1/2)}\right) \\
&\quad + \text{symm}\left(\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\right) \\
&= \text{ symm}\left(\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\right) - \overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_{\epsilon}\overline{\boldsymbol{Z}}^{1/2}. \qquad (5.2.19)
\end{aligned}$$

Since the corrector step satisfies (5.1.37) - (5.1.38) and $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}) \in \mathcal{N}(\beta, (1-\theta)\tau)$, we can apply Lemma 5.2.4 to $\overline{\boldsymbol{X}}$, $\overline{\boldsymbol{Z}}$, $\Delta\overline{\boldsymbol{X}}$, and $\Delta\overline{\boldsymbol{Z}}$. So, with $\gamma = \beta$ and replacing τ' with $(1-\theta)\tau$ and $(\boldsymbol{X}', \boldsymbol{Z}', \Delta\boldsymbol{X}', \Delta\boldsymbol{Z}')$ with $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}, \Delta\overline{\boldsymbol{X}}, \Delta\overline{\boldsymbol{Z}})$, the inequality (5.2.16) divided by τ' becomes

$$\|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2}\|_F \|\overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\|_F \leq \frac{\|\boldsymbol{H}\|_F^2}{2(1-\beta)^2(1-\overline{\theta})\tau}, \qquad (5.2.20)$$

where

$$\boldsymbol{H} = \text{ symm}\left(\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{(-1/2)}\right).$$

In addition, by (5.1.39),

$$\begin{aligned}\|\boldsymbol{H}\|_F &= \|\text{ symm}\left(\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{(-1/2)}\right)\|_F \\ &= \|\overline{\boldsymbol{Z}}^{1/2}(\overline{\boldsymbol{X}} + \Delta\overline{\boldsymbol{X}}_\epsilon)\overline{\boldsymbol{Z}}^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}\|_F \\ &\leq \|\overline{\boldsymbol{Z}}^{1/2}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}\|_F + \|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_\epsilon\overline{\boldsymbol{Z}}^{1/2}\|_F \\ &\leq \beta(1-\overline{\theta})\tau + \overline{\delta}_\epsilon\tau. \quad \text{(since } (\overline{\boldsymbol{X}},\overline{\boldsymbol{Z}}) \in \mathcal{N}(\beta,(1-\theta)\tau)\text{)} \end{aligned} \tag{5.2.21}$$

By (5.2.20) and (5.2.21),

$$\begin{aligned}&\|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2}\|_F\|\overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\|_F \\ &\quad\leq \frac{1}{2(1-\beta)^2(1-\overline{\theta})\tau}\left(\beta(1-\overline{\theta})\tau + \overline{\delta}_\epsilon\tau\right)^2 \\ &\quad=\frac{\beta^2}{2(1-\beta)^2}(1-\overline{\theta})\tau + \frac{\beta}{(1-\beta)^2}\overline{\delta}_\epsilon\tau + \frac{\overline{\delta}_\epsilon^2\tau}{2(1-\beta)^2(1-\overline{\theta})}.\end{aligned} \tag{5.2.22}$$

By Lemma 5.2.4 again, using (5.2.18) divided by τ',

$$\begin{aligned}\frac{\delta'_z}{\tau'} = \|\overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\|_F &= \frac{\|\boldsymbol{H}\|_F}{(1-\beta)(1-\overline{\theta})\tau} \\ &\leq \frac{\beta(1-\overline{\theta})\tau + \overline{\delta}_\epsilon\tau}{(1-\beta)(1-\overline{\theta})\tau} \quad \text{(by (5.2.21))}, \\ &< \frac{\beta}{1-\beta} + \frac{(1-\overline{\theta})(1-2\beta)}{(1-\beta)(1-\overline{\theta})} = 1, \quad \text{(by Lemma 5.2.5)}\end{aligned}$$

so, by (5.2.3)

$$\lambda_{\min}(\overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}) > -1.$$

This implies that $(\boldsymbol{I} + \overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}) \succ 0$, so

$$\boldsymbol{Z}^+ = \overline{\boldsymbol{Z}} + \Delta\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}^{1/2}(\boldsymbol{I} + \overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)})\overline{\boldsymbol{Z}}^{1/2} \succ 0.$$

Therefore, $(\boldsymbol{Z}^+)^{-1/2}$ exists. By defining

$$\boldsymbol{E} = (\boldsymbol{Z}^+)^{1/2}\boldsymbol{X}^+(\boldsymbol{Z}^+)^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}, \quad \boldsymbol{M} = \overline{\boldsymbol{Z}}^{1/2}(\boldsymbol{Z}^+)^{-1/2},$$

we can see that $\boldsymbol{P} = \boldsymbol{MEM}^{-1}$.

Recall that

$$s = \beta^2 - \beta + 1, \quad t = 2\alpha(1-\beta)^2 - \beta^2, \tag{5.2.23}$$

By applying (5.2.6) with these $\boldsymbol{E}$ and $\boldsymbol{M}$, since $\boldsymbol{E} \in \mathcal{S}^n$, we have

$$\begin{aligned}
&\|(\boldsymbol{Z}^+)^{1/2}\boldsymbol{X}^+(\boldsymbol{Z}^+)^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}\|_F \le \|\,\mathrm{symm}\,(\boldsymbol{P})\,\|_F \\
&\quad = \|\,\mathrm{symm}\left(\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\right) - \overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_\epsilon\overline{\boldsymbol{Z}}^{1/2}\|_F \quad \text{(by (5.2.19))} \\
&\quad \le \|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\|_F + \|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_\epsilon\overline{\boldsymbol{Z}}^{1/2}\|_F \\
&\quad \le \|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}\,\overline{\boldsymbol{Z}}^{1/2}\|_F\|\overline{\boldsymbol{Z}}^{(-1/2)}\Delta\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{(-1/2)}\|_F + \|\overline{\boldsymbol{Z}}^{1/2}\Delta\overline{\boldsymbol{X}}_\epsilon\overline{\boldsymbol{Z}}^{1/2}\|_F \\
&\quad \le \frac{\beta^2}{2(1-\beta)^2}(1-\overline{\theta})\tau + \left(\frac{\beta}{(1-\beta)^2} + 1\right)\overline{\delta}_\epsilon\tau + \frac{\overline{\delta}_\epsilon^2\tau}{2(1-\beta)^2(1-\overline{\theta})} \\
&\qquad \text{(by (5.2.22) and (5.1.51) in Condition 5.2)} \\
&\quad = \frac{\tau}{2(1-\beta)^2(1-\overline{\theta})}\left[\beta^2(1-\overline{\theta})^2 + (1-\overline{\theta})(2\beta + 2(1-\beta)^2)\overline{\delta}_\epsilon + \overline{\delta}_\epsilon^2\right] \\
&\qquad \text{(by definition of } s \text{ in (5.2.23))} \\
&\quad < \frac{\tau}{2(1-\beta)^2(1-\overline{\theta})}\left[\beta^2(1-\overline{\theta})^2 + 2(1-\overline{\theta})^2(\beta^2-\beta+1)(\sqrt{s^2+t}-s)\right. \\
&\qquad \left. + (1-\overline{\theta})^2(\sqrt{s^2+t}-s)^2\right] \quad \text{(by Condition 5.2)} \\
&= \frac{\tau}{2(1-\beta)^2(1-\overline{\theta})}\left[\beta^2(1-\overline{\theta})^2 + 2(1-\overline{\theta})^2 s(\sqrt{s^2+t}-s)\right. \\
&\quad \left. + (1-\overline{\theta})^2(s^2+t+s^2-2s\sqrt{s^2+t})\right] \\
&= \frac{(1-\overline{\theta})\tau}{2(1-\beta)^2}\left[\beta^2 + 2s(\sqrt{s^2+t}-s) - 2s(\sqrt{s^2+t}-s) + t\right] \\
&= \frac{(1-\overline{\theta})\tau}{2(1-\beta)^2}(\beta^2+t) = \frac{(1-\overline{\theta})\tau}{2(1-\beta)^2}(2(1-\beta)^2\alpha) \quad \text{(by definition of } t \text{ in (5.2.23))} \\
&= \alpha(1-\overline{\theta})\tau.
\end{aligned}$$

In addition, this implies that

$$\lambda_{\min}((\boldsymbol{Z}^{+})^{1/2}\boldsymbol{X}^{+}(\boldsymbol{Z}^{+})^{1/2} - (1-\overline{\theta})\tau\boldsymbol{I}) \geq -\alpha(1-\overline{\theta})\tau,$$

by (5.2.3), so

$$\lambda_{\min}((\boldsymbol{Z}^{+})^{1/2}\boldsymbol{X}^{+}(\boldsymbol{Z}^{+})^{1/2}) \geq -\alpha(1-\overline{\theta})\tau + (1-\overline{\theta})\tau = (1-\alpha)(1-\overline{\theta})\tau > 0,$$

so $(\boldsymbol{Z}^{+})^{1/2}\boldsymbol{X}^{+}(\boldsymbol{Z}^{+})^{1/2} \succ 0$, and $\boldsymbol{X}^{+} \succ 0$ as well. □

Now, we quantify the bound on the duality gap $\mu = (\boldsymbol{X}\bullet\boldsymbol{Z})/n$. For the analysis, the following properties of Frobenius norm and the trace of a matrix are useful. For a matrix $\boldsymbol{E} \in \mathcal{S}^n$,

$$|\mathrm{tr}\,(\boldsymbol{E})| = \left|\sum_{i=1}^{m} \lambda_i(\boldsymbol{E})\right| \leq \left|\sum_{i=1}^{m} \sigma_i(\boldsymbol{E})\right|,$$

where $\lambda_i(\boldsymbol{E})$ is the i-th eigenvalue and $\sigma_i(\boldsymbol{E})$ is the i-th singular value of $\boldsymbol{E}$.

