Majorization-Minimization Algorithm and Its Applications in Robust Covariance Matrix Estimation

by

Ying SUN

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

July 2016, Hong Kong

Authorization

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

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

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

_____

Ying SUN

July 2016

Majorization-Minimization Algorithm and Its Applications in Robust Covariance Matrix Estimation

by

Ying SUN

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

Prof. Daniel PALOMAR, ECE (Thesis Supervisor)

Prof. Hai YANG, CIVL (Committee Chairperson)

Prof. Matthew McKAY, ECE (Committee Member)

Prof. Danny Hin-Kwok TSANG, ECE (Committee Member)

Prof. Qi QI, IELM (Committee Member)

Prof. Bertram SHI, ECE (Department Head)

The Department of Electronic and Computer Engineering

July 2016

To My Beloved Parents

Acknowledgements

It has been a long way since the starting point of a PhD life, and it is my great honor to express my gratitude towards people who generously assisted me through this journey.

First and foremost, I would like to thank my supervisor, Prof. Daniel P. Palomar. It is him who led me to the world of research, helped and guided me through the difficulties encountered for the past five years. His enthusiasm, sharp intuition, and persistence on research have been a constant inspiration to me. I am also extremely grateful for his patience and a generous offer of discussion time, during which I enjoyed and learned a lot. It was a great experience to work with Daniel and a privilege to have him as a supervisor.

I would also like to thank Dr. Prabhu Babu, who interacted with me closely on the discussions and encouraged me to carry on when facing difficulties. I have been always impressed by his passion for research.

I was fortunate to have opportunities of working with researchers from different areas. These collaborations significantly increased the diversity of my research.

I appreciate a lot working with Prof. Frédéric Pascal, Prof. Guillaume Ginolhac, and Dr. Arnaud Breloy. The topic of covariance estimation in radar systems has led to fruitful results and is still on-going.

The joint work with Shanpu Shen, Dr. Sichao Song, and Prof. Ross D. Murch on the problem of antenna design is also a valuable experience to me. Trying to apply optimization tools to another discipline has broadened my scope of research greatly.

Thanks certainly go to Prof. Gesualdo Scutari, who introduced me to the area of distributed optimization during my visit to Purdue university. Learning and understanding the theory behind optimization algorithms is a great pleasure. I am deeply grateful for his guidance in our daily discussions and the assistance on paper writing.

I would like to thank Prof. Matthew McKay, Prof. Danny Hin-Kwok Tsang, Prof. Qi Qi, Prof. Wing-Kin Ma, and Prof. Hai Yang for serving as my thesis defense committee

members, and Prof. Bing-Yi Jing for being my thesis proposal committee member. I am grateful for the time they spent on reading my thesis and attending the presentation. Their detailed invaluable comments have greatly strengthened the quality of the thesis.

Thanks to all my dearest friends who accompanied me till the end of my PhD. I cherish the moments spent with our convex group members, including Yang Yang, Yiyong and Wei Huang, Mengyi, Junxiao, Zhongju, Benidis, Tianyu, Linlong, Licheng, Ziping, Junyan, Rena, and of course, Daniel. The memories of group running and gathering, the time we spent on discussing and cracking problems, will always be a treasure in my life. I feel extremely lucky to be a member of such a nice group. Sincere thanks to my UST friends Jingyi, Yinghua, Su Pan, Yu Li, Pancy, Liusha, Yu Cheng, Qian Tian, and Crystal, for the sharing of joy and pain, the taste of food and life. Special thanks to Yangmuzi Zhang, Cheng Chen, Yin Xia, and Tianhang Zhang, friends of mine since undergraduate, for the advice and companions in the US and Spain. Thanks also to my friends Amir, Loris, Wenqi Wang, and Monica in Purdue, for the warm welcome and assistance on daily lives during my visit. In addition, I would like to thank my Lab 3116 colleagues: Bunny, Yueping Wu, Xi Zhang, Jian Yu, Nicolas Auguin, Chang Li, Yaming Luo, BenX, Yuanming Shi, Lu Yang, Yuyi Mao, Xianghao Yu, Yinghao Yu, Shibo Chen, Runfa Zhou, Shuqi Chai, Wanting Zhou, Lixiang Lian, Bin Qian, Baojian Zhou, and Shenghong Li. Having a chance of meeting them in the past five years is great.

Finally, I owe my deepest gratitude to my parents Gui Sun and Puyan Xi, for their unconditional love and support in my life.

Table of Contents

List of Tables

List of Figures

Abstract

Covariance estimation has been a fundamental and long existing problem, closely related to various fields including multi-antenna communication systems, social networks, bioinformatics, and financial engineering. Classical estimators, although simple to construct, have been criticized for their inaccurate estimation when the number of samples is small compared to the variable dimension. In this thesis, we study the problem of improving estimation accuracy by regularizing a covariance matrix based on prior information. Two types of regularization methods are considered, namely, shrinking the raw estimator to a known target and imposing a structural constraint on it.

Our study first focuses on the shrinkage covariance estimators with a zero mean. For a family of estimators defined as the maximizer of a penalized likelihood function, sufficient conditions their existence are established, which quantitatively reveal the number of required samples is reduced for estimation. The condition is then particularized for two particular estimators, where we show that it is also necessary. To compute the two estimators, numerical algorithms are devised leveraging the majorization-minimization (MM) algorithm framework, under which convergence is analyzed systematically. The problem is then extended to the joint estimation of mean and covariance matrix.

For applications where the covariance matrix possesses a certain structure, we propose estimating it by maximizing the data likelihood function under the prior structural constraint. First, estimation with a general convex constraint is introduced, along with an efficient algorithm for computing the estimator derived based on MM. Then, the algorithm is tailored to several special structures that enjoy a wide range of applications in signal processing related fields. In addition, two types of non-convex structures are also discussed. The algorithms are proved to converge to a stationary point of the problems. Numerical results show that the proposed estimators outperform the state of the art methods in the sense of achieving a smaller estimation error at a lower computational cost.

Abbreviations

pdf	Probability Distribution Function.
i.i.d.	Independent and Identically Distributed.
MLE	Maximum Likelihood Estimator.
MVE	Minimum Volume Ellipsoid.
MCD	Minimum Covariance Determinant.
PCA	Principal Component Analysis.
CCA	Canonical Component Analysis.
CES	Complex Elliptically Symmetric.
NMSE	Normalized Mean-Square Error.
MM	Majorization-Minimization.
EM	Expectation Maximization.
SDP	Semidefinite Programming.
SOCP	Second-Order Cone Programming.
SVD	Singular Value Decomposition.
LMI	Linear Matrix Inequality.
KKT	Karush Kuhn Tucker.
MIMO	Multiple-Input Multiple-Output.
DOA	Direction-of-Arrival.
MUSIC	MUltiple Signal Classification.

Notations

a, A	Scalar.
$\mathbf{a}$	Vector.
$\mathbf{A}$	Matrix.
$(\cdot)^T$	Transpose.
$(\cdot)^*$	Conjugate.
$(\cdot)^H$	Conjugate transpose.
$\mathrm{Re}\,(\cdot)$	Real part.
a_i	The i-th entry of $\mathbf{a}$.
$A_{i,j}$	The (i-th, j-th) element of matrix $\mathbf{A}$.
$\mathbf{A}_{i,:}$	The i-th row of matrix $\mathbf{A}$.
$\mathbf{A}_{:,j}$	The j-th column of matrix $\mathbf{A}$.
$\mathbb{R}$, $\mathbb{C}$	The set of real and complex numbers, respectively.
$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$	The set of $m \times n$ matrices with real- and complex-valued entries, respectively.
$\mathbb{S}_+^n$	The set of symmetric (Hermitian) positive definite $n \times n$ matrices.
$\mathrm{diag}\,(\mathbf{a})$	Diagonal matrix with $\mathbf{a}$ as its principal diagonal.
$\mathrm{diag}\,(\mathbf{A})$	Vector consisting of the diagonal elements of $\mathbf{A}$.
$\|\mathbf{a}\|_p$	The ℓ_p norm of $\mathbf{a}$.
$\mathbf{1}$	Vector with all elements being 1.
$\mathbf{I}_n$	Identity matrix of dimension $n \times n$.
$\mathrm{rank}(\mathbf{A})$	Rank of matrix $\mathbf{A}$.
$\mathrm{Tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$.
$\det(\mathbf{A})$	Determinant of matrix $\mathbf{A}$.
$\mathrm{vec}\,(\mathbf{A})$	Vector consisting of all the columns of $\mathbf{A}$ stacked.
$\lambda_{\max}(\mathbf{A})$	Maximum eigenvalue of matrix $\mathbf{A}$.
$\lambda_{\min}(\mathbf{A})$	Minimum eigenvalue of matrix $\mathbf{A}$.

$\mathbf{A}^{1/2}$	Hermitian square root of the positive semidefinite matrix $\mathbf{A}$.
$\mathrm{span}(\{\mathbf{A}_i\})$	Vector space spanned by all the column vectors of the matrices in $\{\mathbf{A}_i\}$
$\mathrm{range}\,(\mathbf{A})$	Range space of matrix $\mathbf{A}$.
$\mathbf{A} \succeq \mathbf{B}$,	Matrix $\mathbf{A} - \mathbf{B}$ is positive semidefintie.
$\mathbf{A} \succ \mathbf{B}$	Matrix $\mathbf{A} - \mathbf{B}$ is positive definite.
$\|\cdot\|_F$	Frobenius norm.
$\mathbb{E}(\cdot)$	Expectation.
$\mathrm{Var}\,(\cdot)$	Variance.
$\mathrm{Cov}\,(\cdot)$	Covariance.
$O(\cdot)$	Big-O notation.
$o(\cdot)$	Little-o notation.
$P(\cdot)$	Probability.
$\mathcal{N}\,(\cdot)$	Gaussian distribution.
$\otimes$	Kronecker product.
$\odot$	Hadamard product.
$f'\,(a)$	Derivative of function $f\,(x)$ evaluated at $x = a$.
$\mathsf{d}\,(\cdot)$	Differential operator.
$\nabla f\,(\mathbf{a})$	Gradient of function $f\,(\mathbf{x})$ evaluated at $\mathbf{x} = \mathbf{a}$.
$\lim$	Limit.
$\sup$	Supremum.
$\inf$	Infimum.
$\log\,(\cdot)$	Natural logarithm.

Chapter 1

Introduction

1.1 Motivation

Estimating the covariance matrix is an important problem as it is a fundamental building block of many applications. Examples include direction of arrival estimation [1], anomaly detection in wireless sensor networks [2], space-time adaptive processing detection problem [3], impulsive noise attenuation in image processing [4], high-resolution frequency estimation [5], and portfolio optimization in finance [6]; see [7] for a general review and the references therein. An intuitive and easy approach is to estimate the covariance matrix by sample average, which coincides with the maximum likelihood estimator (MLE) under the assumption that the samples are independent and identically drawn from a Gaussian distribution. However, in many applications the samples follow a distribution with a heavier tail than the Gaussian distribution, either due to the intrinsic mechanism of the application (e.g., financial time series [8]) or the existence of outliers, such as faulty observations. In this case, the Gaussian assumption can result in an unreliable estimate.

A way to address the aforementioned problem is to find a robust covariance matrix estimator that limits the influence of erroneous observations so as to achieve better performance in non-Gaussian scenarios. Various robust estimators, include M-estimators [9], S-estimators [10], MVE (Minimum Volume Ellipsoid) [11], and MCD (Minimum Covariance Determinant) [12] have been proposed in the literature. See [13, 14] for a complete overview. In the seminal work [15], Tyler proposed an M-estimator that estimates the covariance matrix up to a scaling factor for samples drawn from an elliptical distribution with a known mean. It was proved that under the assumption that the samples are independent and identically drawn

from an elliptical distribution, Tyler's estimator is a minimax robust estimator, and is also "distribution-free". Tyler's estimator coincides with the MLE of the scatter matrix by fitting the samples normalized by their length to the angular central Gaussian distribution [16–18].

However, in modern applications we often encounter the situation that the number of samples N is small compared to their dimension K. For example, the sample covariance matrix is singular when $N < K$ regardless of the population parameter, which is problematic for applications that depend on the inverse of a covariance matrix. Unfortunately, Tyler's estimator has the same drawback as the sample covariance matrix. A popular way of addressing this issue in the literature is to shrink the estimator to some target matrix, an identity matrix or a diagonal matrix for example. Based on this idea, a regularized Tyler's estimator that shrinks the estimate towards an identity matrix was first proposed in [19]. A rigorous proof for the existence, uniqueness, and convergence properties is provided in [2], where a systematic way of choosing the regularization parameter was also proposed. However, unlike the Tyler's estimator which is an MLE, the shrinkage estimator was not derived from a meaningful cost function. To overcome this issue, a new scale-invariant shrinkage Tyler's estimator, defined as a minimizer of a penalized cost function, was recently proposed in [20]. By showing that the objective function is geodesically convex, Wiesel proved that any algorithm that converges to the local minimum of the objective function is actually the global minimum. Numerical algorithms are provided for the estimator and simulation results demonstrate the estimator is robust and effective in the sample deficient scenario. Despite the good properties, it was reported that the numerical algorithm may diverge when the objective function is unbounded below [20]. Therefore, we ask the following questions:

- Under what condition will the minimizer (shrinkage estimator) be well-defined?
- If the shrinkage estimator exists (algorithm converges), will it be unique as the Tyler's estimator?

Tyler's estimator assumes a known mean. In practice, there are scenarios in which both mean and covariance are required to be estimated from the samples, from fundamental techniques such as data decorrelation and principal component analysis [21], to high-level applications such as portfolio optimization in financial engineering [22]. A two-step estimation can be done by first subtracting an estimated mean (e.g. sample mean or sample median) from the samples, then feed the centered samples to covariance estimator. However, estimating the

mean without taking into account the correlation of each component may be inadequate, especially in the presence of outliers. For example, as mentioned in [7], a multivariate outlier may have none of its components outlying by just looking at a particular coordinate, and being treated as a normal sample consequently. Moreover, the estimation error of the mean propagates to the covariance estimation since the true mean is substituted by an erroneous estimate. For the above-mentioned reason, a robust estimator that estimates the mean and covariance jointly is desired. A general class of M-estimates, which includes the well-known Huber's estimator and the MLE of the Student t distribution, were considered in [9]. The existence and uniqueness of the estimator were established under the condition that the number of samples is sufficiently large. This motivates us to consider the following question:

- How to generalize shrinkage covariance estimators to the joint mean-covariance estimation problem?

More recently, the fact that covariance matrix in some applications naturally possesses some special structures has received increasing attention. Exploiting the structure information in the estimation process usually implies a reduction in the number of parameters to be estimated, and thus is beneficial to improving the estimation accuracy [23]. Various types of structures have been studied. For example, the Toeplitz structure with applications in time series analysis and array signal processing was considered in [23–25]. A sparse graphical model was studied in [26], where sparsity was imposed on the inverse of the covariance matrix. Banding or tapering the sample covariance matrix was proposed in [27]. A spiked covariance structure, which is closely related to the problem of component analysis and subspace estimation, was introduced in [1]. Other structures such as group symmetry and the Kronecker structure were considered in [28–30]. While the previously mentioned works have shown that imposing a prior structure on the covariance estimator improves its performance in many applications, most of them either assume that the samples follow a Gaussian distribution or attempt to regularize the sample covariance matrix. In the robust covariance estimation context, we attempt to provide an answer to the following questions:

- How to incorporate structural information into robust covariance estimator to obtain a better estimate?

- How to numerically compute the structured estimator efficiently?

1.2 Outline and Contribution

The problem of improving the performance of robust estimators is considered from two aspects in the thesis. One of them is via shrinkage when the number of samples is small compared to the problem dimension, and the other one is via imposing structures on the estimator.

In the first scenario, we consider shrinkage estimators that can be defined as the minimizer of a penalized cost function. Assuming a given mean vector, the new estimator shrinks the original one towards a prior target matrix. The contribution is threefold. First, a sufficient condition for the existence of a class of shrinkage estimators is provided and particularized for Wiesel's estimator proposed in [20], whose existence condition was unknown. We also show that the condition is sufficient in this case. Second, we study another shrinkage Tyler's estimator that has a simpler form, where the necessary and sufficient condition of its existence is also provided. Furthermore, we show that the two shrinkage estimators are equivalent, and they are unique (up to a scaling factor). Third, numerical algorithms to compute the estimators are provided with convergence guarantee.

In the next step, we assume the mean vector is unknown, and the number of samples is small. To avoid the shortcoming of a two-step estimation mentioned before, we propose an estimator that estimates the mean and covariance matrix jointly and performs shrinkage at the same time. Sufficient condition on the existence and uniqueness of the proposed estimator is given. Numerical algorithms are also provided and compared in terms of convergence speed.

In the second scenario, we consider improving the estimation accuracy of Tyler's estimator by forcing the estimator to be structured. This is done by formulating the estimator as the solution of a constrained *non-convex* optimization problem. Computing the solution using off-the-shelf solvers can be computationally demanding, or even infeasible for some types of non-convex constraint set. Leveraging the idea of the MM algorithm, we devise an algorithm for obtaining the estimator when the structural constraint set is convex by iteratively solving a convex problem. The algorithm is then tailored to several special structures, namely, the sum of rank-one matrices, Toeplitz, and banded Toeplitz structures, with improved efficiency. In addition, algorithms are also provided for the spiked structure and Kronecker structure, which are non-convex.

The dissertation is organized as follows. In Chapter 2, we introduce the MM algorithm framework that serves as an optimization tool throughout this dissertation. In Chapter 3, preliminaries on robust covariance estimators are introduced as a basic building block for the

development of theory and algorithms in the dissertation. Chapter 4 discusses the shrinkage robust estimators with a given mean vector in the small sample regime. The problem is extended to joint mean-covariance estimation in Chapter 5. In Chapter 6 we consider the problem of imposing structural constraints on Tyler's estimator. Chapter 7 concludes the main results.

1.3 Publications

Journal Papers

1. Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143-5156, Oct. 2014.

2. Y. Sun, P. Babu, and D. P. Palomar, "Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3096-3109, Jun. 2015.

3. Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3576-3590, Jul., 2016

4. Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in radar systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1986-1998, Apr. 2016

5. K. Benidis, Y. Sun, P. Babu, D. P. Palomar, "Orthogonal Sparse PCA and Covariance Estimation via Procrustes Reformulation", submitted to *IEEE Transactions on Signal Processing*, Jan. 2016.

6. Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithm and Its Applications in Signal Processing, Communications, and Machine Learning," submitted to *IEEE Transactions on Signal Processing*, Jan. 2016.

7. S. Shen, Y. Sun, S. Song, D. P. Palomar, and R. D. Murch, "Successive Boolean optimization of planar pixel antennas," submitted to *IEEE Transactions on Antennas and*

Propagation, May 2016.

Conference Papers

1. Y. Sun, P. Babu, and D. P. Palomar, "Regularized Robust Estimation of Mean and Covariance Matrix under Heavy Tails and Outliers," in *Proc. IEEE 8th Sensor Array and Multichannel Signal Process. Workshop (SAM)*, A Coruña, Spain, June 2014, pp. 125-128.

2. Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed distributions," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 19-24 April 2015, pp. 5693-5697.

3. K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, "Orthogonal Sparse Eigenvectors: A Procrustes Problem," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 20-25 March, 2016.

4. A. Breloy, Y. Sun, P. Babu, D.P. Palomar, "Low-Complexity Algorithms for Low Rank Clutter Parameters Estimation in Radar Systems", accepted in *EUSIPCO*, Budapest, Hungary, 29 Aug. - 2 Sept. 2016.

5. A. Breloy, Y. Sun, P. Babu, G. Ginolhac, D.P. Palomar, F. Pascal, "A robust signal subspace estimator", accepted in *IEEE Workshop on Statistical Signal Processing (SSP)*, Palma de Mallorca, Spain, 26-29 Jun. 2016.

Chapter 2

The Majorization-Minimization Algorithm

Advances in data analysis, machine learning, communications, and signal processing often lead to a complex modeling with a massive amount of data to be processed and a large number of parameters to be inferred. This results in the associated optimization problems getting more complicated than before. While those that appear in textbooks are small-scale examples that may be convex or even have nice closed-form solution, real-world problems we face with are more challenging to handle due to the following difficulties. First of all, some of the problems are high-dimensional. Apart from the ones that are trivially parallelizable, or convex where decomposition techniques can be employed, the general case requires a prohibitively large amount of computational resources (time and storage) or it is even impossible to call off-the-self solvers. Second, the problem may take a complicated form. For example, in some applications, the objective function can be highly non-convex which can cause numerical issues. Nastier ones can even have discontinuous objective functions or non-convex constraint sets, where classical nonlinear programming algorithms cannot be applied. Any one of these obstacles could rule out the possibility of employing a general-purpose solver and expecting it to return a reliable solution. It leaves one no option but to devise optimization algorithms tailored to each of the specific problems that can take advantage of the problem structure. This is where the majorization-minimization (MM) algorithm comes into play.

The idea of the MM algorithm is straightforward to grasp. It states that instead of minimizing the original objective function f, which may take a complicated form, one can minimize an upperbound (surrogate function), $g\left(\cdot|x_t\right)$, of it at the t-th iteration. The surrogate

function is required to touch f (up to a global constant) at $x = x_t$. Consequently, the objective value is forced to decrease. The procedure is shown pictorially in Figure 2.1. A parallel argument can be made for maximization problems by replacing the upperbound minimization step by a lowerbound maximization step. The scheme is referred to as minorization-maximization.

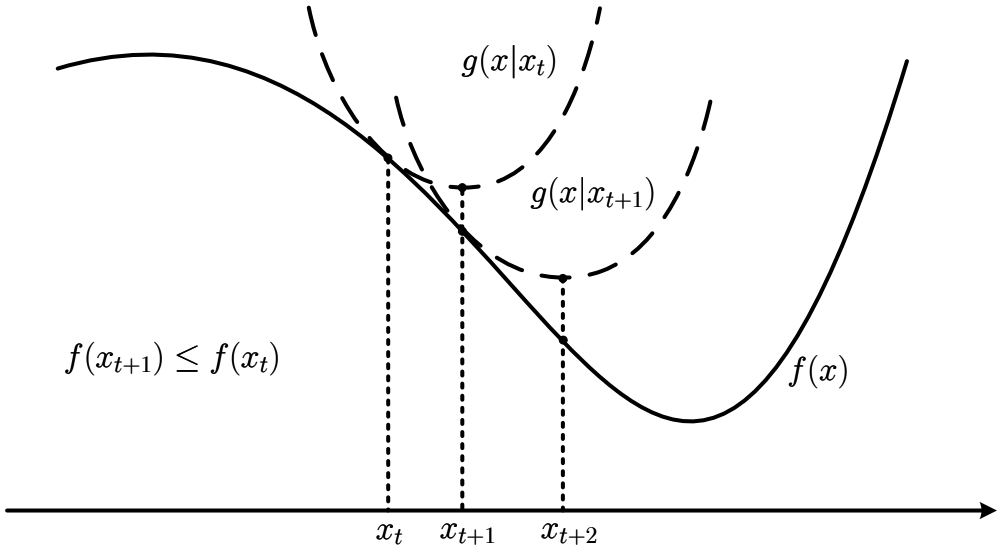


Figure 2.1: The MM procedure.

Although it has received considerable attention recently, MM has a long history that dates back to the 1970s [31], and is closely related to the famous expectation maximization (EM) algorithm [32]. Indeed, EM is a special case of MM applied mainly in maximum likelihood (ML) estimation problems with incomplete-data. The algorithm was systematically introduced in the seminal work [33] by Dempster, Laird, and Rudin in 1977. MM generalizes EM by replacing the E-step, which calculates the conditional expectation of log-likelihood of the complete data set, by a minorization step that finds a surrogate function to locally approximate the objective function. The surrogate function keeps the key property of the E-step by being a global lower bound of the objective. As a consequence, most of the convergence results of EM hold naturally for MM. Compared to EM, MM has a much larger scope of applications.

Despite the simplicity of the algorithm description, the MM framework can be used to design algorithms that conquer the above-mentioned challenges for applications with a complex problem formulation. In the past decades, MM has been applied in fundamental methodologies such as sparse regression analysis to handle non-convex or discontinuous objective functions, as well as high-dimensionality [34–37]; in sparse principal component analysis (PCA) [38] to handle a cardinality constraint; in canonical component analysis

(CCA) [39,40], covariance estimation [20,41–43], and matrix factorization [44,45] to handle non-convex objective functions and constraints. It is also applied to higher level applications such as image processing [46–48], phase retrieval [49], and sequence design [50, 51], where high problem dimensionality is a major source of difficulty.

The success of MM lies in the construction of the surrogate function. Generally speaking, a "good" surrogate function often bears some of the following properties [52]:

- separability in variables (parallel computing);
- convex and smooth functions;
- with a closed-form minimizer, even if the constraint set is complicated.

Consequently, minimizing the surrogate function is efficient and scalable. In the end, a complicated problem is solved by sequentially solving a series of much simpler sub-problems. Nevertheless, a surrogate function that tries to follow the shape of the original objective is also desirable to achieve a fast convergence rate, which often implies it should be sufficiently complicated. These two opposite goals result in finding an appropriate surrogate function a crucial step in designing an efficient MM algorithm, and is achieved only by a thorough analysis of each optimization problem and trial and error.

2.1 Algorithm Framework

2.1.1 The MM Algorithm and Its Convergence

Consider the following generic optimization problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{2.1.1}$$

where $\mathcal{X}$ is a nonempty closed set in $\mathbb{R}^n$ and $f : \mathcal{X} \to \mathbb{R}$ is a continuous function. In addition, we assume that a minimizer of f exists in $\mathcal{X}$ [1].

With an initial point $\mathbf{x}_0 \in \mathcal{X}$, MM generates a sequence of feasible points $(\mathbf{x}_t)_{t\in\mathbb{N}}$ by the following induction. Given $\mathbf{x}_t$, the algorithm requires a surrogate function $g(\cdot|\mathbf{x}_t) : \mathcal{X} \to \mathbb{R}$

[1] The existence of a minimizer can be guaranteed if $\lim_{\substack{\mathbf{x}\in\mathcal{X} \\ \|\mathbf{x}\|\to+\infty}} f(\mathbf{x}) = +\infty$.

satisfying the majorization property:

$$g\left(\mathbf{x}|\mathbf{x}_t\right) \geq f\left(\mathbf{x}\right) + c_t,\ \forall \mathbf{x} \in \mathcal{X}, \tag{2.1.2}$$

where $c_t = g\left(\mathbf{x}_t|\mathbf{x}_t\right) - f\left(\mathbf{x}_t\right)$. That is, the difference of $g\left(\cdot|\mathbf{x}_t\right)$ and f is minimized at $\mathbf{x}_t$.

A subsequent point $\mathbf{x}_{t+1}$ is then generated as

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}} g\left(\mathbf{x}|\mathbf{x}_t\right). \tag{2.1.3}$$

The sequence $\left(f\left(\mathbf{x}_t\right)\right)_{t\in\mathbb{N}}$ is non-increasing since

$$f\left(\mathbf{x}_{t+1}\right) \leq g\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right) - c_t \leq g\left(\mathbf{x}_t|\mathbf{x}_t\right) - c_t = f\left(\mathbf{x}_t\right), \tag{2.1.4}$$

where the first inequality follows from the majorization step (2.1.2), and the second inequality follows from the minimization step (2.1.3). We denote the algorithm mapping defined by steps (2.1.2) and (2.1.3) that sends $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ by $M : \mathbb{R}^n \to \mathbb{R}^n$ in the rest of this section.

We have already shown that the sequence $\left(f\left(\mathbf{x}_t\right)\right)_{t\in\mathbb{N}}$ is monotonically decreasing. Assuming that a minimizer of f exists in $\mathcal{X}$, $\left(f\left(\mathbf{x}_t\right)\right)_{t\in\mathbb{N}}$ is bounded below hence has a limit $f^\star$. The next question is how is $f^\star$ and $\left(\mathbf{x}_t\right)_{t\in\mathbb{N}}$ related to the minimizers of problem (2.1.1).

Without convexity, obtaining the global minimizers of problem (2.1.1) is out of reach in general. Therefore, one usually seek a larger set composed of all stationary points satisfying first order necessary conditions for optimality.

Assuming that the constraint set $\mathcal{X}$ is convex, and directional derivate of f at $\mathbf{x}_t$ along direction $\mathbf{d}$ defined as

$$f'\left(\mathbf{x}_t;\mathbf{d}\right) \triangleq \liminf_{\lambda\downarrow 0} \frac{f\left(\mathbf{x}_t + \lambda\mathbf{d}\right) - f\left(\mathbf{x}_t\right)}{\lambda} \tag{2.1.5}$$

exists, the set of stationary points is given by

$$\mathcal{X}^\star \triangleq \left\{\mathbf{x}|f'\left(\mathbf{x};\mathbf{d}\right) \geq 0,\ \forall \mathbf{x}+\mathbf{d} \in \mathcal{X}\right\}. \tag{2.1.6}$$

The convergence of MM is given in the next theorem.

Theorem 2.1 [53]**.** *Assuming that $g\left(\mathbf{x}|\mathbf{x}_t\right)$ is continuous in both $\mathbf{x}$ and $\mathbf{x}_t$, and $f'\left(\mathbf{x}_t;\mathbf{d}\right) = g'\left(\mathbf{x}_t;\mathbf{d}|\mathbf{x}_t\right),\ \forall \mathbf{x}_t+\mathbf{d} \in \mathcal{X}$, then any limit point of the sequence $\left(\mathbf{x}_t\right)_{t\in\mathbb{N}}$ is a stationary point*

of problem (2.1.1)*. Moreover, if the initial level set, defined as* $\mathcal{X}^0 \triangleq \{\mathbf{x}|f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$*, is compact, then*

$$\lim_{t\to+\infty} d(\mathbf{x}_t, \mathcal{X}^\star) \triangleq \lim_{t\to+\infty} \inf_{\mathbf{x}\in\mathcal{X}^\star} \|\mathbf{x}_t - \mathbf{x}\|_2 = 0.$$

Theorem 2.1 states that under mild assumptions on the surrogate function, the distance of $\mathbf{x}_t$ and $\mathcal{X}^\star$ converges to zero, and $f^\star$ a stationary value, meaning $f^\star = f(\mathbf{x}^\star)$ with $\mathbf{x}^\star \in \mathcal{X}^\star$.