By the Cauchy-Schwarz inequality, for $\boldsymbol{E} \in \mathcal{S}^n$,

$$n\|\boldsymbol{E}\|_F^2 = n\sum_{i=1}^{m} \sigma_i^2(\boldsymbol{E}) \geq \left(\sum_{i=1}^{n} \sigma_i(\boldsymbol{E})\right)^2 \geq (\mathrm{tr}\,(\boldsymbol{E}))^2,$$

so

$$n\|\boldsymbol{E}\|_F^2 \geq (\mathrm{tr}\,(\boldsymbol{E}))^2 \tag{5.2.24}$$

Lemma 5.2.7. *If* $(\mathbf{X},\mathbf{Z}) \in \mathcal{N}(\alpha,\tau)$*, then*

$$(1-\frac{\alpha}{\sqrt{n}})\tau \leq \mu = \frac{1}{n}(\mathbf{X}\bullet\mathbf{Z}) \leq (1+\frac{\alpha}{\sqrt{n}})\tau.$$

Proof. Since $(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I})$ is symmetric, by (5.2.24),

$$\begin{aligned} n\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\|_F^2 &\geq \left(\mathrm{tr}\left(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\right)\right)^2 \\ &= \left(\mathrm{tr}\left(\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}\right) - n\tau\right)^2 \\ &= (\mathrm{tr}\,(\boldsymbol{X}\boldsymbol{Z}) - n\tau)^2 = (\boldsymbol{X}\bullet\boldsymbol{Z} - n\tau)^2. \end{aligned}$$

Thus, since $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{N}(\alpha, \tau)$,

$$(\boldsymbol{X} \bullet \boldsymbol{Z} - n\tau)^2 \leq n\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\|_F^2 \leq n\alpha^2\tau^2,$$

i.e.,

$$\left(\frac{1}{n}(\boldsymbol{X} \bullet \boldsymbol{Z}) - \tau\right)^2 \leq \frac{1}{n}\alpha^2\tau^2,$$

and the rest of the proof is straightforward. □

5.2.3 Summary of the Progress of the Iteration

We have shown that

$$\begin{aligned}
\boldsymbol{r}_p^+ &= (1-\overline{\theta})\boldsymbol{r}_p, \\
\boldsymbol{r}_d^+ &= (1-\overline{\theta})\boldsymbol{r}_d \quad (\boldsymbol{R}_d^+ = (1-\overline{\theta})\boldsymbol{R}_d), \\
(\boldsymbol{X}^+, \boldsymbol{Z}^+) &\in \mathcal{N}(\alpha, (1-\overline{\theta})\tau), \\
(1-\frac{\alpha}{\sqrt{n}})\tau^+ \leq \mu^+ &= \frac{1}{n}(\boldsymbol{X}^+ \bullet \boldsymbol{Z}^+) \leq (1+\frac{\alpha}{\sqrt{n}})\tau^+, \\
\tau^+ &= (1-\overline{\theta})\tau.
\end{aligned}$$

For the k-iteration, let us define ψ_k as

$$\psi_k := \prod_{i=1}^{k}(1-\overline{\theta}_i).$$

Then, τ_k by the algorithm in Table 5.1 becomes

$$\tau_k = \psi_k\tau_0. \qquad (5.2.25)$$

With these variables, we have the following results.

$$\boldsymbol{r}_p^k = \psi_k \boldsymbol{r}_p^0, \tag{5.2.26}$$

$$\boldsymbol{r}_d^k = \psi_k \boldsymbol{r}_d^0 \quad (\boldsymbol{R}_d^k = \psi_k \boldsymbol{R}_d^0), \tag{5.2.27}$$

$$(\boldsymbol{X}^k, \boldsymbol{Z}^k) \in \mathcal{N}(\alpha, \tau_k), \tag{5.2.28}$$

$$(1 - \frac{\alpha}{\sqrt{n}})\tau_k \leq \mu_k = \frac{1}{n}(\boldsymbol{X}^k \bullet \boldsymbol{Z}^k) \leq (1 + \frac{\alpha}{\sqrt{n}})\tau_k. \tag{5.2.29}$$

In order to prove the convergence of $\boldsymbol{r}_p^k$, $\boldsymbol{r}_d^k$, and μ_k to zero, all that remains is to show that the $\overline{\theta}_i$ are bounded away from zero.

5.2.4 Lower Bound on Step Length

In this section, we omit the k in ψ_k, $\boldsymbol{r}_p^k$, and $\boldsymbol{r}_d^k$ whenever it is evident in the context, and let $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$ denote the k-th iterates of our algorithm.

Lemma 5.2.8. *For any* $(\mathbf{X}^*, \mathbf{y}^*, \mathbf{Z}^*) \in \mathcal{F}^*$*, we have*

$$\begin{aligned} \psi(\mathbf{X} \bullet \mathbf{Z}^0 + \mathbf{X}^0 \bullet \mathbf{Z}) = \mathbf{X} \bullet \mathbf{Z} + \psi^2 \mathbf{X}^0 \bullet \mathbf{Z}^0 \\ + \psi(1-\psi)\mathbf{X}^0 \bullet \mathbf{Z}^* + \psi(1-\psi)\mathbf{X}^* \bullet \mathbf{Z}^0 \\ - (1-\psi)\mathbf{X} \bullet \mathbf{Z}^* - (1-\psi)\mathbf{X}^* \bullet \mathbf{Z}, \end{aligned} \tag{5.2.30}$$

Proof. Let us define

$$\boldsymbol{X}' = \boldsymbol{X} - \psi \boldsymbol{X}^0 - (1-\psi)\boldsymbol{X}^*,$$

$$\boldsymbol{y}' = \boldsymbol{y} - \psi \boldsymbol{y}^0 - (1-\psi)\boldsymbol{y}^*,$$

$$\boldsymbol{Z}' = \boldsymbol{Z} - \psi \boldsymbol{Z}^0 - (1-\psi)\boldsymbol{Z}^*.$$

By (5.1.8), (5.2.26) and the primal feasibility of $\boldsymbol{X}^*$,

$$\begin{aligned} \boldsymbol{A}_i \bullet \boldsymbol{X} &= b_i - r_{pi}, \\ \psi \boldsymbol{A}_i \bullet \boldsymbol{X}^0 &= \psi(b_i - r_{pi}^0) = \psi b_i - r_{pi}, \\ (1-\psi)\boldsymbol{A}_i \bullet \boldsymbol{X}^* &= (1-\psi) b_i, \end{aligned}$$

for $i = 1, \ldots, m$, and by (5.1.9), (5.2.27), and the dual feasibility of $(\boldsymbol{y}^*, \boldsymbol{Z}^*)$

$$\begin{aligned} \sum_{i=1}^m y_i \boldsymbol{A}_i + \boldsymbol{Z} &= \boldsymbol{C} - \boldsymbol{R}_d \\ \psi(\sum_{i=1}^m y_i^0 \boldsymbol{A}_i + \boldsymbol{Z}^0) &= \psi(\boldsymbol{C} - \boldsymbol{R}_d^0) = \psi \boldsymbol{C} - \boldsymbol{R}_d \\ (1-\psi)\left(\sum_{i=1}^m y_i^* \boldsymbol{A}_i + \boldsymbol{Z}^*\right) &= (1-\psi)\boldsymbol{C}. \end{aligned}$$

Thus, $(\boldsymbol{X}', \boldsymbol{y}', \boldsymbol{Z}')$ satisfies

$$\boldsymbol{A}_i \bullet \boldsymbol{X}' = 0 \text{ for } i = 1, \ldots, m,$$
$$\sum_{i=1}^m y_i' \boldsymbol{A}_i + \boldsymbol{Z}' = 0.$$

Therefore, $\boldsymbol{X}' \bullet \boldsymbol{Z}' = \boldsymbol{Z}' \bullet \boldsymbol{X}' = -\sum_{i=1}^m y_i'(\boldsymbol{A}_i \bullet \boldsymbol{X}) = 0$, so

$$[\boldsymbol{X} - \psi \boldsymbol{X}^0 - (1-\psi)\boldsymbol{X}^*] \bullet [\boldsymbol{Z} - \psi \boldsymbol{Z}^0 - (1-\psi)\boldsymbol{Z}^*] = 0.$$

By expanding this equation using $\boldsymbol{X}^* \bullet \boldsymbol{Z}^* = 0$, we can obtain (5.2.30). □

For an initial point $(\boldsymbol{X}^0, \boldsymbol{y}^0, \boldsymbol{Z}^0)$ and an optimal solution $(\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{Z}^*) \in \mathcal{F}^*$, we define ζ as

$$\zeta = \frac{\boldsymbol{X}^0 \bullet \boldsymbol{Z}^* + \boldsymbol{X}^* \bullet \boldsymbol{Z}^0}{\boldsymbol{X}^0 \bullet \boldsymbol{Z}^0}. \tag{5.2.31}$$

Lemma 5.2.9. *(Similar to [71, Lemma 3.2 in p.1016].) For any* $(\mathbf{X}^*, \mathbf{y}^*, \mathbf{Z}^*) \in \mathcal{F}^*$*,*

$$\mathbf{X} \bullet \mathbf{Z}^0 + \mathbf{X}^0 \bullet \mathbf{Z} \leq n\tau_0 \left(2 + \zeta + \frac{\alpha}{\sqrt{n}} \right),$$

where ζ *is defined in (5.2.31).*

Proof. By Lemma 5.2.8, since $\boldsymbol{X} \in \mathcal{S}_+^n, \boldsymbol{Z} \in \mathcal{S}_+^n, \boldsymbol{X}^* \in \mathcal{S}_+^n, \boldsymbol{Z}^* \in \mathcal{S}_+^n$ and $\psi \in [0,1]$,

$$\psi(\boldsymbol{X} \bullet \boldsymbol{Z}^0 + \boldsymbol{X}^0 \bullet \boldsymbol{Z}) \leq \boldsymbol{X} \bullet \boldsymbol{Z} + \psi^2 \boldsymbol{X}^0 \bullet \boldsymbol{Z}^0 + \psi(1-\psi)\boldsymbol{X}^0 \bullet \boldsymbol{Z}^* + \psi(1-\psi)\boldsymbol{X}^* \bullet \boldsymbol{Z}^0.$$

Since $\boldsymbol{X}^0 \bullet \boldsymbol{Z}^0 = n\tau_0$, $\boldsymbol{X} \bullet \boldsymbol{Z} \leq (1+\alpha/\sqrt{n})\psi n\tau_0$ by (5.2.29), $\psi^2 \leq \psi$ and $\psi(1-\psi) \leq \psi$,

$$\begin{aligned}\psi(\boldsymbol{X} \bullet \boldsymbol{Z}^0 + \boldsymbol{X}^0 \bullet \boldsymbol{Z}) &\leq (1+\alpha/\sqrt{n})\psi n\tau_0 + \psi n\tau_0 + \psi\zeta n\tau_0 \\ &\leq \psi n\tau_0 \left[(1+\alpha/\sqrt{n}) + 1 + \zeta\right] = \psi n\tau_0 \left(2 + \zeta + \alpha/\sqrt{n}\right).\end{aligned}$$

□

For the proof of the following corollary and lemmas, we frequently use the following inequality (See Horn and Johnson [42, Exercise 20 in Section 5.6]),

$$\|\boldsymbol{M}_1\boldsymbol{M}_2\|_F \leq \min(\|\boldsymbol{M}_1\|_2\|\boldsymbol{M}_2\|_F, \|\boldsymbol{M}_1\|_F\|\boldsymbol{M}_2\|_2), \ \ \forall \boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n\times n}. \tag{5.2.32}$$

In addition, note that the Frobenius norm $\|\boldsymbol{E}\|_F$ for $\boldsymbol{E} \in \mathbb{R}^{n\times n}$ can be alternatively defined as

$$\|\boldsymbol{E}\|_F = \sqrt{\mathrm{tr}\left(\boldsymbol{E}^T\boldsymbol{E}\right)}. \tag{5.2.33}$$

Corollary 5.2.10. *(Similar to [71, Corollary 3.3 in p.1016].)*

$$\|\mathbf{X}^{1/2}(\mathbf{Z}^0)^{1/2}\|_F \leq (n\tau_0)^{1/2}\left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}, \tag{5.2.34}$$

$$\|\mathbf{Z}^{1/2}(\mathbf{X}^0)^{1/2}\|_F \leq (n\tau_0)^{1/2}\left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}, \tag{5.2.35}$$

$$\|\mathbf{X}^{1/2}\|_F \leq \|(\mathbf{Z}^0)^{-1/2}\|_2\,(n\tau_0)^{1/2}\ \left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}, \tag{5.2.36}$$

$$\|\mathbf{Z}^{1/2}\|_F \leq \|(\mathbf{X}^0)^{-1/2}\|_2\,(n\tau_0)^{1/2}\ \left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}, \tag{5.2.37}$$

$$\|\mathbf{X}^{1/2}\mathbf{Z}^{1/2}\|_2^2 = \|\mathbf{Z}^{1/2}\mathbf{X}\mathbf{Z}^{1/2}\|_2 \leq (1+\alpha)\tau, \tag{5.2.38}$$

$$\|\mathbf{X}^{-1/2}\mathbf{Z}^{-1/2}\|_2^2 = \|\mathbf{Z}^{-1/2}\mathbf{X}^{-1}\mathbf{Z}^{-1/2}\|_2 \leq \frac{1}{(1-\alpha)\tau}. \tag{5.2.39}$$

Proof. First, we prove (5.2.34). By (5.2.33),

$$\begin{aligned}
\|\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}\|_F &= \sqrt{\text{tr}\left((\boldsymbol{Z}^0)^{1/2}\boldsymbol{X}(\boldsymbol{Z}^0)^{1/2}\right)} = \sqrt{\text{tr}\left(\boldsymbol{X}\boldsymbol{Z}^0\right)} \\
&\leq \sqrt{\text{tr}\left(\boldsymbol{X}\boldsymbol{Z}^0\right)+\text{tr}\left(\boldsymbol{X}^0\boldsymbol{Z}\right)} \quad \text{(since } \boldsymbol{X}^0 \in \mathcal{S}^n_+, \boldsymbol{Z} \in \mathcal{S}^n_+ \text{)} \\
&= \sqrt{\boldsymbol{X} \bullet \boldsymbol{Z}^0 + \boldsymbol{X}^0 \bullet \boldsymbol{Z}} \\
&\leq (n\tau_0)^{1/2}\left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}. \quad \text{(by Lemma 5.2.9)}
\end{aligned}$$

In a similar way, (5.2.35) can be proved.

Next, we prove (5.2.36).

$$\begin{aligned}
\|\boldsymbol{X}^{1/2}\|_F &= \|\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}(\boldsymbol{Z}^0)^{(-1/2)}\|_F \leq \|\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}\|_F\|(\boldsymbol{Z}^0)^{(-1/2)}\|_2 \quad \text{(by (5.2.32))} \\
&\leq \|(\boldsymbol{Z}^0)^{-1/2}\|_2\,(n\tau_0)^{1/2}\ \left(2+\zeta+\frac{\alpha}{\sqrt{n}}\right)^{1/2}. \quad \text{(by (5.2.34) proven above)}
\end{aligned}$$

In a similar way, we can also prove (5.2.37).

Next, we prove (5.2.38). The equality is satisfied since $\sigma^2_{\max}(\boldsymbol{E}) = \sigma_{\max}(\boldsymbol{E}^T\boldsymbol{E})$ for any matrix $\boldsymbol{E}$. Because $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{N}(\alpha, \tau)$,

$$\begin{aligned}
\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\|_2 &\leq \|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2} - \tau\boldsymbol{I}\|_F \leq \alpha\tau, \\
\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_2 - \tau &\leq \alpha\tau, \\
\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_2 &\leq \tau + \alpha\tau = (1+\alpha)\tau.
\end{aligned}$$

In a similar way, (5.2.39) can be proved. □

For a predictor direction $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Z}$, we define δ_x and δ_z as

$$\delta_x = \|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_F, \tag{5.2.40}$$

$$\delta_z = \tau\|\boldsymbol{Z}^{-1/2}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\|_F. \tag{5.2.41}$$

Then, δ defined in (5.1.41) is bounded by

$$\begin{aligned}
\delta &= \frac{1}{\tau}\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\|_F \\
&\leq \frac{1}{\tau}\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_F\|\boldsymbol{Z}^{-1/2}\Delta\boldsymbol{X}\boldsymbol{Z}^{-1/2}\|_F = \frac{1}{\tau^2}\delta_x\delta_z.
\end{aligned} \tag{5.2.42}$$

Lemma 5.2.11. *(Similar to [71, Lemma 3.4 in pp.1016-1018].) For $(\breve{\mathbf{X}}, \breve{\mathbf{y}}, \breve{\mathbf{Z}}) \in \mathcal{F}$, denote*

$$\mathbf{T} = \psi\left[\mathbf{Z}^{1/2}(\mathbf{X}^0 - \breve{\mathbf{X}})\mathbf{Z}^{1/2} + \text{ symm}\left(\mathbf{Z}^{1/2}\mathbf{X}(\mathbf{Z}^0 - \breve{\mathbf{Z}})\mathbf{Z}^{-1/2}\right)\right] - \mathbf{Z}^{1/2}(\mathbf{X} + \Delta\mathbf{X}_\epsilon)\mathbf{Z}^{1/2},$$

$$\mathbf{T}_x = \psi\mathbf{Z}^{1/2}(\mathbf{X}^0 - \breve{\mathbf{X}})\mathbf{Z}^{1/2},$$

$$\mathbf{T}_z = \psi\mathbf{Z}^{-1/2}(\mathbf{Z}^0 - \breve{\mathbf{Z}})\mathbf{Z}^{-1/2}.$$

Then,

$$\delta_x = \|\mathbf{Z}^{1/2}\Delta\mathbf{X}\mathbf{Z}^{1/2}\|_F \leq \|\mathbf{T}_x\|_F + \frac{\|\mathbf{T}\|_F}{1-\alpha}, \tag{5.2.43}$$

$$\delta_z = \tau\|\mathbf{Z}^{-1/2}\Delta\mathbf{Z}\mathbf{Z}^{-1/2}\|_F \leq \tau\|\mathbf{T}_z\|_F + \frac{\|\mathbf{T}\|_F}{1-\alpha}. \tag{5.2.44}$$

Proof. We will use Lemma 5.2.4 with $(\boldsymbol{X}', \boldsymbol{y}', \boldsymbol{Z}') = (\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$, $(\Delta\boldsymbol{X}', \Delta\boldsymbol{y}', \Delta\boldsymbol{Z}') = (\Delta\boldsymbol{X} + \psi(\boldsymbol{X}^0 - \breve{\boldsymbol{X}}), \Delta\boldsymbol{y} + \psi(\boldsymbol{y}^0 - \breve{\boldsymbol{y}}), \Delta\boldsymbol{Z} + \psi(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}}))$, $\gamma = \alpha$, and $\tau' = \tau$. For a predictor direction $(\Delta\boldsymbol{X}, \Delta\boldsymbol{y}, \Delta\boldsymbol{Z})$, by (5.1.5) and (5.1.8),

$$\begin{aligned} \boldsymbol{A}_i \bullet \Delta\boldsymbol{X} &= r_{pi}, \\ \psi(\boldsymbol{A}_i \bullet \boldsymbol{X}^0) &= \psi(b_i - r_{pi}^0) = \psi b_i - r_{pi}, \end{aligned}$$

and since $\breve{\boldsymbol{X}}$ is feasible,

$$\psi(\boldsymbol{A}_i \bullet \breve{\boldsymbol{X}}) = \psi b_i,$$

for $i = 1, \ldots, m$. Hence, $\boldsymbol{A}_i \bullet (\Delta\boldsymbol{X} + \psi(\boldsymbol{X}^0 - \breve{\boldsymbol{X}})) = 0$.