2.1.2 The Cyclic Block MM Algorithm and Its Convergence

The idea of majorizing f by a surrogate function can also be applied blockwise. Specifically, $\mathbf{x}$ is partitioned into m blocks as $\mathbf{x} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)$, where each n_i-dimensional block $\mathbf{x}^{(i)} \in \mathcal{X}_i$ and $\mathcal{X} = \prod_{i=1}^m \mathcal{X}_i$. At the $(t+1)$-th iteration, $\mathbf{x}^{(i)}$ is updated by solving the following problem:

$$\begin{aligned} \underset{\mathbf{x}^{(i)}}{\text{minimize}} \quad & g_i\left(\mathbf{x}^{(i)}|\mathbf{x}_t\right) \\ \text{subject to} \quad & \mathbf{x}^{(i)} \in \mathcal{X}_i, \end{aligned} \tag{2.1.7}$$

where $i = (t \bmod m) + 1$. Inheriting from the MM algorithm, the surrogate function $g_i\left(\mathbf{x}^{(i)}|\mathbf{x}_t\right)$ is also required to satisfy the tightness and global upperbound assumptions, i.e.,
(B1) $f(\mathbf{x}_t) = g_i\left(\mathbf{x}_t^{(i)}|\mathbf{x}_t\right), \forall\, \mathbf{x}_t \in \mathcal{X}, \forall\, i$;
(B2) $f\left(\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}_t^{(m)}\right) \leq g_i\left(\mathbf{x}^{(i)}|\mathbf{x}_t\right), \forall\, \mathbf{x}^{(i)} \in \mathcal{X}_i, \forall \mathbf{x}_t \in \mathcal{X}, \forall\, i$.

In short, at each iteration, the block MM applies the ordinary MM algorithm to one block while keeping the value of the other blocks fixed. The blocks are updated in cyclic order.

The block MM algorithm generalizes the MM algorithm and provides more flexibility in designing surrogate functions. Moreover, in some cases the surrogate function can approximate f better than treating the variables as a single block, thus result in a faster convergence [54].

It is also worth mentioning that the block MM algorithm generalizes the block coordinate descent (BCD) algorithm [55], where at the $(t+1)$-th iteration, $\mathbf{x}^{(i)}$ is updated by solving the problem

$$\begin{aligned} \underset{\mathbf{x}^{(i)}}{\text{minimize}} \quad & f\left(\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}_t^{(m)}\right) \\ \text{subject to} \quad & \mathbf{x}^{(i)} \in \mathcal{X}_i. \end{aligned} \tag{2.1.8}$$

Notice that by letting $g_i\left(\mathbf{x}_t^{(i)}|\mathbf{x}_t\right)$ equal to $f\left(\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}_t^{(m)}\right)$, block MM algorithm

reduces to BCD algorithm.

Similar to the argument of MM, it can be derived from assumptions (B1) and (B2) that the sequence $\left(f\left(\mathbf{x}_t\right)\right)_{t\in\mathbb{N}}$ generated by the block MM algorithm is nonincreasing.

Assuming in addition that

(B3) f is continuous and $\mathcal{X}$ is convex;

(B4) $g_i\left(\mathbf{x}_t^{(i)}|\mathbf{x}_t\right)$ is continuous in both $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_t$, $\forall i$;

(B5) $f'\left(\mathbf{x}_t;\mathbf{d}_i^0\right)=g_i'\left(\mathbf{x}_t^{(i)};\mathbf{d}_i|\mathbf{x}_t\right),\ \forall\,\mathbf{x}_t^{(i)}+\mathbf{d}_i\in\mathcal{X}_i,\ \mathbf{d}_i^0\triangleq\left(\mathbf{0};\ldots;\mathbf{d}_i;\ldots;\mathbf{0}\right),\ \forall\, i$,

the authors of [53] proved the following convergence result.

Theorem 2.2 [53]**.** *A limit point* $\mathbf{x}^\star$ *of* $\{\mathbf{x}_t\}$ *is a stationary point of* (2.1.1) *in either of the following two cases:*

(1) if f *is regular at* $\mathbf{x}^\star$*,* $g_i\left(\mathbf{x}_t^{(i)}|\mathbf{x}_t\right)$ *is quasi-convex in* $\mathbf{x}_t^{(i)}$*, and problem* (2.1.7) *has a unique solution for any* $\mathbf{x}_t\in\mathcal{X}$*;*

(2) if f *is regular at* $\mathbf{x}^\star$ *with respect to the coordinates* $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$*, the level set* $\mathcal{X}^0=\{\mathbf{x}|f\left(\mathbf{x}\right)\leq f\left(\mathbf{x}_0\right)\}$ *is compact, and problem* (2.1.7) *has a unique solution any* $\mathbf{x}_t\in\mathcal{X}$ *for at least* $(n-1)$ *blocks.*

Compared to Theorem 2.1 for single block MM, Theorem 2.2 imposes more restrictive assumptions on the surrogate function by either requiring it to be quasi-convex or to have a unique minimizer.

2.1.3 Extensions

MM can be combined with many other algorithm frameworks, leading to various extension, as briefly described below.

Instead of finding a minimizer of $g\left(\cdot|\mathbf{x}_t\right)$ at each iteration, which may be computationally intractable, we can find a point that satisfies $g\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)\leq g\left(\mathbf{x}_t|\mathbf{x}_t\right)$ instead (*i.e.*, just making an improvement). This leads to the generalized EM (GEM) algorithm [33] in the context of ML estimation. The subsequent point $\mathbf{x}_{t+1}$ can be found by taking a gradient, Newton, or quasi-Newton step, which is closely related to MM acceleration schemes [56–58].

In block MM, the cyclic update ordering can be generalized to the "essential cyclic rule" [59] in the sense that each block is updated at least once within a given number of iterations [53, 60, 61]. Other sweeping schemes include the Gauss-Southwell update rule, maximum improvement update rule, as well as randomized update rule [53].

An incremental MM was proposed in [62] for minimizing an objective function of the form $f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$, which is closely related to stochastic optimization with f being the empirical average. The algorithm assumes only one of the f_i's is observed at each iteration, and the surrogate of f is updated based on f_i and the history of the algorithm recursively.

In [60], MM has been made more flexible by weakening the global upperbound condition of the surrogate function to a local upperbound.

2.2 Acceleration Schemes

While the idea of MM appears simple, it sometimes suffers from a slow convergence speed [32, 52] when the surrogate function approximates f poorly. To alleviate this shortcoming, MM accelerators are often employed so that the algorithm can have a fast convergence rate. Various types of accelerators have been proposed in the literature, including those derived based on multivariate Aitken's method [63], conjugate gradient acceleration [56], Newton and quasi-Newton type acceleration [58, 64, 65], over-relaxation [66–69]. In this section, we briefly outline the acceleration schemes. An overview in the context of EM can be found in Chapter 4 of [70].

We begin with the general idea of line search type nonlinear programming algorithms. To minimize a function f, at the current point $\mathbf{x}_t$ one first determine a descent direction $\mathbf{d}_t$, then a step-size α_t that decreases the objective function. MM can be interpreted in this way by identifying $\mathbf{d}_t = M(\mathbf{x}_t) - \mathbf{x}_t$ and $\alpha_t = 1$.

Line search type accelerators use a different α_t other than one while fixing the direction $\mathbf{d}_t$. For instance, α_t in [71] was determined by the two previous steps based on Aitken's method. However, changing α_t in this way may destroy the monotonicity of the algorithm. A constant step-size $\alpha_t \equiv \alpha$ was adopted in over-relaxation methods [66–69], and the optimal α was provided in [69]. Nevertheless, computing α is generally a difficult problem. To address these issues, α_t was suggested to be computed in an inexact line search fashion so that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ is guaranteed [69, 72, 73].

Another class of accelerators also change the descent direction $\mathbf{d}_t$. To ensure the objective value is nonincreasing, $\mathbf{x}_{t+1}$ need not to be a global minimizer of $g(\cdot|\mathbf{x}_t)$. Instead, one can solve (2.1.3) inexactly by taking a Newton step. This leads to the EM gradient algorithm [57]. A quasi-Newton accelerator proposed in [58] improves it by adding an approximate of the

Hessian of $H(\mathbf{x}|\mathbf{x}_t) \triangleq f(\mathbf{x}) - g(\mathbf{x}|\mathbf{x}_t)$ to $\nabla g^2(\mathbf{x}|\mathbf{x}_t)$ in the Newton step (assuming both $\nabla^2 H(\mathbf{x}|\mathbf{x}_t)$ and $\nabla g^2(\mathbf{x}|\mathbf{x}_t)$ exist). In [56], the generalized gradient algorithm was applied to minimize f by treating $M(\mathbf{x}_t) - \mathbf{x}_t$ as the generalized gradient. An overview and comparison of above-mentioned ones can be found in [72].

Finally we introduce a class of accelerators based on finding a fixed point of M, which is a stationary point of f if the regularity conditions (A1), (A2), and (A3.1) hold. Assuming that M is continuously differentiable, it is known that the Newton's method provides quadratic convergence in the vicinity of a fixed point. Define $F(\mathbf{x}) = M(\mathbf{x}) - \mathbf{x}$, a Newton step of finding a zero of F is given by[2]

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - (\nabla F(\tilde{\mathbf{x}}_t))^{-1} F(\tilde{\mathbf{x}}_t),$$

where ∇F is the Jacobian of F. While $F(\tilde{\mathbf{x}}_t)$ can be evaluated by the MM step, the Jacobian $\nabla F(\tilde{\mathbf{x}}_t)$ is hard to obtain in general (unless $M(\mathbf{x})$ has an explicit form) and need to be approximated based on the previous iterates $(\tilde{\mathbf{x}}_{t'})_{0 \le t' \le t}$. The STEM accelerator proposed in [74] approximates $\nabla F(\tilde{\mathbf{x}}_t)$ by a scaled identity matrix. Aitken [63] and SQUAREM accelerator [74] approximate $\nabla F(\tilde{\mathbf{x}}_t)$ by the secant method. More recently, an accelerator was proposed in [75] based on quasi-Newton method.

We point out that Newton type algorithms for nonlinear programming converge only in the vicinity of a stationary point, therefore accelerators based on Newton's iteration are often executed after a few MM steps so that $\mathbf{x}_t$ falls into the convergence region. It is also worth mentioning that the MM acceleration schemes are developed for unconstrained optimization problems (except for the cases that the constraint can be eliminated by reparameterization). For a constrained optimization problem, it is generally not true that the point returned by accelerators will be feasible.

In practice, both the convergence rate and the computational cost per iteration should be taken into account in choosing an appropriate accelerator.

2.3 Inequalities and Surrogate Function Construction

Most of the existing techniques to construct surrogate functions can be categorized as convexity-based methods and special-inequality-based methods [32,52]. In this section, we review these

[2]The sequence $(\tilde{\mathbf{x}}_t)_{t \in \mathbb{N}}$ should be distinguished from the MM sequence $(\mathbf{x}_t)_{t \in \mathbb{N}}$.

techniques and provide the upperbounds (and lowerbounds for maximization problems) for a variety of "basic functions" that frequently appear.

2.3.1 First Order Taylor Expansion

Suppose f can be decomposed as

$$f(\mathbf{x}) = f_0(\mathbf{x}) + f_{\text{ccv}}(\mathbf{x}), \tag{2.3.1}$$

where f_{ccv} is a differentiable concave function.

As f_{ccv} is concave, linearizing it leads to the following inequality:

$$f_{\text{ccv}}(\mathbf{x}) \leq f_{\text{ccv}}(\mathbf{x}_t) + \nabla f_{\text{ccv}}(\mathbf{x}_t)^T (\mathbf{x} - \mathbf{x}_t). \tag{2.3.2}$$

Therefore, an upperbound of f can be easily constructed as

$$f(\mathbf{x}) \leq f_0(\mathbf{x}) + \nabla f_{\text{ccv}}(\mathbf{x}_t)^T \mathbf{x} + \text{const.}$$

Example 2.1. The function $\log(x)$ can be upperbounded as

$$\log(x) \leq \log(x_t) + \frac{1}{x_t}(x - x_t) \tag{2.3.3}$$

with equality achieved at $x = x_t$.

Example 2.2. The function $\log\det(\mathbf{\Sigma})$ can be upperbounded as

$$\log\det(\mathbf{\Sigma}) \leq \log\det(\mathbf{\Sigma}_t) + \text{Tr}\left(\mathbf{\Sigma}_t^{-1}(\mathbf{\Sigma} - \mathbf{\Sigma}_t)\right) \tag{2.3.4}$$

with equality achieved at $\mathbf{\Sigma} = \mathbf{\Sigma}_t$.

Example 2.3. The function $\text{Tr}(\mathbf{S}\mathbf{X}^{-1})$ with both $\mathbf{S}$ and $\mathbf{X}$ in $\mathbb{S}_{++}$ can be lowerbounded as

$$\text{Tr}\left(\mathbf{S}\mathbf{X}^{-1}\right) \geq \text{Tr}\left(\mathbf{S}\mathbf{X}_t^{-1}\right) - \text{Tr}\left(\mathbf{X}_t^{-1}\mathbf{S}\mathbf{X}_t^{-1}(\mathbf{X} - \mathbf{X}_t)\right) \tag{2.3.5}$$

with equality achieved at $\mathbf{X} = \mathbf{X}_t$.

Example 2.4 [76]**.** The function $\mathrm{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right)$ with $\mathbf{Y} \in \mathbb{S}_{++}$ can be lowerbounded as

$$\mathrm{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right) \geq 2\mathrm{Tr}\left(\mathbf{X}_t^T\mathbf{Y}_t^{-1}\mathbf{X}\right) - \mathrm{Tr}\left(\mathbf{Y}_t^{-1}\mathbf{X}_t\mathbf{X}_t^T\mathbf{Y}_t^{-1}\mathbf{Y}\right) + \mathrm{const.} \tag{2.3.6}$$

with equality achieved at $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_t, \mathbf{Y}_t)$.

Proof. The function $\mathrm{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right)$ is jointly convex in $\mathbf{X}$ and $\mathbf{Y}$, and therefore lowerbounded by its linear expansion around $(\mathbf{X}_t, \mathbf{Y}_t)$, which implies (2.3.6). □

Remark 2.1. It is worth emphasizing that although the upperbounds in the previously listed examples are constructed by linearization, it can generate a rich family of upperbounds that are not necessarily linear in the variables.

More generally, given a convex, a linear, and a concave function, f_{cvx}, f_{lin}, and f_{ccv}, respectively, if their values and gradients are equal at some $\mathbf{x}_t$, then, for any $\mathbf{x}$,

$$f_{\mathrm{ccv}}\left(\mathbf{x}\right) \leq f_{\mathrm{lin}}\left(\mathbf{x}\right) \leq f_{\mathrm{cvx}}\left(\mathbf{x}\right), \tag{2.3.7}$$

as illustrated in Figure 2.2.

The relation (2.3.7) states that a concave function can be upperbounded by a convex function. The linear bound (2.3.2) is a special case and is also the tightest convex upperbound.

Example 2.5. The concave function $|x|^p\,,\ 0 < p \leq 1$, can be upperbounded as[3]

$$|x|^p \leq \frac{p}{2}\left|x_t\right|^{p-2} x^2 + \mathrm{const}, \tag{2.3.8}$$

providing that $x_t \neq 0$.

Inequality (2.3.8) plays an important role in the iteratively reweighted least squares (IRLS) algorithm that applied in regression analysis [77], sparse representation and regularization problems [40], and sensor network localization problem [78]. In all these applications, a quadratic upperbound is preferred to a linear one in the majorization step, even though a linear one would be tighter, so that the minimization step admits a solution that is easy to compute.