Also, by (5.1.6) and (5.1.9)

$$\begin{aligned} \left(\sum_{i=1}^{m} \Delta y_i \boldsymbol{A}_i\right) + \Delta\boldsymbol{Z} &= \boldsymbol{R}_d, \\ \psi\left[\left(\sum_{i=1}^{m} y_i^0 \boldsymbol{A}_i\right) + \boldsymbol{Z}^0\right] &= \psi(\boldsymbol{C} - \boldsymbol{R}_d^0) = \psi\boldsymbol{C} - \boldsymbol{R}_d, \\ \psi\left[\left(\sum_{i=1}^{m} \breve{y}_i \boldsymbol{A}_i\right) + \breve{\boldsymbol{Z}}\right] &= \psi\boldsymbol{C}. \end{aligned}$$

Thus, $(\Delta\boldsymbol{y} + \psi(\boldsymbol{y}^0 - \breve{\boldsymbol{y}}), \Delta\boldsymbol{Z} + \psi(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}}))$ satisfies (5.2.12).

In addition, since $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{N}(\alpha, \tau)$, we can use Lemma 5.2.4 by replacing γ with α, τ' with τ, $(\boldsymbol{X}', \boldsymbol{y}', \boldsymbol{Z}')$ with $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$, and $(\Delta\boldsymbol{X}', \Delta\boldsymbol{y}', \Delta\boldsymbol{Z}')$ with $(\Delta\boldsymbol{X} + \psi(\boldsymbol{X}^0 - \breve{\boldsymbol{X}}), \Delta\boldsymbol{y} + \psi(\boldsymbol{y}^0 - \breve{\boldsymbol{y}}), \Delta\boldsymbol{Z} + \psi(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}}))$. Then, using (5.1.35), $\boldsymbol{H}$ in Lemma 5.2.4 becomes $\boldsymbol{T}$.

Therefore, from Lemma 5.2.4, using (5.2.17) and (5.2.18), we have the following inequalities,

$$\|\boldsymbol{Z}^{1/2}(\Delta\boldsymbol{X}+\psi(\boldsymbol{X}^0-\breve{\boldsymbol{X}}))\boldsymbol{Z}^{1/2}\|_F \le \frac{\|\boldsymbol{T}\|_F}{1-\alpha},$$
$$\tau\|\boldsymbol{Z}^{-1/2}(\Delta\boldsymbol{Z}+\psi(\boldsymbol{Z}^0-\breve{\boldsymbol{Z}}))\boldsymbol{Z}^{-1/2}\|_F \le \frac{\|\boldsymbol{T}\|_F}{1-\alpha}.$$

Hence,

$$\begin{aligned}
\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_F &\le \frac{\|\boldsymbol{T}\|_F}{1-\alpha}+\psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0-\breve{\boldsymbol{X}})\boldsymbol{Z}^{1/2}\|_F,\\
&= \frac{\|\boldsymbol{T}\|_F}{1-\alpha}+\|\boldsymbol{T}_x\|_F\\
\tau\|\boldsymbol{Z}^{-1/2}\Delta\boldsymbol{Z}\boldsymbol{Z}^{-1/2}\|_F &\le \frac{\|\boldsymbol{T}\|_F}{1-\alpha}+\tau\psi\|\boldsymbol{Z}^{-1/2}(\boldsymbol{Z}^0-\breve{\boldsymbol{Z}})\boldsymbol{Z}^{-1/2}\|_F\\
&= \frac{\|\boldsymbol{T}\|_F}{1-\alpha}+\tau\|\boldsymbol{T}_z\|_F.
\end{aligned}$$

□

Lemma 5.2.12. *For given any* $(\breve{\mathbf{X}},\breve{\mathbf{y}},\breve{\mathbf{Z}})\in\mathcal{F}$, *we have*

$$\delta \le \left(\frac{1-\alpha}{1-\alpha-q}\right)^2\left[\frac{(3-\alpha)\,(2+\zeta+\alpha/\sqrt{n})}{(1-\alpha)^2}nd_0+\sqrt{n}\left(\frac{1+\alpha}{1-\alpha}\right)\right]^2, \qquad (5.2.45)$$

where

$$d_0=\max\left(\|(\mathbf{X}^0)^{-1/2}(\mathbf{X}^0-\breve{\mathbf{X}})(\mathbf{X}^0)^{-1/2}\|_F,\|(\mathbf{Z}^0)^{-1/2}(\mathbf{Z}^0-\breve{\mathbf{Z}})(\mathbf{Z}^0)^{-1/2}\|_F\right).$$

Proof. First, we calculate bounds on $\|\boldsymbol{T}_x\|$, $\|\boldsymbol{T}_z\|$, and $\|\boldsymbol{T}\|$ in Lemma 5.2.11.

By Corollary 5.2.10, we have

$$\begin{aligned}
\|\boldsymbol{T}_x\|_F &= \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0-\breve{\boldsymbol{X}})\boldsymbol{Z}^{1/2}\|_F\\
&= \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0)^{1/2}(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0-\breve{\boldsymbol{X}})(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0)^{1/2}\boldsymbol{Z}^{1/2}\|_F\\
&\le \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0)^{1/2}\|_F^2\ \ \|(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0-\breve{\boldsymbol{X}})(\boldsymbol{X}^0)^{-1/2}\|_F\\
&\le \psi n\tau_0(2+\zeta+\alpha/\sqrt{n})d_0 = n\tau(2+\zeta+\alpha/\sqrt{n})d_0,
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{T}_z\|_F &= \psi\|\boldsymbol{Z}^{-1/2}(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}})\boldsymbol{Z}^{-1/2}\|_F \\
&\leq \psi\|\boldsymbol{Z}^{-1/2}\boldsymbol{X}^{-1/2}\|_2^2 \;\; \|\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}\|_F^2 \;\; \|(\boldsymbol{Z}^0)^{-1/2}(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}})(\boldsymbol{Z}^0)^{-1/2}\|_F \\
&\leq \frac{\psi n\tau_0\,(2+\zeta+\alpha/\sqrt{n})}{(1-\alpha)\tau} d_0 = nd_0 \frac{(2+\zeta+\alpha/\sqrt{n})}{(1-\alpha)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\boldsymbol{T}\|_F &\leq \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0 - \breve{\boldsymbol{X}})\boldsymbol{Z}^{1/2}\|_F + \psi\|\boldsymbol{Z}^{1/2}\boldsymbol{X}(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}})\boldsymbol{Z}^{-1/2}\|_F + \|\boldsymbol{Z}^{1/2}(\boldsymbol{X} + \Delta\boldsymbol{X}_\epsilon)\boldsymbol{Z}^{1/2}\|_F \\
&= \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0)^{1/2}(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0 - \breve{\boldsymbol{X}})(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0)^{1/2}\boldsymbol{Z}^{1/2}\|_F \\
&\quad + \psi\|\boldsymbol{Z}^{1/2}\boldsymbol{X}^{1/2}\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}(\boldsymbol{Z}^0)^{-1/2}(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}})(\boldsymbol{Z}^0)^{-1/2}(\boldsymbol{Z}^0)^{1/2}\boldsymbol{X}^{1/2}\boldsymbol{X}^{-1/2}\boldsymbol{Z}^{-1/2}\|_F \\
&\quad + \|\boldsymbol{Z}^{1/2}(\boldsymbol{X} + \Delta\boldsymbol{X}_\epsilon)\boldsymbol{Z}^{1/2}\|_F \\
&\leq \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0)^{1/2}\|_F^2\|(\boldsymbol{X}^0)^{-1/2}(\boldsymbol{X}^0 - \breve{\boldsymbol{X}})(\boldsymbol{X}^0)^{-1/2}\|_F \\
&\quad + \psi\|\boldsymbol{Z}^{1/2}\boldsymbol{X}^{1/2}\|_2\|\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0)^{1/2}\|_F^2\|(\boldsymbol{Z}^0)^{-1/2}(\boldsymbol{Z}^0 - \breve{\boldsymbol{Z}})(\boldsymbol{Z}^0)^{-1/2}\|_F\|\boldsymbol{X}^{-1/2}\boldsymbol{Z}^{-1/2}\|_2 \\
&\quad + \sqrt{n}\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_2 + \|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}_\epsilon\boldsymbol{Z}^{1/2}\|_F \\
&\leq \psi n\tau_0(2+\zeta+\alpha/\sqrt{n})d_0 + \psi n\tau_0(2+\zeta+\alpha/\sqrt{n})d_0\sqrt{\frac{1+\alpha}{1-\alpha}} \\
&\quad + \sqrt{n}(1+\alpha)\tau + \delta_\epsilon\tau \quad \text{(by definition of } \delta_\epsilon \text{ in (5.1.43))} \\
&\leq n\tau d_0(2+\zeta+\alpha/\sqrt{n}) + n\tau d_0(2+\zeta+\alpha/\sqrt{n})\left(\frac{1+\alpha}{1-\alpha}\right) + \sqrt{n}(1+\alpha)\tau + \delta_\epsilon\tau \\
&\leq \tau\left[\frac{2nd_0(2+\zeta+\alpha/\sqrt{n})}{1-\alpha} + \sqrt{n}(1+\alpha)\right] + \delta_\epsilon\tau \\
&\leq \tau\left[\frac{2nd_0(2+\zeta+\alpha/\sqrt{n})}{1-\alpha} + \sqrt{n}(1+\alpha)\right] + \delta_x q. \text{ (by Condition 5.1)}
\end{aligned}$$

For simple notation, let C_x, C_z, and C_0 denote

$$\begin{aligned} C_x &= n(2+\zeta+\alpha/\sqrt{n})d_0, \\ C_z &= n\frac{(2+\zeta+\alpha/\sqrt{n})}{1-\alpha}d_0, \\ C_0 &= \frac{2nd_0(2+\zeta+\alpha/\sqrt{n})}{1-\alpha} + \sqrt{n}(1+\alpha), \end{aligned}$$

then we can rewrite the bounds on $\|\boldsymbol{T}_x\|_F$, $\|\boldsymbol{T}_z\|$, and $\|\boldsymbol{T}\|$ as

$$\|\boldsymbol{T}_x\|_F \le C_x\tau, \quad \|\boldsymbol{T}_z\|_F \le C_z, \quad \|\boldsymbol{T}\|_F \le C_0\tau + \delta_x q,$$

By Lemma 5.2.11 and the bounds on $\|\boldsymbol{T}_x\|_F$ and $\|\boldsymbol{T}\|_F$ above, we have

$$\begin{aligned} \delta_x &\le \|\boldsymbol{T}_x\|_F + \frac{\|\boldsymbol{T}\|_F}{1-\alpha} \le C_x\tau + \frac{C_0\tau+\delta_x q}{1-\alpha}, \\ & \left(\frac{1-\alpha-q}{1-\alpha}\right)\delta_x \le C_x\tau + \frac{C_0\tau}{1-\alpha}, \\ \delta_x &\le \left(\frac{1-\alpha}{1-\alpha-q}\right)\left(C_x + \frac{C_0}{1-\alpha}\right)\tau. \quad \text{(since } 1-\alpha-q>0 \text{ by (5.1.50))} \end{aligned}$$

In addition, by Lemma 5.2.11 and the bounds on $\|\boldsymbol{T}_z\|_F$ and $\|\boldsymbol{T}\|_F$ above, we have

$$\begin{aligned} \delta_z &\le C_z\tau + \frac{C_0\tau+\delta_x q}{1-\alpha} \le \left(C_z + \frac{C_0}{1-\alpha}\right)\tau + \frac{q}{1-\alpha}\delta_x \\ &\le \left(C_z + \frac{C_0}{1-\alpha}\right)\tau + \frac{q}{1-\alpha-q}\left(C_x + \frac{C_0}{1-\alpha}\right)\tau. \quad \text{(by the bound of } \delta_x \text{ above)} \end{aligned}$$

Finally, by (5.2.42),

$$\begin{aligned} \delta &\le \frac{1}{\tau^2}\delta_x\delta_z \\ &\le \left(\frac{1-\alpha}{1-\alpha-q}\right)\left(C_x + \frac{C_0}{1-\alpha}\right)\left(C_z + \frac{C_0}{1-\alpha} + \frac{q}{1-\alpha-q}\left(C_x + \frac{C_0}{1-\alpha}\right)\right) \\ &\le \left(\frac{1-\alpha}{1-\alpha-q}\right)\left(C_x + \frac{C_0}{1-\alpha}\right)\left(C_z + \frac{C_0}{1-\alpha}\right) + \frac{q(1-\alpha)}{(1-\alpha-q)^2}\left(C_x + \frac{C_0}{1-\alpha}\right)^2. \end{aligned}$$

By definitions of C_x, C_z, and C_0, since $0 < \alpha < 1/2$,

$$\left(C_x + \frac{C_0}{1-\alpha}\right) < \left(C_z + \frac{C_0}{1-\alpha}\right),$$

so we have

$$\begin{aligned}
\delta &\leq \left(\frac{1-\alpha}{1-\alpha+q} + \frac{q(1-\alpha)}{(1-\alpha-q)^2}\right)\left(C_z + \frac{C_0}{1-\alpha}\right)^2 \\
&= \left(\frac{(1-\alpha)(1-\alpha-q)+q(1-\alpha)}{(1-\alpha+q)^2}\right)\left(C_z + \frac{C_0}{1-\alpha}\right)^2 \\
&= \left(\frac{1-\alpha}{1-\alpha+q}\right)^2\left(C_z + \frac{C_0}{1-\alpha}\right)^2 \\
&= \left(\frac{1-\alpha}{1-\alpha+q}\right)^2\left[\frac{(3-\alpha)\left(2+\zeta+\alpha/\sqrt{n}\right)}{(1-\alpha)^2}nd_0 + \sqrt{n}\left(\frac{1+\alpha}{1-\alpha}\right)\right]^2,
\end{aligned}$$

and we obtain (5.2.45). □

Since δ is bounded, $\widehat{\theta}$ defined by (5.1.46) is bounded away from 0. Thus, the step length $\overline{\theta} \in [\widehat{\theta}, \breve{\theta}]$ is also bounded away from 0.

5.2.5 Polynomial Complexity

We prove that our algorithm converges in $O(n \ln(\epsilon_0/\epsilon))$ iterations, the same as the (unreduced) algorithm of [71], where

$$\epsilon_0 = \max\left(\boldsymbol{X}_0 \bullet \boldsymbol{Z}_0, \|\boldsymbol{r}_p^0\|, \|\boldsymbol{r}_d^0\|\right)$$

and ϵ is the required tolerance on

$$\max\left(\boldsymbol{X}^k \bullet \boldsymbol{Z}^k, \|\boldsymbol{r}_p^k\|, \|\boldsymbol{r}_d^k\|\right).$$

Again, we omit the index k for simplicity of notation.

Lemma 5.2.13. *Suppose that* $\mathbf{X}^0 = \mathbf{Z}^0 = \rho\mathbf{I}$ *where* $\rho > 0$ *is a constant such that* $\|\mathbf{X}^*\|_2 \leq \rho$ *and* $\|\mathbf{Z}^*\|_2 \leq \rho$ *for* $(\mathbf{X}^*, \mathbf{y}^*, \mathbf{Z}^*) \in \mathcal{F}^*$*. Then the predictor step length* $\overline{\theta}_k \in [\widehat{\theta}_k, \breve{\theta}_k]$ *satisfies*

$$\overline{\theta}_k \geq \frac{1}{wn},$$

where

$$w = 1 + \frac{hq}{(\beta - \alpha)} + \sqrt{\frac{h(h+3.5)}{\beta - \alpha}},$$

and $h = 13/(0.5 - q)$*.*

Proof. By Lemma 5.2.9, we have

$$\rho(\text{tr}\,(\boldsymbol{X}) + \text{tr}\,(\boldsymbol{Z})) \leq (2 + \zeta + \alpha/\sqrt{n})n\tau_0 = (2 + \zeta + \alpha/\sqrt{n})n\rho^2,$$

so

$$\sum_{i=1}^{n}(\lambda_i(\boldsymbol{X}) + \lambda_i(\boldsymbol{Z})) \leq (2 + \zeta + \alpha/\sqrt{n})n\rho.$$

From (5.1.45), we have

$$\alpha/\sqrt{n} \leq \alpha \leq 1/2.$$

Since $\boldsymbol{X}^* \bullet \boldsymbol{Z}^* = 0$,

$$\begin{aligned}\zeta &= (\boldsymbol{Z}^* \bullet \boldsymbol{X}^0 + \boldsymbol{X}^* \bullet \boldsymbol{Z}^0)/(\boldsymbol{X}^0 \bullet \boldsymbol{Z}^0)\\ &= (\text{tr}\,(\boldsymbol{X}^*) + \text{tr}\,(\boldsymbol{Z}^*))/(n\rho) \leq 1,\end{aligned}$$

which implies

$$\|\boldsymbol{X}^{1/2}\|_F^2 + \|\boldsymbol{Z}^{1/2}\|_F^2 = \sum_{i=1}^{n}(\lambda_i(\boldsymbol{X}) + \lambda_i(\boldsymbol{Z})) \leq (3 + \alpha/\sqrt{n})\rho n \leq 3.5\rho n. \qquad (5.2.46)$$

In addition, we can see

$$\|\boldsymbol{X}^0-\boldsymbol{X}^*\|_2\le\rho,\;\;\|\boldsymbol{Z}^0-\boldsymbol{Z}^*\|_2\le\rho. \tag{5.2.47}$$