In the last example, we show that inequality (2.3.7) can be used to construct lowerbounds for maximization problems.

[3]The result also holds for $1 < p \leq 2$ although $|x|^p$ is convex.

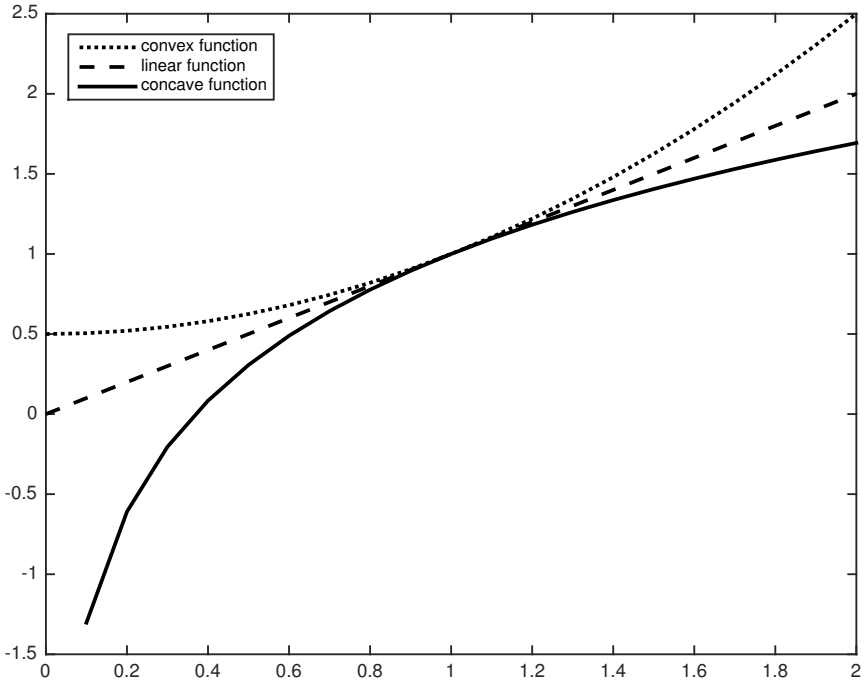


Figure 2.2: Surrogate function construction technique by first order Taylor expansion: a concave function can upperbound a linear function, which can be upperbounded by a convex function.

Example 2.6 [79]**.** A monomial $\prod_{i=1}^{n} x_i^{\alpha_i}$, where $x_i \geq 0$, $\forall i$, can be lowerbounded as

$$\prod_{i=1}^{n} x_i^{\alpha_i} \geq \prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i} \left(1 + \sum_{i=1}^{n} \alpha_i \log x_i - \sum_{i=1}^{n} \alpha_i \log x_i^t\right) \tag{2.3.9}$$

with equality achieved at $x_i = x_i^t$.

Proof. Inequality (2.3.3) implies that

$$\begin{aligned}\log\left(\prod_{i=1}^{n} x_i^{\alpha_i}\right) \leq & \log\left(\prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i}\right) \\ & + \left(\prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i}\right)^{-1} \left(\prod_{i=1}^{n} x_i^{\alpha_i} - \prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i}\right).\end{aligned}$$

Rearranging the terms we have (2.3.9). □

The advantage of applying inequality (2.3.9) is that the surrogate function is now separable in the variables, which can be optimized individually if the constraints are also separable.

2.3.2 Convexity Inequality

From the definition of convexity we have the following inequality:

$$f_{\text{cvx}}\left(\sum_{i=1}^{n} w_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} w_i f_{\text{cvx}}\left(\mathbf{x}_i\right), \tag{2.3.10}$$

where $\sum_{i=1}^{n} w_i = 1, w_i \geq 0. \forall i = 1, \ldots, n$. Equality is achieved when all the $\mathbf{x}_i$'s are equal.

Example 2.7 (Jensen's Inequality)**.** Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function and $\mathbf{x}$ be a random variable that take values in $\mathcal{X}$. Assuming that $\mathbb{E}(\mathbf{x})$ and $\mathbb{E}(f(\mathbf{x}))$ are finite, then

$$\mathbb{E}(f(\mathbf{x})) \geq f(\mathbb{E}(\mathbf{x})).$$

Jensen's inequality plays a key role in showing EM is a particular case of MM.

Particularizing (2.3.10) for the concave function $\log$, we have the following inequality.

Example 2.8. The function $\sum_{i=1}^{n} \alpha_i \log f_i(x)$ with $\alpha_i > 0$ can be upperbounded as

$$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \log f_i(x_t) + \left(\sum_{i=1}^{n} \alpha_i\right) \log\left(\frac{\sum_{i=1}^{n} \alpha_i \frac{f_i(x)}{f_i(x_t)}}{\sum_{i=1}^{n} \alpha_i}\right), \tag{2.3.11}$$

where $f_i(x) > 0, \forall i$. Equality is achieved at $x = x_t$.

Inequality (2.3.11) creates an upperbound for $\sum_{i=1}^{n} \alpha_i \log f_i(x)$ that merges the summation inside the $\log$ function.

Recall that by applying inequality (2.3.3) (linearizing the log terms) we can obtain an alternative upperbound that is linear in the $f_i(x)$'s as

$$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \left(\log f_i(x_t) + \frac{1}{f_i(x_t)}\Big(f_i(x) - f_i(x_t)\Big)\right). \tag{2.3.12}$$

However, (2.3.11) is much tighter and is therefore preferred to (2.3.12) for a faster convergence rate, provided that the surrogate function is easy to minimize. Figure 2.3 compares the two upperbounds (2.3.11) and (2.3.12) for function $\sum_{i=1}^{n} \alpha_i \log(1 + c_i x)$ as an illustration.

Particularizing inequality (2.3.10) for convex functions we have the following bounds.

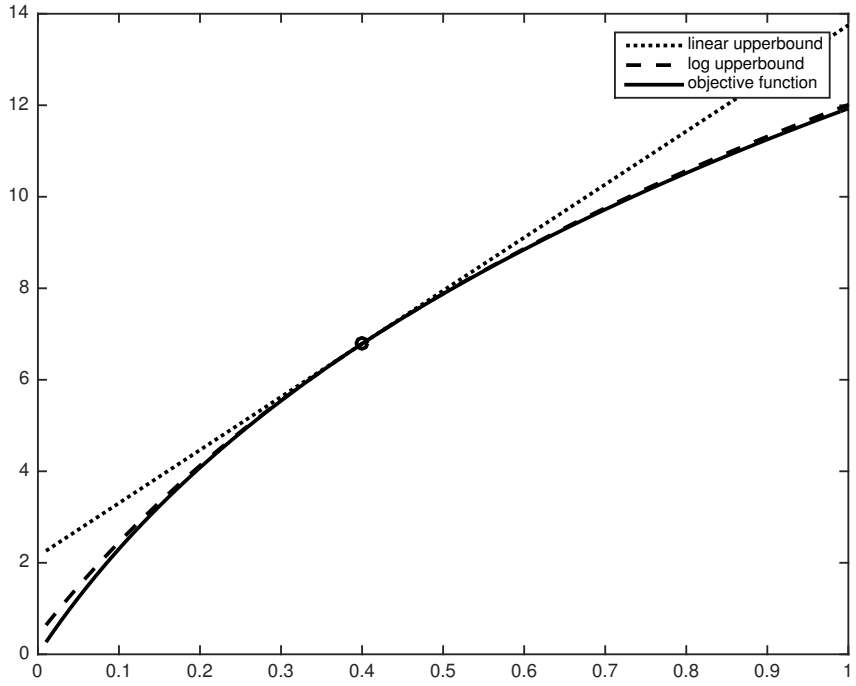


Figure 2.3: Objective function: $f(x) = 3\log(1+x) + 5\log(1+3x) + 1.5\log(1+6x)$; log upperbound: upperbound given by (2.3.11); linear upperbound: upperbound given by (2.3.12).

Example 2.9. The function $\frac{1}{\sum_{i=1}^n a_i x_i}$ with $a_i > 0$ and $x_i > 0$ can be upperbounded as

$$\frac{1}{\sum_{i=1}^n a_i x_i} \leq \frac{\sum_{i=1}^n a_i \left(x_i^t\right)^2 x_i^{-1}}{\left(\sum_{i=1}^n a_i x_i^t\right)^2} \tag{2.3.13}$$

with equality achieved at $x_i = x_i^t$, $\forall i = 1, \ldots, n$.

Generalizing (2.3.13) to an arbitrary convex function f yields the following inequality.

Example 2.10 [52]**.** The convex function $f\left(\mathbf{a}^T\mathbf{x}\right)$ can be upperbounded as

$$f\left(\mathbf{a}^T\mathbf{x}\right) \leq \sum_{i=1}^n \alpha_i f\left(\frac{a_i}{\alpha_i}\left(x_i - x_i^t\right) + \mathbf{a}^T\mathbf{x}_t\right), \tag{2.3.14}$$

where $\alpha_i > 0$, $\sum_{i=1}^n \alpha_i = 1$. Moreover, if the elements of $\mathbf{a}$ and $\mathbf{x}_t$ are positive, letting $\alpha_i = \frac{a_i x_i^t}{\mathbf{a}^T\mathbf{x}_t}$ leads to a different upperbound as

$$f\left(\mathbf{a}^T\mathbf{x}\right) \leq \sum_{i=1}^n \frac{a_i x_i^t}{\mathbf{a}^T\mathbf{x}_t} f\left(\frac{\mathbf{a}^T\mathbf{x}_t}{x_i^t} x_i\right). \tag{2.3.15}$$

Inequalities (2.3.14) and (2.3.15) were proposed and applied in medical imaging in [80, 81].

2.3.3 Construction by Second Order Taylor Expansion

Lemma 2.1 (Descent Lemma [55])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with a Lipschitz continuous gradient and Lipschitz constant L (we say that ∇f is L-Lipschitz henceforth). Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 . \tag{2.3.16}$$

The descent lemma implies that if a function has a Lipschitz continuous gradient, then there exists a *separable* quadratic upperbound for it.

More generally, if a function f has bounded curvature, i.e., there exists a matrix $\mathbf{M}$ such that $\mathbf{M} \succeq \nabla^2 f(\mathbf{x})$, $\forall\, \mathbf{x} \in \mathcal{X}$, then the following inequality implied by the Taylor's theorem [76] holds:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{1}{2} (\mathbf{x} - \mathbf{y})^T \mathbf{M} (\mathbf{x} - \mathbf{y}) . \tag{2.3.17}$$

Particularizing (2.3.17) for $f(\mathbf{x}) = \mathbf{x}^H \mathbf{L} \mathbf{x}$ gives the following inequality.

Example 2.11. The quadratic form $\mathbf{x}^H \mathbf{L} \mathbf{x}$, where $\mathbf{L}$ is a Hermitian matrix, can be upper-bounded as

$$\mathbf{x}^H \mathbf{L} \mathbf{x} \leq \mathbf{x}^H \mathbf{M} \mathbf{x} + 2\mathrm{Re}\left(\mathbf{x}^H (\mathbf{L} - \mathbf{M}) \mathbf{x}_t\right) + \mathbf{x}_t^H (\mathbf{M} - \mathbf{L}) \mathbf{x}_t, \tag{2.3.18}$$

where $\mathbf{M} \succeq \mathbf{L}$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$[4].

Example 2.11 shows that by inequality (2.3.18), we can replace the matrix $\mathbf{L}$ by $\mathbf{M}$ that possesses some desired properties. In particular, by setting $\mathbf{M}$ as a diagonal matrix the surrogate function will become separable in the x_i's.

[4] Wirtinger calculus is applied for complex-valued matrix differentials [82].

2.3.4 Arithmetic-Geometric Mean Inequality

The arithmetic-geometric mean inequality states that [76]

$$\prod_{i=1}^{n} z_i^{\alpha_i} \leq \sum_{i=1}^{n} \frac{\alpha_i}{\|\boldsymbol{\alpha}\|_1} z_i^{\|\boldsymbol{\alpha}\|_1}, \tag{2.3.19}$$

where z_i and α_i are nonnegative numbers. Equality is achieved when the z_i's are equal.

Inequality (2.3.19) has been applied in the literature to construct upperbounds in geometric and signomial programming. For example, letting $z_i = x_i/x_i^t$ for $\alpha_i > 0$ and $z_i = x_i^t/x_i$ for $\alpha_i < 0$ we have the following inequality.

Example 2.12 [79]**.** A monomial $\prod_{i=1}^{n} x_i^{\alpha_i}$ can be upperbounded as

$$\prod_{i=1}^{n} x_i^{\alpha_i} \leq \left(\prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i}\right) \sum_{i=1}^{n} \frac{|\alpha_i|}{\|\boldsymbol{\alpha}\|_1} \left(\frac{x_i}{x_i^t}\right)^{\|\boldsymbol{\alpha}\|_1 \mathrm{sgn}(\alpha_i)}. \tag{2.3.20}$$

Equality is achieved at $x_i = x_i^t, \forall i$.

Together with the lowerbound (2.3.9) for a monomial, these two inequalities serve as the basic ingredients for deriving an MM algorithm for signomial programming [79].

Example 2.13 [83]**.** A posynomial $\sum_{i=1}^{n} u_i(\mathbf{x})$, where $u_i(\mathbf{x})$ is a monomial, can be lower bounded as

$$\sum_{i=1}^{n} u_i(\mathbf{x}) \geq \prod_{i=1}^{n} \left(\frac{u_i(\mathbf{x})}{\alpha_i}\right)^{\alpha_i}, \tag{2.3.21}$$

where $\alpha_i = \frac{u_i(\mathbf{x}_t)}{\prod_{i=1}^{n} u_i(\mathbf{x}_t)}$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

Inequality (2.3.21) lowerbounds a posynomial by a monomial, which can be used in solving complementary geometric programming (GP) with the objective function being the ratio of posynomials.

Example 2.14. The ℓ_2-norm $\|\mathbf{x}\|_2$ can be upperbounded as

$$\|\mathbf{x}\|_2 \leq \frac{1}{2}\left(\|\mathbf{x}_t\|_2 + \|\mathbf{x}\|_2^2 / \|\mathbf{x}_t\|_2\right) \tag{2.3.22}$$

given that $\|\mathbf{x}_t\|_2 \neq 0$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

2.3.5 Cauchy-Schwartz Inequality

Cauchy-Schwartz inequality states that

$$\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 .$$

Equality is achieved when $\mathbf{x}$ and $\mathbf{y}$ are collinear.

Example 2.15. The function $\left|\mathbf{a}^H\mathbf{x}\right|$ can be lowerbounded as

$$\left|\mathbf{a}^H\mathbf{x}\right| \geq \mathrm{Re}\left(\mathbf{x}_t^H\mathbf{a}\mathbf{a}^H\mathbf{x}\right) / \left|\mathbf{a}^H\mathbf{x}_t\right| \tag{2.3.23}$$

with equality achieved at $\mathbf{x} = \mathbf{x}_t$.

Proof. For two complex numbers $z_1 = u_1 + iv_1$ and $z_2 = u_2 + iv_2$, we have

$$\begin{aligned}\mathrm{Re}\left(z_1 z_2^*\right) &= u_1u_2 + v_1v_2 \\ &\leq \sqrt{u_1^2 + v_1^2} \cdot \sqrt{u_2^2 + v_2^2}\end{aligned}$$

by Cauchy-Schwartz inequality. Letting $z_1 = \mathbf{a}^H\mathbf{x}$ and $z_2 = \mathbf{a}^H\mathbf{x}_t$ yields the desired inequality. □

Example 2.16. The function $\|\mathbf{x}\|_2$ can be lowerbounded as

$$\|\mathbf{x}\|_2 \geq \mathbf{x}^T\mathbf{x}_t / \|\mathbf{x}_t\|_2 \tag{2.3.24}$$

given that $\|\mathbf{x}_t\|_2 \neq 0$. Equality achieved at $\mathbf{x} = \mathbf{x}_t$.

Together with inequality (2.3.24), these two inequalities provide a quadratic upperbound and a linear lowerbounds for the ℓ_2-norm on the whole space except the origin.

2.3.6 Schur Complement

The Schur complement condition for a positive definite matrix states that if $\mathbf{C} \succ \mathbf{0}$, then

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \succeq \mathbf{0}$$

if and only if the Schur complement of $\mathbf{C}$,

$$\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T \succeq \mathbf{0}. \tag{2.3.25}$$

Inequality (2.3.25) provides a way to upperbound the inverse of a matrix.

Example 2.17 [43]**.** Assuming $\mathbf{P} \succ \mathbf{0}$, the matrix $\left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}$ can be upperbounded as

$$\mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1} \succeq \left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}, \tag{2.3.26}$$

where $\mathbf{R}_t = \mathbf{A}\mathbf{P}_t\mathbf{A}^H$. Equality is achieved at $\mathbf{P} = \mathbf{P}_t$.

Inequality (2.3.26) can also be derived based on convexity inequality, see [84].

Particularizing (2.3.26) for $\mathbf{P} = \operatorname{diag}(p_1, \ldots, p_n)$ and $\mathbf{A} = \left[\sqrt{a_1}, \ldots, \sqrt{a_n}\right]$ we get

$$\frac{1}{\sum_{i=1}^n a_i p_i} \leq \frac{\sum_{i=1}^n a_i \left(p_i^t\right)^2 p_i^{-1}}{\left(\sum_{i=1}^n a_i p_i^t\right)^2},$$

which is the same as (2.3.13) but based on a different derivation.

2.3.7 Generalization

Based on the inequalities that upperbound simple functions provided before, one can generate a surrogate function for more complicated objective functions by majorizing f more than once. Specifically, one can find a sequence of functions $g^{(1)}(\cdot|\mathbf{x}_t), \ldots, g^{(k)}(\cdot|\mathbf{x}_t)$ satisfying

$$\begin{aligned} &g^{(i)}(\mathbf{x}_t|\mathbf{x}_t) = g^{(i+1)}(\mathbf{x}_t|\mathbf{x}_t) \\ &g^{(i)}(\mathbf{x}|\mathbf{x}_t) \leq g^{(i+1)}(\mathbf{x}|\mathbf{x}_t), \, \forall \mathbf{x} \in \mathcal{X}, \, \forall i = 1, \ldots, k-1. \end{aligned}$$

It can be seen easily that $g^{(k)}(\cdot|\mathbf{x}_t)$ majorizes f. The surrogate function $g^{(i)}(\cdot|\mathbf{x}_t)$ usually gets simpler and simpler until it satisfies some desired properties such as:

- the inner minimization problem is convex;
- the inner minimization problem has a closed-form solution, especially when the constraint set is nonconvex;
- the inner problem can be solved in parallel.

2.4 Conclusion

In this section, we have introduced the preliminaries on the MM algorithm. From a theoretical perspective, we have introduced the general algorithm framework, its convergence conditions, as well as acceleration schemes. Practical techniques of surrogate function construction have also been provided. It serves as one of the building blocks of for the algorithm design throughout the thesis. In particular, we highlight Examples 2.1, 2.2, and 2.17, which are instrumental in devising MM algorithms for covariance estimation problems considered in the rest sections.

Chapter 3

Preliminaries on Robust Covariance Estimation

In Chapter 4, 5, and 6[1], we will consider the problem of estimating the covariance matrix with a number N of K-dimensional *i.i.d.* samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from an elliptical distribution with pdf of the form

$$f(\mathbf{x}) = \det(\mathbf{R}_0)^{-\frac{1}{2}} g\left((\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{R}_0^{-1} (\mathbf{x}-\boldsymbol{\mu}_0)\right) \tag{3.0.1}$$

with location and scatter parameter $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ in $\mathbb{R}^K \times \mathbb{S}_{++}^K$. The nonnegative function $g : \mathbb{R}_+ \to \mathbb{R}_+$, which is called the density generator, determines the shape of the pdf. For most of the popularly used distributions, e.g., the Gaussian and the Student's *t*-distribution, g is a decreasing function and determines the decay of the tails of the distribution.

3.1 Robust Covariance Matrix Estimation

Given $\boldsymbol{\mu}_0$, our problem of interest is to estimate the covariance matrix. Since one can always center the pdf by defining $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}_0$, without loss of generality in the rest of the thesis we will assume $\boldsymbol{\mu}_0 = \mathbf{0}$ when $\boldsymbol{\mu}_0$ is given. We use the notation P_N and f for the empirical and the sampling distributions, respectively.

It is known that the covariance matrix of an elliptical distribution takes the form $c_g \mathbf{R}_0$ with c_g being a constant that depends on g [13], hence it is unlikely to have a good covariance

[1] In Chapter 6, the model is generalized to complex-valued random variables.

estimator without prior knowledge of g. Instead of trying to find the parametric form of g and get an estimator of $c_g\mathbf{R}_0$, we are interested in estimating the normalized covariance matrix $\frac{\mathbf{R}_0}{\mathrm{Tr}(\mathbf{R}_0)}$.

The commonly adopted sample covariance matrix, which also happens to be the maximum likelihood estimator for the normal distribution, estimates $c_g\mathbf{R}_0$ asymptotically. However, it is sensitive to outliers. This motivates the research of estimators robust to outliers and, in fact, many researchers in the statistics literature have addressed this problem by proposing various robust covariance estimators like M-estimators [9], S-estimators [10], MVE [11], and MCD [12] to name a few, see [13, 14] for a complete overview.

As an example, in the work [9], Maronna analyzed the properties of the M-estimators, which are given as the solution $\mathbf{R}$ to the equation

$$\mathbf{R} = \frac{1}{N}\sum_{i=1}^{N} u\left(\mathbf{x}_i^T\mathbf{R}^{-1}\mathbf{x}_i\right)\mathbf{x}_i\mathbf{x}_i^T \tag{3.1.1}$$

where the choice of function $u(\cdot)$ determines a whole family of different estimators. Under some technical conditions on $u(s)$ (i.e., $u(s) \geq 0$ for $s > 0$ and nonincreasing, and $su(s)$ is strictly increasing), Maronna proved that there exists a unique $\mathbf{R}$ that solves (3.1.1), and gave an iterative algorithm to arrive at that solution. He also established its consistency and robustness. A number of well known estimators take the form (3.1.1) and in [9] Maronna gave two examples, with one being the MLE for multivariate Student's t-distribution, and the other being the Huber's estimator [85]. Both of them are popular for handling heavy tails and outliers in the data.

For all the robust covariance estimators, there is a tradeoff between their efficiency, which measures the variance (estimation accuracy) of the estimator, and robustness, which quantifies the sensitivity of the estimator to outliers. As these two quantities are opposed in nature, a considerable effort has to be put in designing estimators that achieve the right balance between these two quantities. In [15], Tyler dealt with this problem by proposing an estimator that is distribution-free, with its variance independent of the parametric form of the underlying distribution. In addition, it is also considered as the "most robust" estimator, in the sense that its maximum asymptotic variance is less than the maximum asymptotic variance of any other consistent and uniformly asymptotically normal estimator.

Tyler's estimator of $\mathbf{R}$ is given as the solution of the fixed-point equation (4.1.1), where

the results of [9] cannot be applied since $su(s) = K$ is not strictly increasing. Tyler established the conditions for the existence of a solution to (4.1.1), as well as the fact that the estimator is unique up to a positive scale factor, in the sense that $\mathbf{R}$ solves (4.1.1) if and only if $c\mathbf{R}$ solves (4.1.1) for some positive scalar c. The estimator was shown to be consistent and asymptotically normal with its asymptotic standard deviation independent of g [15].

Tyler's fixed-point equation (4.1.1) can be alternatively interpreted as follows. Consider the normalized samples defined as $\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, it is known that the pdf of $\mathbf{s}$ takes the form [16–18]

$$f(\mathbf{s}) = \frac{\Gamma\left(\frac{K}{2}\right)}{2\pi^{K/2}} \det(\mathbf{R})^{-\frac{1}{2}} \left(\mathbf{s}^T\mathbf{R}^{-1}\mathbf{s}\right)^{-K/2}. \tag{3.1.2}$$

Given N samples from the normalized distribution $\{\mathbf{s}_i\}_{i=1}^N$, the MLE of $\mathbf{R}$ can be obtained by minimizing the negative log-likelihood function

$$L(\mathbf{R}) = \frac{N}{2}\log\det(\mathbf{R}) + \sum_{i=1}^{N}\frac{K}{2}\log\left(\mathbf{s}_i^T\mathbf{R}^{-1}\mathbf{s}_i\right) \tag{3.1.3}$$

which is equivalent to minimizing

$$L^{\text{Tyler}}(\mathbf{R}) = \frac{N}{2}\log\det(\mathbf{R}) + \sum_{i=1}^{N}\frac{K}{2}\log\left(\mathbf{x}_i^T\mathbf{R}^{-1}\mathbf{x}_i\right). \tag{3.1.4}$$

If a minimizer $\hat{\mathbf{R}} \succ \mathbf{0}$ of the function $L^{\text{Tyler}}(\mathbf{R})$ exists, it needs to satisfy the first order optimality condition $\nabla L^{\text{Tyler}}\left(\hat{\mathbf{R}}\right) = \mathbf{0}$, which can be equivalently written as

$$\mathbf{R} = \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{R}^{-1}\mathbf{x}_i}. \tag{3.1.5}$$

In [15, 16], the authors provided the following conditions for existence of a nonsingular solution of (3.1.5): (i) no $\mathbf{x}_i$ lies on the origin, and (ii) for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{\dim(S)}{K}$, where $P_N(S) \triangleq \frac{\sum_{i=1}^N 1_{\{\mathbf{x}_i \in S\}}}{N}$ stands for the proportion of samples in S. The second condition states that the samples should be sufficiently spread out in the whole $\mathbb{R}^K$ space. Intuitively, it avoids the possibility of using a singular covariance matrix to capture the variation of the samples if they are concentrated in some proper subspace. Assuming that the samples has a continuous pdf, condition (ii) simplifies to $N > K$.

To arrive at the estimator satisfying (3.1.5), Tyler proposed the following iterative algorithm:

$$\begin{aligned}\tilde{\mathbf{R}}_{t+1} &= \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{R}_t^{-1}\mathbf{x}_i}\\ \mathbf{R}_{t+1} &= \frac{\tilde{\mathbf{R}}_{t+1}}{\mathrm{Tr}\left(\tilde{\mathbf{R}}_{t+1}\right)}\end{aligned} \tag{3.1.6}$$

that converges to the unique (up to a positive scale factor) solution of (3.1.5).

The robustness of Tyler's estimator to outliers with a large magnitude can be understood intuitively as follows: by normalizing the samples, i.e., $\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, the estimator is not sensitive to the magnitude of samples, only their direction can affect the performance.

In Chapter 4, we consider adapting a family of M-estimators for high-dimensional estimation problem. For notation convenience, we shall use $\boldsymbol{\Sigma}_0$ instead of $\mathbf{R}_0$ therein.

3.2 Robust Estimation of Mean and Covariance Matrix

The M-estimator (3.1.1) described in the previous section assumes a given $\boldsymbol{\mu}_0$, which can be substituted by an estimate $\hat{\boldsymbol{\mu}}$. This leads to a two-step estimation procedure, where the shape of the distribution cannot be taken into account when estimating $\boldsymbol{\mu}_0$.

To address this problem, the estimating equation (3.1.1) was be extended to jointly estimating $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$ as [9]

$$\frac{1}{N}\sum_{i=1}^{N} u_1\left(\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right) = 0 \tag{3.2.1}$$

$$\frac{1}{N}\sum_{i=1}^{N} u_2\left(\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right)\mathbf{x}_i\mathbf{x}_i^T = \mathbf{R}. \tag{3.2.2}$$

Among them a widely adopted one is the MLE of the Student's t-distribution [86, 87]. This is defined as the minimizer of the negative log-likelihood function of a t-distribution with degree of freedom ν, which takes the following form:

$$L^{\nu}\left(\boldsymbol{\mu},\mathbf{R}\right) = \frac{N}{2}\log\det\left(\mathbf{R}\right) + \frac{K+\nu}{2}\sum_{i=1}^{N}\log\left(\nu+\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right).$$

By setting the gradient of $L^{\nu}\left(\boldsymbol{\mu},\mathbf{R}\right)$ to zero, it can be derived that the estimator satisfies the following system of fixed-point equations:

$$\begin{aligned}\frac{\nu+K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i-\boldsymbol{\mu}}{\nu+\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)}&=\mathbf{0}\\ \frac{\nu+K}{N}\sum_{i=1}^{N}\frac{\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T}{\nu+\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)}&=\mathbf{R},\end{aligned} \tag{3.2.3}$$

which is a special case of (3.2.2).

The solution of (3.2.3) can be interpreted as a weighted sample average, for which the weight decreases as the sample gets farther away from the center, i.e., the weight is inversely proportional to $d_i^2=\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)$. This property indicates that the estimator is less sensitive to outliers than the sample average. The degree of down-weighting increases as ν decreases. The properties of the MLE of the Student's *t*-distribution are well-studied in the literature [9, 13, 88, 89]. Under the condition that for any hyperplane H with $0\leq\dim\left(H\right)\leq K-1$, $P_N\left(H\right)<\frac{\dim(H)+\nu}{K+\nu}$, the solution $\left(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}}\right)$ to (3.2.3) exists, and it is unique when $\nu\geq 1$ [88].

In Chapter 5, we are going to adapt the *M*-estimator defined as the solution of (3.2.3) with $\nu=1$, which corresponds to the MLE of a Cauchy distribution, to high-dimensional estimation problem. For completeness, we first introduce some of its properties that serve as the basis of analysis in the high dimension regime.

Following the idea of [88] and [90], we construct the augmented samples $\bar{\mathbf{x}}_i=\left[\mathbf{x}_i;1\right]\in\mathbb{R}^{K+1}$ and define the following variable:

$$\mathbf{R}=\begin{bmatrix}\mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu}\\ \boldsymbol{\mu}^T & 1\end{bmatrix}. \tag{3.2.4}$$

The negative log-likelihood function of the Cauchy distribution $L^1\left(\boldsymbol{\mu},\mathbf{R}\right)$ ($L\left(\boldsymbol{\mu},\mathbf{R}\right)$ hereafter) can be equivalently expressed as:

$$L\left(\boldsymbol{\mu},\mathbf{R}\right)=L\left(\mathbf{R}\right)=\frac{N}{2}\log\det\left(\mathbf{R}\right)+\frac{K+1}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\mathbf{R}^{-1}\bar{\mathbf{x}}_i\right),$$

which can be virtually viewed as fitting $\bar{\mathbf{s}}_i=\frac{\bar{\mathbf{x}}_i}{\|\bar{\mathbf{x}}_i\|_2}$ to a $(K+1)$-dimensional angular central Gaussian distribution (3.1.2) with the center being zero.