By (5.2.46), (5.2.47) and Corollary 5.2.10,

$$\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0-\boldsymbol{X}^*)\boldsymbol{Z}^{1/2}\|_2\le\|\boldsymbol{Z}^{1/2}\|_F^2\|\boldsymbol{X}^0-\boldsymbol{X}^*\|_2\le 3.5\rho^2n, \tag{5.2.48}$$

$$\begin{aligned}\|\boldsymbol{Z}^{1/2}\boldsymbol{X}(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\boldsymbol{Z}^{-1/2}\|_2 &\le\|(\boldsymbol{Z}^{1/2}\boldsymbol{X}^{1/2})\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\boldsymbol{X}^{1/2}(\boldsymbol{X}^{-1/2}\boldsymbol{Z}^{-1/2})\|_2\\ &\le\|(\boldsymbol{Z}^{1/2}\boldsymbol{X}^{1/2})\|_2\|(\boldsymbol{X}^{-1/2}\boldsymbol{Z}^{-1/2})\|_2\|\boldsymbol{X}^{1/2}\|_F^2\|(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\|_2\\ &\le\left(\sqrt{\frac{1+\alpha}{1-\alpha}}\right)3.5\rho^2n\le 6.1\rho^2n. \end{aligned}\tag{5.2.49}$$

By (5.2.48), (5.2.49), and Corollary 5.2.10, in Lemma 5.2.11 with $(\breve{\boldsymbol{X}},\breve{\boldsymbol{y}},\breve{\boldsymbol{Z}})=(\boldsymbol{X}^*,\boldsymbol{y}^*,\boldsymbol{Z}^*)$,

$$\|\boldsymbol{T}_x\|_F \le \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0-\boldsymbol{X}^*)\boldsymbol{Z}^{1/2}\|_F\le 3.5\psi\rho^2n=3.5n\tau, \tag{5.2.50}$$

$$\begin{aligned}\tau\|\boldsymbol{T}_z\|_F &\le \tau\psi\|(\boldsymbol{Z}^{-1/2}\boldsymbol{X}^{-1/2})\boldsymbol{X}^{1/2}(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\boldsymbol{X}^{1/2}(\boldsymbol{X}^{-1/2}\boldsymbol{Z}^{-1/2})\|_F\\ &\le \tau\psi\|(\boldsymbol{Z}^{-1/2}\boldsymbol{X}^{-1/2})\|_2^2\|\boldsymbol{X}^{1/2}\|_F^2\|(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\|_F\\ &\le 3.5\tau\psi\rho^2n/(0.5\tau)=7n\tau. \end{aligned}\tag{5.2.51}$$

Similarly, by (5.2.48), (5.2.49), (5.2.38), and (5.1.43)

$$\begin{aligned}\|\boldsymbol{T}\|_F &\le \psi\|\boldsymbol{Z}^{1/2}(\boldsymbol{X}^0-\boldsymbol{X}^*)\boldsymbol{Z}^{1/2}\|_F+\psi\|\boldsymbol{Z}^{1/2}\boldsymbol{X}(\boldsymbol{Z}^0-\boldsymbol{Z}^*)\boldsymbol{Z}^{-1/2}\|_F\\ &\quad+\|\boldsymbol{Z}^{1/2}\boldsymbol{X}\boldsymbol{Z}^{1/2}\|_F+\|\boldsymbol{Z}^{1/2}\Delta\boldsymbol{X}_\epsilon\boldsymbol{Z}^{1/2}\|_F\\ &\le (3.5\psi\rho^2n+6.1\psi\rho^2n+1.5n\tau)+\delta_\epsilon\tau\\ &\le 11.1n\tau+\delta_xq.\quad \text{(by Condition 5.1)}\end{aligned}$$

By the bound of δ_x in Lemma 5.2.11,

$$\begin{aligned}
\|\boldsymbol{T}\|_F &\leq 11.1n\tau + \left(\|\boldsymbol{T}_x\|_F + \frac{\|\boldsymbol{T}\|_F}{1-\alpha}\right) q, \\
\left(\frac{1-\alpha-q}{1-\alpha}\right) \|\boldsymbol{T}\|_F &\leq 11.1n\tau + q\|\boldsymbol{T}_x\|_F, \\
\|\boldsymbol{T}\|_F &\leq \left(\frac{1-\alpha}{1-\alpha-q}\right) (11.1n\tau + q\|\boldsymbol{T}_x\|_F) .
\end{aligned}$$

Furthermore, by the bound of $\|\boldsymbol{T}_x\|_F$ above, we have

$$\|\boldsymbol{T}\|_F \leq \left(\frac{1-\alpha}{1-\alpha-q}\right) (11.1n\tau + 3.5qn\tau) = \left(\frac{1-\alpha}{1-\alpha-q}\right) (11.1 + 3.5q)n\tau. \quad (5.2.52)$$

By Lemma 5.2.11 with (5.2.50) and (5.2.52),

$$\begin{aligned}
\delta_x &\leq \|\boldsymbol{T}_x\|_F + \frac{\|\boldsymbol{T}\|_F}{1-\alpha} \leq 3.5n\tau + \frac{(11.1+3.5q)n\tau}{1-\alpha-q} \\
&\leq 3.5n\tau + \frac{(11.1+3.5q)n\tau}{0.5-q} \quad \text{(since } \alpha < 0.5\text{)} \\
&\leq \left(3.5 + \frac{11.1+3.5q}{0.5-q}\right) n\tau = \left(\frac{12.85}{0.5-q}\right) n\tau,
\end{aligned}$$

so, by the definition of h,

$$\delta_x \leq hn\tau. \quad (5.2.53)$$

Similarly, by Lemma 5.2.11 with (5.2.51) and (5.2.52),

$$\begin{aligned}
\delta_z &\leq \tau\|\boldsymbol{T}_z\|_F + \frac{\|\boldsymbol{T}\|_F}{1-\alpha} \leq 7n\tau + \frac{(11.1+3.5q)n\tau}{1-\alpha-q} \\
&\leq 7n\tau + \frac{(11.1+3.5q)n\tau}{0.5-q} \quad \text{(since } \alpha < 0.5\text{)} \\
&\leq \left(7 + \frac{11.1+3.5q}{0.5-q}\right) n\tau = \left(3.5 + \frac{12.85}{0.5-q}\right) n\tau,
\end{aligned}$$

so, by the definition of h,

$$\delta_z \leq (h + 3.5)n\tau. \quad (5.2.54)$$

Therefore, by (5.2.42), (5.2.53), and (5.2.54),

$$\delta \leq \frac{1}{\tau^2}\delta_x\delta_z \leq \frac{1}{\tau^2}(hn\tau)\left((h+3.5)n\tau\right) \leq h(h+3.5)n^2. \tag{5.2.55}$$

By Condition 5.1 and (5.2.53),

$$\delta_\epsilon \leq \frac{q}{\tau}\delta_x \leq \frac{q}{\tau}(hn\tau) = qnh. \tag{5.2.56}$$

By the definition of $\widehat{\theta}$ in (5.1.46),

$$\begin{aligned}
\widehat{\theta} &= \frac{2(\beta-\alpha)}{\sqrt{(\beta-\alpha+\delta_\epsilon)^2+4\delta(\beta-\alpha)}-(\beta-\alpha+\delta_\epsilon)} \\
&= \frac{2}{\sqrt{\left(1+\frac{\delta_\epsilon}{\beta-\alpha}\right)^2+\frac{4\delta}{\beta-\alpha}}-\left(1+\frac{\delta_\epsilon}{\beta-\alpha}\right)} \\
&\geq \frac{2}{\sqrt{\left(1+\frac{\delta_\epsilon}{\beta-\alpha}\right)^2+\frac{4\delta}{\beta-\alpha}}} \\
&\geq \frac{2}{\left(1+\frac{\delta_\epsilon}{\beta-\alpha}\right)+\sqrt{\frac{4\delta}{\beta-\alpha}}} \quad \text{(since } \sqrt{x}+\sqrt{y} \geq \sqrt{x+y}\text{)} \\
&= \frac{1}{\frac{1}{2}\left(1+\frac{\delta_\epsilon}{\beta-\alpha}\right)+\sqrt{\frac{\delta}{\beta-\alpha}}} \geq \frac{1}{\left(n+\frac{\delta_\epsilon}{\beta-\alpha}\right)+\sqrt{\frac{\delta}{\beta-\alpha}}}.
\end{aligned}$$

Finally, by the bound of δ and δ_ϵ in (5.2.55) and (5.2.56), we have

$$\begin{aligned}
\widehat{\theta} &\geq \frac{1}{\left(n+\frac{qnh}{\beta-\alpha}\right)+\sqrt{\frac{h(h+3.5)n^2}{\beta-\alpha}}} \\
&\geq \frac{1}{n\left(1+\frac{hq}{\beta-\alpha}+\sqrt{\frac{h(h+3.5)}{\beta-\alpha}}\right)} = \frac{1}{wn}.
\end{aligned}$$

□

Note that if $q = 0$, then constraint reduction is not performed . In that case,

$$w = 1 + \sqrt{(26 \times 29.5)/(\beta - \alpha)} \leq 1 + (29/\sqrt{\beta - \alpha}),$$

and $\overline{\theta}$ has the lower bound same as the unreduced algorithm by [71, Theorem 3.8].

Lemma 5.2.14. *Define* $\epsilon_k = \max(\mathbf{X}^k \bullet \mathbf{Z}^k, \|\mathbf{r}_p^k\|, \|\mathbf{r}_d^k\|)$. *The algorithm in Section 5.2 converges in* $O(n \ln(\epsilon_0/\epsilon))$ *iterations for a given tolerance* ϵ *where* $\epsilon_0 = \max(n\tau_0, \|\mathbf{r}_p^0\|, \|\mathbf{r}_d^0\|)$.