It is known that if $\left(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}}\right)$ satisfies estimating equation (3.2.3) with $\{\mathbf{x}_i\}$ being *i.i.d.* samples from an elliptical distribution, then as $N\to+\infty$, the asymptotic values $(\boldsymbol{\mu}_\infty,\mathbf{R}_\infty)$

will converge in probability to the unique solution of the system of equations

$$\begin{aligned}(K+1)\,E_f\left\{\frac{\mathbf{x}-\boldsymbol{\mu}}{1+(\mathbf{x}-\boldsymbol{\mu})^T\,\mathbf{R}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})}\right\}&=\mathbf{0}\\(K+1)\,E_f\left\{\frac{(\mathbf{x}-\boldsymbol{\mu})\,(\mathbf{x}-\boldsymbol{\mu})^T}{1+(\mathbf{x}-\boldsymbol{\mu})^T\,\mathbf{R}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})}\right\}&=\mathbf{R}\end{aligned} \tag{3.2.5}$$

with the relation $\boldsymbol{\mu}_\infty = \boldsymbol{\mu}_0$ and $\mathbf{R}_\infty = c\mathbf{R}_0$ for some scalar c that depends on f [13].

From the above result, we conclude that if the mean of the underlying distribution f exists, we can estimate it by $\hat{\boldsymbol{\mu}}$. However, the covariance matrix can only be estimated up to a scaling factor that depends on f, which is unknown. For this reason, we focus on the joint estimation of the mean and covariance matrix normalized by its trace, i.e., $\boldsymbol{\mu}_0$ and $\mathbf{R}_0/\mathrm{Tr}\,(\mathbf{R}_0)$.

Chapter 4

Regularized Tyler's Scatter Estimator: Existence, Uniqueness, and Algorithms

The problem of estimating the covariance matrix in the scenario when the number of samples is insufficient has been studied intensively assuming normally distributed samples, see [91–95] for examples. One approach is to perform linear shrinkage by combining the SCM with a target matrix $\mathbf{T}$, which can be an identity matrix or some knowledge-aided prior. However, when the underlying distribution is heavy-tailed, or the samples are contaminated by outliers, SCM is often criticized to perform poorly. As introduced in Chapter 3, M-estimators alleviate this problem and improves the estimation accuracy by down-weighting the contribution of samples with a large magnitude. In this Chapter, we discuss the problem of robustly estimating the covariance matrix with insufficient samples by merging the idea of shrinkage and robust estimation.

4.1 Introduction

In this chapter, we study the problem of adapting the Tyler's estimator for covariance matrix [15] to the regime where the number of samples is small compared to their dimension. Recall that given a number of N samples $\{\mathbf{x}\}_{i=1}^{N}$ in $\mathbb{R}^K$, Tyler's estimator is defined as the solution of the fixed-point equation

$$\mathbf{\Sigma} = \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}. \tag{4.1.1}$$

It is proved that the existence of the estimator requires $N > K$, assuming that any set of K samples are linearly independent. If this preliminary assumption is violated, the fixed point iteration algorithm implied by (4.1.1) will fail to converge. It means that the estimator is not applicable to scenarios where $\mathbf{x}_i$ lies in a high-dimensional space, while the number of available samples is scarce.

In order to solve this problem, a regularized Tyler's estimator was proposed in [2] that shrinks the estimate towards an identity matrix. A rigorous proof for the existence, uniqueness of the estimator, as well as the convergence of the algorithm was provided in [2], where a systematic way of choosing the regularization parameter was also proposed. However, it was pointed out in [20] that the estimator was not derived from a meaningful cost function. To address this issue, a scale-invariant shrinkage Tyler's estimator, defined as a minimizer of a penalized cost function, was recently proposed in [20]. By showing that the objective function is geodesically convex, Wiesel proved that any algorithm that converges to the local minimum of the objective function is actually the global minimum. Numerical algorithms are provided for the estimator and simulation results demonstrate the estimator is robust and effective in the low sample support scenario. Despite the good properties, the existence and the uniqueness of the estimator remains unclear. As a consequence, the algorithm derived to obtain the estimator sometimes fails to converge.

In this chapter, we study the family of shrinkage Tyler's estimators that can be defined as the minimizer of a penalized cost function. A sufficient condition for the existence of shrinkage estimators with a general cost function is provided. Particularizing the result for shrinkage Tyler's estimator proposed in [20] (Wiesel's estimator), we obtain that under the condition $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ and the shrinkage target matrix being positive definite, the estimator exists, where α_0 controls the amount of penalty added to the cost function, $P_N(S)$ stands for the proportion of samples contained in a proper subspace S. Assuming that the samples are drawn from a continuous distribution, the condition simplifies to $N > \frac{K}{1+\alpha_0}$. Compared to the existence condition for Tyler's estimator, which is $P_N(S) < \frac{\dim(S)}{K}$, or $N > K$ under continuity assumption, this result clearly demonstrates that regularization can relax the requirement on the number of samples, hence shows its capability of handling high dimensional estimation problems. In addition, we propose another shrinkage Tyler's estimator defined as the minimizer of Kullback-Leibler divergence (KL divergence) penalized cost

function. We show that the estimator exists under the same condition as Wiesel's estimator. Moreover, we prove that for both Wiesel's and KL shrinkage estimators, the existence condition is not only sufficient, but also necessary if $\alpha_0 > 0$. Finally, we show that the two estimators are actually equivalent. It is worth mentioning that the KL shrinkage estimator coincides with the estimator proposed in [3], where the same condition $N > \frac{K}{1+\alpha_0}$ is also independently derived with the target being a scaled identity matrix, assuming that the samples are complex-valued and linearly independent. To numerically compute the estimators, algorithms are derived based on MM, where the convergence can be analyzed systematically.

This chapter is organized as follows: In Section 3.1, we briefly review Tyler's estimator for samples drawn from the elliptical family. In Section 4.2, the two types of shrinkage estimators, i.e., one proposed in [20] and another derived based on KL divergence are considered, and a rigorous proof for the existence and uniqueness of the estimators is provided. Algorithms based on majorization-minimization are presented in Section 4.3. Numerical examples follow in Section 4.4, and we conclude in Section 4.5.

Note

The boundary of the open set $\mathbb{S}_{++}^K$ is conventionally defined as $\mathbb{S}_{+}^K \backslash \mathbb{S}_{++}^K$, which contains all rank deficient matrices in $\mathbb{S}_{+}^K$. For notation simplicity, we also include matrices with all eigenvalues $\lambda \to +\infty$ into the boundary of $\mathbb{S}_{++}^K$. Therefore a sequence of matrices $\mathbf{\Sigma}^k$ converges to the boundary of $\mathbb{S}_{++}^K$ iff $\lambda_{\max}^k \to +\infty$ or $\lambda_{\min}^k \to 0$. In the rest of the thesis, we will use the statement "$\mathbf{\Sigma}$ converges" equivalently as "a sequence of matrices $\mathbf{\Sigma}^k$ converges".

4.2 Regularized Covariance Matrix Estimation

The regularity conditions for the existence of Tyler's estimator calls for the number of samples satisfying $N \geq K + 1$ [16, 88]. For applications where the number of samples violates this requirement, the technique of shrinkage is usually used to address the issue of singularity.

Two estimators were proposed in [19] and [2] leveraging the idea of regularizing an estimator via diagonal loading [93, 96]. In [19], the authors address the problem by adding a

scaled identity matrix per iteration to algorithm (3.1.6) as

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{t+1} &= \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \rho \mathbf{I} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)}.\end{aligned} \tag{4.2.1}$$

A slightly different one was later proposed in [2] as the limit of $\mathbf{\Sigma}_t$ generated by the following iteration:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{t+1} &= \frac{1}{1+\alpha_0} \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0} \mathbf{I} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)},\end{aligned} \tag{4.2.2}$$

where the uniqueness of the estimator was proved based on concave Perron-Frobenius theory, and a method to choose the regularization weight α_0 was also given. However, unlike the shrinkage estimator for SCM, which can be derived based on minimizing a penalized likelihood function, they do not have such an interpretation.

In contrast, the author of [20] took a different route and derived a shrinkage Tyler's estimator that has an interpretation based on minimizing the penalized negative log-likelihood function

$$L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right) = \frac{2}{N} L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right) + \alpha_0 h^{\mathrm{target}}\left(\mathbf{\Sigma}\right) \tag{4.2.3}$$

where $h^{\mathrm{target}}\left(\mathbf{\Sigma}\right) = K \log\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right) + \log\det\left(\mathbf{\Sigma}\right)$ is a function with minimum attained at the desired target matrix $\mathbf{T}$, hence it will shrink the solution of (4.2.3) towards the target. By showing the cost function $L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right)$ is geodesically convex, the author proved that any local minimum over the set of positive definite matrices is a global minimum [20]. He then derived an iterative algorithm based on majorization-minimization that monotonically decreases the cost function at each iteration:

$$\mathbf{\Sigma}_{t+1} = \frac{1}{1+\alpha_0} \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0} \frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}. \tag{4.2.4}$$

Even though it is shown in [20] that the cost function is convex in geodesic space, the existence and uniqueness of the global minimizer remains unknown. Moreover, it is mentioned there that for some values of α_0 the cost function becomes unbounded below and the

iterations do not converge.

In this section, we address the following points: (i) we give the missing interpretation based on minimizing a cost function for the estimator defined in (4.2.2) *without the trace normalization step*, and we also prove its existence and uniqueness[1]; (ii) we prove the iteration in (4.2.4) with an additional trace normalization step converges to a unique point and also establish the conditions on the regularization parameter α_0 to ensure the existence of the solution. For both cases, the cost function takes the form of penalized negative log-likelihood function with different penalizing functions. Our methodology for the proofs hinges on techniques used by Tyler in [16, 88].

We start with a proof of existence for a minimizer of a general penalized negative log-likelihood function in the following theorem, the proof of existence of the two aforementioned cases $L^{\text{Tyler}}(\mathbf{\Sigma})$ and $L^{\text{Wiesel}}(\mathbf{\Sigma})$ are just special cases of the general result.

The idea of proving the existence is to establish the regularity conditions under which the cost function takes value $+\infty$ on the boundary of the set $\mathbb{S}_{++}^K$, a minimum then exists by the continuity of the cost function. To establish the main result in Theorem 4.1, the following lemma is needed.

Lemma 4.1. *For any continuous function f defined on the set $\mathbb{S}_{++}$, there exists a $\hat{\mathbf{\Sigma}} \succ \mathbf{0}$ such that $f\left(\hat{\mathbf{\Sigma}}\right) \leq f(\mathbf{\Sigma})\ \forall \mathbf{\Sigma} \succ \mathbf{0}$ if $f(\mathbf{\Sigma}) \to +\infty$ on the boundary of the set $\mathbb{S}_{++}$.*

Proof. The conclusion follows from the continuity of f and the Weierstrass extreme value theorem. □

Definition 4.1. For any continuous function $f(s)$ defined on $s > 0$, define the quantities

$$a_f = \sup\left\{a | s^{a/2} \exp\left(-f(s)\right) \to 0 \text{ as } s \to +\infty\right\} \tag{4.2.5}$$

and

$$a_f^{'} = \inf\left\{a | s^{a/2} \exp\left(-f(s)\right) \to 0 \text{ as } s \to 0\right\} \tag{4.2.6}$$

In this chapter, we are particularly interested in the functions $f(s) = c\log s$ and $f(s) = cs$ with some positive scalar $c < +\infty$. For $f(s) = c\log s$, $a_f = a_f^{'} = 2c$ and, for $f(s) = cs$, $a_f = +\infty$, $a_f^{'} = 0$. We restrict our attention to the case $a_f \geq 0$.

[1]Note that by eliminating the trace normalization step the new iteration will lead to an estimator different from (4.2.2).

Consider the penalized cost function takes the general form

$$\tilde{L}(\mathbf{\Sigma}) = L^{\rho}(\mathbf{\Sigma}) + h(\mathbf{\Sigma}), \tag{4.2.7}$$

with original cost function

$$L^{\rho}(\mathbf{\Sigma}) = \frac{N}{2}\log\det(\mathbf{\Sigma}) + \sum_{i=1}^{N}\rho\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \tag{4.2.8}$$

where ρ is a continuous function, and the penalty term

$$h(\mathbf{\Sigma}) = \alpha\log\det(\mathbf{\Sigma}) + \sum_{l=1}^{L}\alpha_l h_l\left(\mathrm{Tr}\left(\mathbf{A}_l^T\mathbf{\Sigma}^{-1}\mathbf{A}_l\right)\right) \tag{4.2.9}$$

where $\mathrm{Tr}\left(\mathbf{A}_l^T\mathbf{\Sigma}^{-1}\mathbf{A}_l\right)$ measures the difference between $\mathbf{\Sigma}$ and the positive semidefinite matrix $\mathbf{A}_l\mathbf{A}_l^T$. h_l is, in general, an increasing function that increases the penalty as $\mathbf{\Sigma}$ deviates from $\mathbf{A}_l\mathbf{A}_l^T$, which is considered to be the prior target that we wish to shrink $\mathbf{\Sigma}$ to[2].

We first give an intuitive argument on the condition that ensures the existence of the estimator. Since the estimator $\hat{\mathbf{\Sigma}}$ is defined as the minimizer of the penalized loss function, it exists if $\tilde{L}(\mathbf{\Sigma}) \to +\infty$ on the boundary of $\mathbb{S}_{++}^K$ by Lemma 4.1, and clearly $\hat{\mathbf{\Sigma}}$ is nonsingular. We infer $\mathbf{\Sigma}$ by the samples $\{\mathbf{x}_i\}$, if the samples are concentrated on some subspace, naturally we "guess" the distribution is degenerate, i.e., $\hat{\mathbf{\Sigma}}$ is singular. Therefore, the samples are required to be sufficiently spread out in the whole space so that the inference leads to a nonsingular $\hat{\mathbf{\Sigma}}$. Under the case when we have a prior information that $\mathbf{\Sigma}$ should be close to the matrix $\mathbf{A}_l\mathbf{A}_l^T$, to ensure $\hat{\mathbf{\Sigma}}$ being nonsingular we need to distribute more $\mathbf{x}_i$'s in the null space of $\mathbf{A}_l\mathbf{A}_l^T$ and hence less in the range of $\mathbf{A}_l\mathbf{A}_l^T$. To formalize this intuition, we give the following theorem.

Theorem 4.1. *For cost function $L^{\rho} : \mathbb{S}_{++} \to \mathbb{R}$ given in* (4.2.7)*, define a_{ρ} and $a_{\rho}^{'}$ for ρ, a_l and $a_l^{'}$ for $\alpha_l h_l$'s according to* (4.2.5) *and* (4.2.6)*. Then $\tilde{L}(\mathbf{\Sigma}) \to +\infty$ on the boundary of the set $\mathbb{S}_{++}^K$ if the following conditions are satisfied:*

(i) no $\mathbf{x}_i$ lies on the origin;

[2]We consider shrinkage to multiple targets in the general formulation, which is shown to have better performance than a single target in some applications. See [97] and the references therein.

(ii) for any proper subspace S

$$P_N(S) < \min\left\{1 - \frac{(N+2\alpha)(K-\dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}, \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}\right\}$$

where sets ω *and* υ *are defined as* $\omega = \{l|\mathbf{A}_l \subseteq S\}$, $\upsilon = \{l|\mathbf{A}_l \nsubseteq S\}$;

(iii) $\left(-\frac{N}{2} - \alpha\right) K + \frac{a_\rho^{'}}{2} N + \frac{1}{2}\sum_l a_l^{'} < 0$ *and* $\frac{a_\rho}{2} N - \left(\frac{N}{2} + \alpha\right) K + \frac{1}{2}\sum_l a_l > 0$.

Proof. See Appendix 4.6.1. □

Remark 4.1. Condition (i) avoids the scenario when $\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i$ takes value 0 and $\rho(s)$ is undefined at $s = 0$, for example $\rho(s) = \log(s)$ for the log-likelihood function. The first part in condition (ii), $P_N(S) < 1 - \frac{(N+2\alpha)(K-\dim(S))-\sum_{l\in\upsilon} a_l}{a_\rho N}$, ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ under the case that some but not all eigenvalues λ_j of $\boldsymbol{\Sigma}$ tend to zero, and the second part in condition (ii), $P_N(S) < \frac{(N+2\alpha)\dim(S)-\sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}$, ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ under the case that some but not all eigenvalues λ_j of $\boldsymbol{\Sigma}$ tend to positive infinity. Together they force $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when $\frac{\lambda_{\max}}{\lambda_{\min}} \to +\infty$. The first part of condition (iii) ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when all $\lambda \to +\infty$ and the second part ensures $\tilde{L}(\boldsymbol{\Sigma}) \to +\infty$ when all $\lambda \to 0$.

Corollary 4.1. *Assuming the population distribution* $f(\cdot)$ *is continuous, and the matrices* $\mathbf{A}_l$ *are full rank, condition (ii) in Theorem 4.1 simplifies to:*

$$\begin{cases} \sum_l a_l - (N+2\alpha)(K-d) > a_\rho(d-N) \\ \alpha > \frac{a_\rho^{'} - N}{2} \end{cases}, \ \forall 1 \le d \le K-1.$$

Proof. The conclusion follows from the following two facts: given that the sampling distribution f is continuous, and no $\mathbf{x}_i$ lies on the origin, any $1 \le d < K$ sample points define a proper subspace S with $\dim(S) = d$ with probability one; and since the $\mathbf{A}_l$'s are full rank, the set $\omega = \emptyset$. □

Under the regularity conditions provided in Theorem 4.1, Lemma 4.1 implies a minimizer $\hat{\boldsymbol{\Sigma}}$ of $\tilde{L}(\boldsymbol{\Sigma})$ exists and is positive definite, therefore it needs to satisfy the condition $\frac{\partial \tilde{L}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \mathbf{0}$.

We then show how Theorem 4.1 works for Tyler's estimator defined as the nonsingular minimizer of (3.1.4). Notice that the loss function $L^{\text{Tyler}}(\boldsymbol{\Sigma})$ is scale-invariant, we have $L^{\text{Tyler}}(c\boldsymbol{\Sigma}_0) = L^{\text{Tyler}}(\boldsymbol{\Sigma}_0)$ for any $\boldsymbol{\Sigma}_0 \in \mathbb{S}_{++}$ and any $c > 0$. This implies that there are cases when $\boldsymbol{\Sigma}$ goes to the boundary of $\mathbb{S}_{++}^K$ and $L^{\text{Tyler}}(\boldsymbol{\Sigma})$ will not go to positive infinity.

Due to this reason, condition (iii) is violated in Theorem 4.1. To handle the scaling issue, we introduce a trace constraint $\mathrm{Tr}\left(\mathbf{\Sigma}\right)=1$.

For the Tyler's problem of minimizing (3.1.4), we seek the condition that ensures $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)$ goes to infinity when $\mathbf{\Sigma}$ goes to the boundary of the set $\{\mathbf{\Sigma}|\mathbf{\Sigma}\succ\mathbf{0},\ \mathrm{Tr}\left(\mathbf{\Sigma}\right)=1\}$ relative to $\{\mathbf{\Sigma}|\mathbf{\Sigma}\succeq\mathbf{0},\ \mathrm{Tr}\left(\mathbf{\Sigma}\right)=1\}$. The condition implies that there is a unique minimizer $\hat{\mathbf{\Sigma}}$ that minimizes $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)$ over the set $\{\mathbf{\Sigma}|\mathbf{\Sigma}\succ\mathbf{0},\ \mathrm{Tr}\left(\mathbf{\Sigma}\right)=1\}$, which is equivalent to the existence of a unique (up to a positive scale factor) minimizer $\mathbf{\Sigma}^{\star}$ that minimizes $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)$ over the set $\mathbb{S}_{++}^{K}$ since $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)$ is scale-invariant.

The constraint $\mathrm{Tr}\left(\mathbf{\Sigma}\right)=1$ excludes the case that any of $\lambda_j\to+\infty$ and the case all $\lambda_j\to 0$, hence we only need to let $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)\to+\infty$ under the case that some but not all $\lambda_j\to 0$, which corresponds to the condition $P_N\left(S\right)<1-\frac{(N+2\alpha)(K-\dim(S))-\sum_{l\in v}a_l}{a_\rho N}$ in Theorem 4.1. For Tyler's cost function $L^{\mathrm{Tyler}}\left(\mathbf{\Sigma}\right)$, we have $\rho\left(s\right)=\frac{K}{2}\log s$ and $\alpha=0$, $a_\rho=a_\rho^{'}=K$, therefore Theorem 4.1 leads to the condition on the samples: $P_N\left(S\right)<\frac{\dim(S)}{K}$, or $N\geq K+1$ if the population distribution f is continuous, which reduces to the condition given in [16].

4.2.1 Regularization via Wiesel's penalty

In [20], Wiesel proposed two scale-invariant regularization penalties of the following form:

$$\begin{aligned} h^{\mathrm{identity}}\left(\mathbf{\Sigma}\right)&=K\log\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\right)\right)+\log\det\left(\mathbf{\Sigma}\right)\\ h^{\mathrm{target}}\left(\mathbf{\Sigma}\right)&=K\log\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{\Sigma}\right). \end{aligned} \tag{4.2.10}$$

Wiesel justified the choice of the above mentioned penalty functions by showing that the minimizer is some scaled multiple of $\mathbf{I}$ (or $\mathbf{T}$). Thus adding this penalty terms to the Tyler's cost function would yield estimators that are shrunk towards $\mathbf{I}$ (or $\mathbf{T}$). In the rest of this subsection we consider the general case h^{target} only, where the penalty term shrinks $\mathbf{\Sigma}$ to scalar multiples of $\mathbf{T}$, and we make the assumption that $\mathbf{T}$ is positive definite, which is reasonable since a nonsingular estimator $\hat{\mathbf{\Sigma}}$ is desired. The cost function is restated below for convenience

$$\begin{aligned} L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right)=&\log\det\left(\mathbf{\Sigma}\right)+\frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right)\\ &+\alpha_0\left(K\log\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{\Sigma}\right)\right). \end{aligned} \tag{4.2.11}$$

Minimizing $L^{\text{Wiesel}}(\mathbf{\Sigma})$ gives the fixed-point condition

$$\mathbf{\Sigma} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)}. \tag{4.2.12}$$

Recall that in the absence of regularization (i.e., $\alpha_0 = 0$), a solution to the fixed-point equation exists under the condition $P_N(S) < \frac{\dim(S)}{N}$. With the regularization, however, it is not clear. We start giving a result for the uniqueness and then come back to the existence.

Theorem 4.2. *If equation* (4.2.12) *has a solution, then it is unique up to a positive scale factor.*

Proof. It's easy to see if $\mathbf{\Sigma}$ solves (4.2.12), $c\mathbf{\Sigma}$ is also a solution for $c > 0$. Without loss of generality assume $\mathbf{\Sigma} = \mathbf{I}$ is a solution, otherwise define $\tilde{\mathbf{x}}_i = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}_i$ and $\tilde{\mathbf{T}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{T}\mathbf{\Sigma}^{-\frac{1}{2}}$, and that there exists another solution $\mathbf{\Sigma}_1$. Denote the eigenvalues of $\mathbf{\Sigma}_1$ as $\lambda_1 \geq \cdots \geq \lambda_K$ with at least one strictly inequality, then under the condition that $\mathbf{T}$ is positive definite

$$\begin{aligned}
\mathbf{\Sigma}_1 &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_1^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\mathbf{\Sigma}_1^{-1}\mathbf{T}\right)} \\
&\prec \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_1^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\text{Tr}\left(\lambda_1^{-1}\mathbf{T}\right)} \\
&= \lambda_1\mathbf{I}
\end{aligned}$$

where the inequality follows from the fact that $\text{Tr}\left(\mathbf{S}\mathbf{\Sigma}_1^{-1}\right) > \text{Tr}\left(\lambda_1^{-1}\mathbf{S}\right)$ for any positive definite matrix $\mathbf{S}$ and the last equality follows from the assumption that $\mathbf{I}$ is a solution to (4.2.12). We have the contradiction $\lambda_1 < \lambda_1$, hence all the eigenvalues of $\mathbf{\Sigma}_1$ should be equal, i.e., $\mathbf{\Sigma}_1 = \lambda\mathbf{I}$. □

Before establishing the existence condition, we give an example when the solution to (4.2.12) does not exist for illustration.

Example 4.1. Consider the case when all $\mathbf{x}_i$'s are aligned in one direction. Eigendecompose $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and choose $\mathbf{u}_1$ to be aligned with the $\mathbf{x}_i$'s, let $\lambda_1 \to +\infty$ while others $0 < c \leq \lambda < +\infty$. Ignoring the constant terms, the boundedness of $L^{\text{Wiesel}}(\mathbf{\Sigma})$ is equivalent to the boundedness of $(1+\alpha_0-K)\log\lambda_1$, hence it is unbounded below if $\alpha_0 < K-1$.

The example shows that $L^{\text{Wiesel}}(\mathbf{\Sigma})$ can be unbounded below, implying that (4.2.12) has

no solution if the data are too concentrated and α_0 is small. The following theorem gives the exact tradeoff between data dispersion and the choice of α_0.

Theorem 4.3. *A unique solution to* (4.2.12) *exists (up to a positive scale factor) if the following conditions are satisfied:*

(i) no $\mathbf{x}_i$ *lies on the origin;*

(ii) for any proper subspace $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$,

and they are the global minima of the loss function (4.2.11).

Proof. We start by rewriting the function including a scale factor $\frac{N}{2}$ w.r.t. (4.2.11) for convenience:

$$\begin{aligned} L^{\text{Wiesel}}(\mathbf{\Sigma}) =& \frac{N}{2}\log\det(\mathbf{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ &+ \frac{\alpha_0 N}{2}\left(K\log\left(\operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right) + \log\det(\mathbf{\Sigma})\right). \end{aligned} \tag{4.2.13}$$

Invoke Theorem 4.1 with $\rho(s) = \frac{K}{2}\log(s)$, $h_1(s) = K\log(s)$, $\alpha = \alpha_1 = \frac{\alpha_0 N}{2}$ and $\mathbf{A}_1 = \mathbf{T}^{\frac{1}{2}}$, hence $a_\rho = a_\rho^{'} = K$ and $a_1 = a_1^{'} = 2\alpha K$. By the same reasoning as for the Tyler's loss function, the condition $P_N(S) < 1 - \frac{(N+2\alpha)(K-\dim(S))-\sum_{l\in v} a_l}{a_\rho N}$, which is $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ since $\mathbf{T}$ is full rank, ensures the existence of a unique solution to (4.2.12) under the constraint $\mathbf{\Sigma} \succ \mathbf{0}$ and $\operatorname{Tr}(\mathbf{\Sigma}) = 1$. Hence a unique (up to a positive scale factor) solution to (4.2.12) exists on the set of $\mathbb{S}_{++}^K$ by the scale-invariant property of $L^{\text{Wiesel}}(\mathbf{\Sigma})$. □

To make the existence condition easy to check, we use Corollary 4.1, Theorem 4.3 then simplifies to $\alpha_0 > \frac{K}{N} - 1$ or, equivalently $N > \frac{K}{1+\alpha_0}$, from which we can see that compared to the condition without regularization shrinkage allows less number of samples, and the minimum number depends on α_0.

At last, we show that the condition $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$ is also necessary in the following proposition.

Proposition 4.1. *If* (4.2.12) *admits a solution on* $\mathbb{S}_{++}^K$, *then for any proper subspace* $S \subseteq \mathbb{R}^K$, $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$, *provided that* $\mathbf{T}$ *is positive definite and* $\alpha_0 > 0$.

Proof. For a proper subspace S, define $\mathbf{P}$ as the orthogonal projection matrix associated to S, i.e., $\mathbf{P}\mathbf{x} = \mathbf{x}$, $\forall \mathbf{x} \in S$. Assume the solution is $\mathbf{I}$. Multiplying both sides of equation (4.2.12) by matrix $\mathbf{I} - \mathbf{P}$ and taking the trace we have

$$K - \dim(S) = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T(\mathbf{I}-\mathbf{P})\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0 K}{1+\alpha_0}\frac{\operatorname{Tr}(\mathbf{T}-\mathbf{T}\mathbf{P})}{\operatorname{Tr}(\mathbf{T})}$$

If $\mathbf{x}_i \in S$, then $\mathbf{x}_i^T (\mathbf{I} - \mathbf{P}) \mathbf{x}_i = 0$, and $\mathbf{x}_i^T (\mathbf{I} - \mathbf{P}) \mathbf{x}_i \leq \mathbf{x}_i^T \mathbf{x}_i$ if $\mathbf{x}_i \notin S$. Moreover, $\mathrm{Tr}(\mathbf{TP}) > 0$ since $\mathbf{T}$ is positive definite. This therefore implies

$$K - \dim(S) < \frac{1}{1+\alpha_0} \frac{K}{N} (N - N P_N(S)) + \frac{\alpha_0 K}{1+\alpha_0}.$$

Rearranging the terms yields $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$. □

4.2.2 Regularization via Kullback-Leibler Divergence Penalty

An ideal penalty term should increase as $\mathbf{\Sigma}$ deviates from the prior target $\mathbf{T}$. Wiesel's penalty function discussed in the last subsection satisfies this property and, in this subsection, we propose another penalty that has this property. The resulting estimator takes a simpler form than Wiesel's estimator, and is similar to Chen's estimator defined in (4.2.2).

The penalty that we choose is the KL divergence between $\mathcal{N}_\Sigma(\mathbf{0}, \mathbf{\Sigma})$ and $\mathcal{N}_T(\mathbf{0}, \mathbf{T})$, i.e., two zero-mean Gaussians with covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$, respectively. The formula for the KL divergence is as follows [98, 99]

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

Ignoring the constant terms results in the following loss function:

$$L^{\mathrm{KL}}(\mathbf{\Sigma}) = \log\det(\mathbf{\Sigma}) + \frac{K}{N} \sum_{i=1}^{N} \log\left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i\right) \quad + \alpha_0 \left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) + \log\det(\mathbf{\Sigma})\right) \tag{4.2.14}$$

with the following fixed-point condition:

$$\mathbf{\Sigma} = \frac{1}{1+\alpha_0} \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0} \mathbf{T}. \tag{4.2.15}$$

Unlike the penalty function discussed in the last subsection, KL divergence penalty encourages shrinkage towards $\mathbf{T}$ without scaling ambiguity since the minimizer of the KL divergence penalty is just $\mathbf{T}$. Notice that (4.2.15) is different from Chen's estimator (4.2.2) by eliminating the second trace normalization step.