Proof. By (5.2.26)-(5.2.29), we know

$$\begin{aligned}
\epsilon_k &\leq \max((1 + \alpha/\sqrt{n})n\tau_k, \|\boldsymbol{r}_p^k\|, \|\boldsymbol{r}_d^k\|) \\
&\leq \psi_k \max((1 + \alpha/\sqrt{n})n\tau_0, \|\boldsymbol{r}_p^0\|, \|\boldsymbol{r}_d^0\|) \\
&\leq \psi_k (1 + \alpha/\sqrt{n})n\epsilon_0.
\end{aligned}$$

On the other hand, by the definition of ψ_k and Lemma 5.2.13,

$$\psi_k = \prod_{i=1}^{k}(1 - \overline{\theta}_i) \leq (1 - \frac{1}{wn})^k.$$

So,

$$\epsilon_k \leq \left(1 - \frac{1}{wn}\right)^k (1 + \alpha/\sqrt{n})n\epsilon_0.$$

Thus, if

$$\left(1 - \frac{1}{wn}\right)^K (1 + \alpha/\sqrt{n})n\epsilon_0 \leq \epsilon \tag{5.2.57}$$

after K iterations, then $\epsilon_K \leq \epsilon$. So, we compute the minimum K to satisfy (5.2.57). By taking $\ln$ on both sides,

$$K \ln\left(1 - \frac{1}{wn}\right) + \ln\left[(1 + \alpha/\sqrt{n})n\epsilon_0\right] \leq \ln \epsilon$$

if and only if

$$K \ln\left(1 - \frac{1}{wn}\right) \leq \ln \epsilon - \ln\left[(1+\alpha/\sqrt{n})n\epsilon_0\right]$$
$$= \ln(\epsilon/\epsilon_0) - \ln\left[(1+\alpha/\sqrt{n})n\right] \leq \ln(\epsilon/\epsilon_0)$$

Hence, $\epsilon_K \leq \epsilon$ if

$$K \geq \frac{\ln(\epsilon_0/\epsilon)}{-\ln\left(1 - \frac{1}{wn}\right)}.$$

By the fact

$$\frac{-1}{\ln\left(1 - \frac{1}{wn}\right)} \to wn, \quad \text{as } n \text{ increases},$$

$K = O(n \ln(\epsilon_0/\epsilon))$. □

5.3 Conclusion

We proposed an infeasible *predictor-corrector* interior point method with adaptive constraint reduction for diagonal block structured SDP problems. By proving its global convergence and polynomial complexity $O(n \ln(\epsilon_0/\epsilon))$, we verify that our adaptive criteria guarantee correct selection of inactive constraint blocks.

We finish this chapter with a comment about the super-linear local convergence. Kojima, Shida and Shindoh [50] showed that the *predictor-corrector* algorithm has the super-linear local convergence if the generated sequence converges tangentially to the central path. As noted in [50], the tangential convergence can be achieved by repeating the corrector step of the algorithm by Potra and Sheng [71] until $(\boldsymbol{X}^+, \boldsymbol{Z}^+)$ moves into $\mathcal{N}(g(\tau_k), \tau)$ for a given $g(\tau_k)$ such that $g(\tau_k) \to 0$ as $k \to \infty$. Since our algorithm is based on the one by Potra and Sheng, we expect that a similar modification can be easily adopted for super-linear local convergence.

Chapter 6

Conclusion

In this dissertation, we applied the matrix reduction method to three different optimization problems: total least squares, covariance matrix estimation, and solving semidefinite programming problems. The matrix reduction has different purposes for these problems. In total least squares, we want to eliminate the noise contained in raw data in order to better estimate the parameters in a linear model. In covariance matrix estimation, matrix reduction removes undesirable transient or noisy factors to improve the quality of the estimate. In semidefinite programming, inactive constraints are removed from a working constraint set by matrix reduction when we compute a search direction.

For each problem, we proposed a method to determine the reduction intensity, considering the distinct purpose and assumptions of the particular problem. In total least squares, we studied the asymptotic behavior of the smallest singular values corresponding to the noise. This led us to determine the point of truncation by observing the dispersion of the smallest singular values, measured by the coefficient of variation. From this study, we achieved the following results:

1. We proved convergence properties for the singular values corresponding to error terms as $m \to \infty$.

2. We developed an algorithm for determining the weight for the two terms in the minimization function.

3. We developed an algorithm for determining the rank of the true model matrix.

4. We developed an algorithm to find consistent estimate for the *Errors-in-Variables* problem with weaker assumptions than in previous work.

In covariance matrix estimation, we found an optimal intensity which minimizes the difference between the correlation matrix of the noise and an identity matrix. In this study, we made the following contributions:

1. We developed an algorithm for Tikhonov filtered covariance matrix estimation.

2. We put all previous factor-based covariance estimations into a common framework.

3. We performed empirical experiments using the stock return data from 1958 to 2006.

   (a) In terms of minimizing risks, Tikhonov estimate performs as well as the most competitive estimates so far.

   (b) For not enough historical data, Tikhonov estimate outperforms all the other estimates.

(c) In terms of risk prediction, the risk predicted by Tikhonov estimate is the closest to the realized risk.

In semidefinite programming, we chose the reduced constraint blocks to ensure that iterates remain in the designated neighborhood of a central path. In this study, we obtained the following results:

1. We developed an adaptive constraint-reduced *predictor-corrector* algorithm for SDP.

2. We proved the global convergence of the algorithm.

3. We proved polynomial complexity of the algorithm, which is the first result for such *primal-dual* constraint reduced interior-point-methods.

4. These results also hold when applying the algorithm to LP, QP, QCQP, and SOCP.

Before finishing this dissertation, we suggest the following future studies for the discussed problems. First, the proposed matrix reduction method in total least squares problems is effective only when the number of noise terms is greater than 1, since we cannot measure the dispersion with a single singular value. Thus, an alternative approach is required for problems with a single right hand side and a full rank data matrix. Second, we evaluated the value of using our covariance matrix estimate in the MV portfolio problem. Even though the experiments were performed in many different settings, the evaluation was still restricted to the portfolio problem. In order to extend the applications of our covariance matrix estimate, we could investigate its effectiveness using data sets from a variety of applications. Third, in semidefinite programming, the *predictor-*

corrector algorithm proposed in Chapter 5 computes the Schur complement matrix twice for each iteration. Since most of the practical implementations reuse the Schur complement matrix in the corrector step, the current algorithm is not so practical in this aspect. To make the algorithm more practical, we need to prove the global convergence of an algorithm that reuses the Schur complement matrix, or demonstrate experimental effectiveness of an algorithm that solves the corrector problem using the predictor matrix as a preconditioner. We might also generalize our results to cone programming, perhaps using the work of Schurr et al. [78].

Bibliography

[1] Farid Alizadeh, Jean-Pierre A. Haeberly, and Michael L. Overton. Complementarity and nondegeneracy in semidefinite programming. *Mathematical Programming*, 77(1):111–128, 1997.

[2] Christine Bachoc and Frank Vallentin. New upper bounds for kissing number from semidefinite programming. *Journal of The American Mathematical Society*, 21(3):909–924, 2007.

[3] M. S. Bartlett. Tests of significance in factor analysis. *British Journal of Psychology*, 3(2):77–85, 1950.

[4] C. Bender. Bestimmung der grössten Anzahl gleich Kugeln, welche sich auf eine Kugel von demselben Radius, wie die übrigen, auflegen lassen. *Math. Physik*, 56:302–306, 1974.

[5] Christoffer Bengtsson and Jan Holst. On portfolio selection: Improved covariance matrix estimation for Swedish asset returns. In *31st Meeting, Euro Working Group on Financial Modeling*. Hermes Center, Nov 2002.

[6] Michael J. Best and Robert R. Grauer. On the sensitivity of mean-variance-efficient portfolios to changes in asset means : Some analytical and computational results. *Review of Financial Studies*, 4(2):315–342, 1991.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[8] Richard P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall, 1973.

[9] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[10] Louis K.C. Chan, Jason Karceski, and Josef Lakonishok. On portfolio optimization: Forecasting covariances and choosing the risk model. *Review of Financial Studies*, 12(5):937—974, 1999.

[11] T. F. Chan. An improved algorithm for computing the singular value decomposition. *ACM Transactions on Mathematical Software*, 80:72–78, 1982.

[12] Vijay K. Chopra, Chris R. Hensel, and Andrew L. Turner. Massaging mean-variance inputs: Returns from alternative global investment strategies in the 1980s. *Management Science*, 39(7):845–855, 1993.

[13] Vijay K. Chopra and William T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

[14] T. Conlon, H. J. Ruskin, and M. Crane. Random matrix theory and fund of funds portfolio optimisation. *Physica A*, 382(2):565–576, August 2007. DOI:10.1016/j.physa.2007.04.039.

[15] Gregory Connor and Robert A. Korajczyk. A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4):1263–1291, Sep 1993.

[16] G. Dantzig and Y. Ye. A build-up interior-point method for linear programming: Affine scaling form. Technical report, Stanford University, 1991.

[17] Etienne de Klerk. *Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications*. Kluwer Academic Publisers, 2002.

[18] Etienne de Klerk, Dmitrii V. Pasechnik, and Renata Sotirov. On semidefinite programming relaxations of the traveling salesman problem. *SIAM Journal on Optimization*, 19(4):1559–1573, 2008.

[19] Victor DeMiguel, Lorenzo Garlappi, Francisco J. Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

[20] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy. *Review of Financial Studies*, 22(5):1915–1953, 2009.

[21] Victor DeMiguel and Francisco J. Nogales. Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577, 2009.

[22] D. den Hertog, C. Roos, and T. Terlaky. Adding and deleting constraints in the path-following method for lp. *Advances in Optimization and Approximation (D. Z. Du and J. Sun, eds.), Kluwer Academic Publishers*, pages 166–185, 1994.

[23] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G.W. Stewart. *LINPACK Users' Guide*. SIAM, Philadelphia, PA, 1979.

[24] Edwin J. Elton and Martin J. Gruber. Estimating the dependence structure of share prices – implications for portfolio selection. *Journal of Finance*, 28(5):1203–1232, 1973.

[25] R.D. Fierro, L. Vanhamme, and S. Van Huffel. Total least squares algorithms based on rank-revealing complete orthogonal decompositions. In *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pages 99–116. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.

[26] Katsuki Fujisawa, Masakazu Kojima, and Kazuhide Nakata. Exploiting sparsity in primal-dual interior-point methods for semidefinite programming. *Mathematical Programming*, 79(1):235–253, 1997.

[27] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunder. Methods for modifying matrix factorizations. *Mathematics of Computation*, 28(126):505–535, 1974.

[28] Leon Jay Gleser. Estimation in a multivariate ”errors in variables” regression model: Large sample results. *The Annals of Statistics*, 9(1):24–44, 1981.

[29] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[30] Gene H. Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J., Series B, Numerical Analysis*, 2(2):205–224, 1965.

[31] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, 1996.

[32] Louis Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, 1954.

[33] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992.

[34] Per Christian Hansen, James G. Nagy, and Dianne P. O’Leary. *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM, 2006.

[35] Per Christian Hansen and Dianne P. O’Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993. DOI:10.1137/0914086.

[36] Richard J. Hanson. A numerical method for solving Fredholm integral equations of the first kind using singular values. *SIAM Journal on Numerical Analysis*, 8(3):616–622, 1971.

[37] Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer, 2003.

[38] Christoph Helmberg, Franz Rendl, Robert J. Vanderbei, and Henry Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, 1996.

[39] Iveta Hnětynková, Martin Plešinger, Diana Maria Sima, Zdeněk Strakoš, and Sabine Van Huffel. The total least squares problem in $AX \simeq B$. A new classification with the relationship to the classical works. Technical report, Institute of Computer Science, Academy of Sciences of the Czech Republic, 2010.

[40] A. E. Hoerl and R. W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[41] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[42] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[43] J. Edward Jackson. *A User's Guide to Principal Components*. Wiley-IEEE, 2003.

[44] Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58(4):1651–1684, 2003.

[45] Benjamin Jansen. *Interior Point Techniques in Optimization*. Kluwer Academic Publisers, 1997.

[46] Jin Hyuk Jung, Dianne P. O'Leary, and André L. Tits. Adaptive constraint reduction for training support vector machines. *Electronic Transactions on Numerical Analysis*, 31:156–177, 2008.

[47] Jin Hyuk Jung, Dianne P. O'Leary, and André L. Tits. Adaptive constraint reduction for convex quadratic programming. *Computational Optimization and Applications*, 2010. DOI:10.1007/s10589-010-9324-8.

[48] J. A. Kaliski and Y. Ye. A decomposition variant of the potential reduction algorithm for linear programming. *Management Science*, 39:757–776, 1993.

[49] Benjamin F. King. Market and industry factors in stock price behavior. *Journal of Business*, 39(1):139–190, Jan 1966. Part 2: Supplement on Security Prices.

[50] Masakazu Kojima, Masayuki Shida, and Susumu Shindoh. Local convergence of predictor-corrector infeasible-interior-point algorithms for SDPs and SDLCPs. *Mathematical Programming*, 80(2):129–160, 1998.

[51] Masakazu Kojima, Susumu Shindoh, and Shinji Hara. Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, 7(1):86–125, 1997.

[52] J. Kwapień, S. Drożdż, and P. Oświęcimka. The bulk of the stock market correlation matrix is not pure noise. *Physica A*, 359(1):589–606, January 2006. DOI:10.1016/j.physa.2005.05.090.

[53] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, Aug 1999. DOI:10.1103/PhysRevLett.83.1467.

[54] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–397, 2000. DOI:10.1142/S0219024900000255.

[55] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

[56] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[57] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia, 1998.

[58] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, Feb 1965.

[59] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[60] Madan Lal Mehta. *Random Matrices*. Academic Press, New York, 3rd edition, 2004. Cited by [69].

[61] Renato D.C. Monteiro. Primal-dual path-following algorithms for semidefinite programming. *SIAM Journal on Optimization*, 7(3):663–678, 1996.

[62] Jan Mossin. Equilibrium in a capital asset market. *Econometrica*, 34(4):768–783, Oct 1966.

[63] Christopher C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[64] Christopher C. Paige and Zdeněk Strakoš. Scaled total least squares fundamentals. *Numerische Mathematik*, 91:117–146, 2002.

[65] Sungwoo Park and Dianne P. O'Leary. Implicitly-weighted total least squares. *Linear Algebra and its Applications*, 2010. DOI:10.1016/j.laa.2010.06.020.

[66] Sungwoo Park and Dianne P. O'Leary. Portfolio selection using Tikhonov filtering to estimate the covariance matrix. *SIAM Journal on Financial Mathematics*, 1:932–961, 2010.

[67] David L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery*, 9(1):84–97, 1962.

[68] Vasiliki Plerou, Parameswaran Gopikrishnan, Luís A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical Review E*, 60(6):6519–6529, Dec 1999. DOI:10.1103/PhysRevE.60.6519.

[69] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral, Thomas Guhr, and H. Eugene Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126.1–066126.18, June 2002. DOI:10.1103/PhysRevE.65.066126.

[70] Martin Plešinger. *The Total Least Squares Problem and Reduction of Data in $AX \approx B$*. PhD thesis, Technical University of Liberec, Liberec, Czech Republic, March 2008.

[71] Florian A. Potra and Rongqin Sheng. A superlinearly convergent primal-dual infeasible-interior-point algorithm for semidefinite programming. *SIAM Journal on Optimization*, 8(4):1007–1028, 2006.

[72] Bhaskar D. Rao. Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework. In Sabine Van Huffel, editor, *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pages 11 – 20. SIAM, 1997.

[73] Richard Roll and Stephen A. Ross. An empirical investigation of the arbitrage pricing theory. *Journal of Finance*, 35(5):1073–1103, Dec 1980.

[74] Stephen A. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.

[75] Bert W. Rust. Parameter selection for constrained solutions to ill-posed problems. In *Modeling the Earth's Systems: Physical to Infrastructural*, volume 32, pages 333–347. 32nd Symposium on the Interface, Computing Science and Statistics, Apr 2000.

[76] Bert W. Rust and Dianne P. O'Leary. Residual periodograms for choosing regularization parameters for ill-posed problems. *Inverse Problems*, 24:034005 (30 pages), 2008. DOI:10.1088/0266-5611/24/3/034005.

[77] Alexander Schrijver. A comparison of the Delsarte and Loviász bounds. *IEEE Transactions on Information Theory*, IT-25(4):425–429, 1979.

[78] Simon P. Schurr, Dianne P. O'Leary, and André L. Tits. A polynomial-time interior point method for conic optimization, with inexact barrier evaluations. *SIAM Journal on Optimization*, 20(1):548–571, 2009.

[79] William F. Sharpe. A simplified model for portfolio analysis. *Management Science*, 9(2):277–293, 1963.

[80] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, Sep 1964.

[81] G. W. Stewart. Updating a rank-revealing ULV decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14:494–499, 1993.

[82] G. W. Stewart and Ji guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[83] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-posed Problems*. John Wiley & Sons, 1977.

[84] A. L. Tits, P. A. Absil, and W. Woessner. Constraint reduction for linear programs with many constraints. *SIAM Journal on Optimization*, 17(1):119–146, 2006.

[85] 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. Technical report, Carnegie Mellon University, 2010.

[86] K. Tone. An active-set strategy in an interior point method for linear programming. *Mathematical Programming*, 59(3):345–360, 1993.

[87] Jack L. Treynor. Toward a theory of the market value of risky assets. Technical report, 1961. Unpublished manuscript, Subsequently published as [**?**, Chapter 2].

[88] Charles Trzcinka. On the number of factors in the arbitrage pricing model. *Journal of Finance*, 41(2):347–368, June 1986.

[89] S. Twomey. On the numerical solution of Fredholm integral equations of the first kind by inversion of the linear system produced by quadrature. *Journal of the Association for Computing Machinery*, 10(1):97–101, 1963.

[90] Sabine Van Huffel and Joos Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia, 1991.

[91] W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

[92] Jhacova Ashira Williams. The use of preconditioning for training support vector machines. Master's thesis, University of Maryland, 2008.

[93] Luke B. Winternitz, Stacey O. Nicholls, André L. Tits, and Dianne P. O'Leary. A constraint-reduced variant of Mehrotra's predictor-corrector algorithm. *Computational Optimization and Applications*, 2011.

[94] S. Wold. Cross validatory estimation of the number of components in factor and principal component analysis. *Technometrics*, 20:397–405, 1978.

[95] Y. Ye. An $O(n^3L)$ potential reduction algorithm for linear programming. *Mathematical Programming*, 50(2):239–258, 1991.

[96] Qing Zhao, Stefan E. Karisch, Franz Rendl, and Henry Wolkowicz. Semidefinite programming relaxations for the quadratic assignment problem. *Journal of Combinatorial Optimization*, 2(1):71–109, 1998.