Theorem 4.4. *If equation* (4.2.15) *has a solution, then it is unique.*

Proof. Without loss of generality, we assume $\mathbf{\Sigma} = \mathbf{I}$ solves (4.2.15). Assume there is another

matrix $\mathbf{\Sigma}_1$ that solves (4.2.15), and denote the largest eigenvalue of $\mathbf{\Sigma}_1$ as λ_1 and suppose $\lambda_1 > 1$. We then have the following contradiction:

$$\begin{aligned}\mathbf{\Sigma}_1 &\preceq \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_1^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T} \\ &< \frac{\lambda_1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0\lambda_1}{1+\alpha_0}\mathbf{T} = \lambda_1\mathbf{I}\end{aligned}$$

which gives contradiction $\lambda_1 < \lambda_1$, hence $\lambda_1 \leq 1$. Similarly, suppose the smallest eigenvalue of $\mathbf{\Sigma}_1$ satisfies $\lambda_K < 1$. We then have

$$\begin{aligned}\mathbf{\Sigma}_1 &\succeq \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_K^{-1}\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T} \\ &> \frac{\lambda_K}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{x}_i} + \frac{\alpha_0\lambda_K}{1+\alpha_0}\mathbf{T} = \lambda_K\mathbf{I}\end{aligned}$$

which is a contradiction and hence $\lambda_K \geq 1$, from which $\mathbf{\Sigma}_1 = \mathbf{I}$ follows. □

Theorem 4.5. *A unique solution of* (4.2.15) *exists, if*

(i) no $\mathbf{x}_i$ *lies on the origin;*

(ii) $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$,

and it is the global minimum of loss function (4.2.14).

Proof. Equivalently, we can define

$$L^{\mathrm{KL}}(\mathbf{\Sigma}) = \frac{N}{2}\log\det(\mathbf{\Sigma}) + \frac{K}{2}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) + \frac{\alpha_0 N}{2}\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) + \log\det(\mathbf{\Sigma})\right). \tag{4.2.16}$$

Invoke Theorem 4.1 with $\rho(s) = \frac{K}{2}\log(s)$, $h_1(s) = s$, $\alpha = \alpha_1 = \frac{\alpha_0 N}{2}$ and $\mathbf{A}_1 = \mathbf{T}^{\frac{1}{2}}$, hence $a_\rho = a_\rho^{'} = K$, $a_1 = +\infty$, $a_1^{'} = 0$. Since $\mathbf{T}$ is full rank and $a_1 = +\infty$, condition (ii) reduces to $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$. Condition (iii) is satisfied, hence an interior minimum exists. Furthermore, it is the unique minimum, hence it is global. □

Remark 4.2. The only difference between the regularized estimator discussed in this subsection and the estimator defined in (4.2.2) is the extra normalizing step in (4.2.2). With the trace normalization step, it is proved in [2] that the iteration implied by (4.2.2) converges to a unique solution without any assumption of the data. However, the iteration implied by (4.2.15), which is based on minimizing a negative log-likelihood function penalized via the

KL divergence function, requires some regularity conditions to be satisfied (cf. Theorem 4.5). According to Corollary 4.1, the condition simplifies to $\alpha_0 > \frac{K}{N} - 1$ if the population distribution is continuous.

Proposition 4.2. *If equation* (4.2.15) *admits a solution on* $\mathbb{S}_{++}^K$*, then for any proper subspace* $S \subseteq \mathbb{R}^K$*,* $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$*, provided that* $\mathbf{T}$ *is positive definite and* $\alpha_0 > 0$*.*

Proof. Multiply both sides of equation (4.2.15) by $\mathbf{T}^{-\frac{1}{2}}$ and define $\tilde{\mathbf{\Sigma}} = \mathbf{T}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{T}^{-\frac{1}{2}}$, $\tilde{\mathbf{x}}_i = \mathbf{T}^{-\frac{1}{2}}\mathbf{x}_i$ yields

$$\tilde{\mathbf{\Sigma}} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T}{\tilde{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\tilde{\mathbf{x}}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{I}. \tag{4.2.17}$$

The rest of the proof follows the same reasoning as Proposition 4.1. □

Finally, we show in the following proposition that the Wiesel's shrinkage estimator defined as solution to (4.2.12) and KL shrinkage estimator defined as solution to (4.2.15) are equivalent.

Proposition 4.3. *The solution to fixed point equation* (4.2.15) *solves* (4.2.12) *and, conversely, any solution of* (4.2.12) *solves* (4.2.15) *with a proper scale factor.*

Proof. If α_0 is zero, the statement is trivial. We consider the case $\alpha_0 \neq 0$. Following the argument of previous proposition we arrive at equation (4.2.17). It has been shown in [3] that the unique solution $\tilde{\mathbf{\Sigma}}$ to (4.2.17) satisfies $\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}^{-1}\right) = K$ given $\alpha_0 > 0$, hence $\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) = K$. Substitute it into equation (4.2.12) yields exactly equation (4.2.15) with solution $\mathbf{\Sigma}$, which indicates $\mathbf{\Sigma}$ solves (4.2.12). The second part of the proposition follows from the fact that Wiesel's fixed-point equation (4.2.12) has a unique solution up to a positive scale factor. □

4.3 Algorithms

In this section, we provide numerical algorithms that lead to the shrinkage estimators described above based on MM. For any continuous differentiable function f, we define $f(\mathbf{y}) = +\infty$ when $\lim_{\mathbf{x}\to\mathbf{y}} f(\mathbf{x}) = +\infty$. We briefly recall the MM procedure described in Chapter 2 below for convenience.

To minimize a *continuously differentiable* function f over a closed convex set $\mathcal{X}$, at current point $\mathbf{x}_t$, MM finds a surrogate function $g(\cdot|\mathbf{x}_t)$ that satisfies the following properties:

$$\begin{aligned} &f(\mathbf{x}_t) = g(\mathbf{x}_t|\mathbf{x}_t) \\ &f(\mathbf{x}) \leq g(\mathbf{x}|\mathbf{x}_t) \ \forall \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{4.3.1}$$

The surrogate function $g(\mathbf{x}|\mathbf{x}_t)$ is assumed to be continuous in $\mathbf{x}$ and $\mathbf{x}_t$. Variable $\mathbf{x}$ is then updated as $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\mathbf{x}_t)$.

4.3.1 Regularization via Wiesel's Penalty

In [20], Wiesel derived Tyler's iteration (3.1.6) but without the trace normalization step based on the MM algorithm, with surrogate function $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ for (3.1.4) defined as

$$g(\mathbf{\Sigma}|\mathbf{\Sigma}_t) = \frac{N}{2}\log\det(\mathbf{\Sigma}) + \sum_{i=1}^{N}\frac{K}{2}\frac{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \text{const.} \tag{4.3.2}$$

A positive definite stationary point of $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ satisfies the first equation of (3.1.6). By the same technique, to solve the problem

$$\begin{aligned} \underset{\mathbf{\Sigma}}{\text{minimize}} \quad & \log\det(\mathbf{\Sigma}) + \tfrac{K}{N}\textstyle\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \\ & +\alpha_0\left(K\log\left(\operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right) + \log\det(\mathbf{\Sigma})\right) \\ \text{subject to} \quad & \mathbf{\Sigma} \succeq \mathbf{0}. \end{aligned} \tag{4.3.3}$$

Wiesel derived the iteration (4.2.4) by majorizing (4.2.11) with function

$$(1+\alpha_0)\log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0 K}{\operatorname{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}\operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) + \text{const.} \tag{4.3.4}$$

It is worth pointing out that if we perform the change of variable $\boldsymbol{\psi} = \mathbf{\Sigma}^{-1}$ in $L^{\text{Wiesel}}(\mathbf{\Sigma})$ and linearize the term $\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right)$, this also leads to the same iteration (4.2.4).

In the rest of this subsection, we prove the convergence of the iteration (4.2.4) proposed by Wiesel, but with an additional trace normalization step, i.e., our modified iteration takes

Algorithm 4.1 Wiesel's shrinkage estimator

1. Initialize $\mathbf{\Sigma}_0$ as an arbitrary positive definite matrix.
2. Iterate
$$\tilde{\mathbf{\Sigma}}_{t+1} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}$$
$$\mathbf{\Sigma}_{t+1} = \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)}$$
until convergence.

the form:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{t+1} &= \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)} \\ \mathbf{\Sigma}_{t+1} &= \frac{\tilde{\mathbf{\Sigma}}_{t+1}}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)}.\end{aligned} \tag{4.3.5}$$

Denote the set $\mathcal{S} = \{\mathbf{\Sigma} | \mathrm{Tr}\left(\mathbf{\Sigma}\right) = 1,\ \mathbf{\Sigma} \succeq \mathbf{0}\}$.

Lemma 4.2. *The set* $\mathcal{X}^0 = \left\{\mathbf{\Sigma} | L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right) \leq L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}_0\right)\right\} \cap \mathcal{S}$ *is a compact set.*

Proof. The constraint $\mathrm{Tr}\left(\mathbf{\Sigma}\right) = 1$ implies the set $\mathcal{X}^0$ is bounded. The set is closed follows easily from the fact that $L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right) \to +\infty$ when $\mathbf{\Sigma}$ tends to be singular. □

Lemma 4.3. *The* $\tilde{\mathbf{\Sigma}}_{t+1}$ *given in* (4.3.5) *is the unique minimizer of surrogate function* (4.3.4).

Proof. For surrogate function (4.3.4), its value goes to positive infinity when $\frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})} \to +\infty$, since it majorizes $L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right)$ and $L^{\mathrm{Wiesel}}\left(\mathbf{\Sigma}\right) \to +\infty$ in this case. Now consider the case when $\frac{\lambda_{\max}}{\lambda_{\min}} = O\left(1\right)$. Define $\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}/\lambda_{\min}$, then function (4.3.4) can be rewritten as

$$\begin{aligned}&\left(1+\alpha_0\right)\left(\log\det\left(\bar{\mathbf{\Sigma}}\right) + K\log\lambda_{\min}\right) \\ &+ \frac{K}{N}\sum_{i=1}^{N}\lambda_{\min}^{-1}\frac{\mathbf{x}_i^T\bar{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0 K}{\mathrm{Tr}\left(\mathbf{\Sigma}_t^{-1}\mathbf{T}\right)}\mathrm{Tr}\left(\lambda_{\min}^{-1}\bar{\mathbf{\Sigma}}^{-1}\mathbf{T}\right) + \mathrm{const.}\end{aligned} \tag{4.3.6}$$

The terms $\log\det\left(\bar{\mathbf{\Sigma}}\right)$, $\frac{\mathbf{x}_i^T\bar{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i}$ and $\mathrm{Tr}\left(\bar{\mathbf{\Sigma}}^{-1}\mathbf{T}\right)$ are all constants bounded away from both 0 and $+\infty$. It is easy to see that when $\lambda_{\min} \to 0$ or $\lambda_{\min} \to +\infty$, (4.3.6) goes to $+\infty$. Therefore we conclude that the value of Wiesel's surrogate function (4.3.4) goes to $+\infty$ when $\mathbf{\Sigma}$ approaches the boundary of $\mathbb{S}_+^K$. The fact that $\tilde{\mathbf{\Sigma}}_{t+1}$ given in (4.3.5) is the unique solution to the stationary equation implies that it is the unique minimizer of (4.3.4) on the set $\mathbb{S}_+^K$. □

Proposition 4.4. *The sequence* $\{\mathbf{\Sigma}_t\}$ *generated by Algorithm 4.1 converges to the global minimizer of problem* (4.3.3).

Proof. It is proved in Theorem 4.3 that under the conditions provided in Theorem 4.3, the minimizer $\hat{\mathbf{\Sigma}}$ of problem

$$\begin{aligned} \underset{\mathbf{\Sigma}\succ\mathbf{0}}{\text{minimize}} \quad & \log\det\left(\mathbf{\Sigma}\right)+\frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right)+\alpha_0\left(K\log\left(\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{\Sigma}\right)\right) \\ \text{subject to} \quad & \mathrm{Tr}\left(\mathbf{\Sigma}\right)=1 \end{aligned} \tag{4.3.7}$$

exists and is unique, furthermore, it solves problem (4.3.3). It is also proved that the objective function $L^{\text{Wiesel}}\left(\mathbf{\Sigma}\right)\rightarrow+\infty$ on the boundary of the set $\mathcal{S}$. We now show that the sequence $\{\mathbf{\Sigma}_t\}$ converges to unique minimizer of (4.3.7).

Denote the surrogate function in general as $g\left(\mathbf{\Sigma}|\mathbf{\Sigma}_t\right)$, by Lemma 4.3 we therefore have the following inequality

$$L^{\text{Wiesel}}\left(\mathbf{\Sigma}_t\right)= \ g\left(\mathbf{\Sigma}_t|\mathbf{\Sigma}_t\right)\geq g\left(\tilde{\mathbf{\Sigma}}_{t+1}|\mathbf{\Sigma}_t\right)\geq L^{\text{Wiesel}}\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)=L^{\text{Wiesel}}\left(\mathbf{\Sigma}_{t+1}\right),$$

which means $\left\{L^{\text{Wiesel}}\left(\mathbf{\Sigma}_t\right)\right\}$ is a non-increasing sequence.

Assume that there exists converging subsequence $\mathbf{\Sigma}_{t_j}\rightarrow\mathbf{\Sigma}_\infty$, then

$$\begin{aligned} & g\left(\mathbf{\Sigma}|\mathbf{\Sigma}_{t_j}\right)\geq g\left(\tilde{\mathbf{\Sigma}}_{t_j+1}|\mathbf{\Sigma}_{t_j}\right)\geq L^{\text{Wiesel}}\left(\tilde{\mathbf{\Sigma}}_{t_j+1}\right) \\ =& L^{\text{Wiesel}}\left(\mathbf{\Sigma}_{t_j+1}\right)\geq L^{\text{Wiesel}}\left(\mathbf{\Sigma}_{t_{j+1}}\right)=g\left(\mathbf{\Sigma}_{t_{j+1}}|\mathbf{\Sigma}_{t_{j+1}}\right),\forall\mathbf{\Sigma}\succ\mathbf{0}. \end{aligned}$$

Letting $j\rightarrow+\infty$ results in

$$g\left(\mathbf{\Sigma}|\mathbf{\Sigma}_\infty\right)\geq g\left(\mathbf{\Sigma}_\infty|\mathbf{\Sigma}_\infty\right)\ \forall\mathbf{\Sigma}\succ\mathbf{0},$$

which implies that the directional derivative $L^{\text{Wiesel}'}\left(\mathbf{\Sigma}_\infty;\mathbf{\Delta}\right)\geq\mathbf{0},\forall\mathbf{\Sigma}_\infty+\mathbf{\Delta}\succ\mathbf{0}$. The limit $\mathbf{\Sigma}_\infty$ is nonsingular since if $\mathbf{\Sigma}_\infty$ is singular $L^{\text{Wiesel}}\left(\mathbf{\Sigma}_\infty\right)=+\infty$, but $L^{\text{Wiesel}}\left(\mathbf{\Sigma}_\infty\right)\leq L^{\text{Wiesel}}\left(\mathbf{\Sigma}_0\right)<+\infty$ given that $\mathbf{\Sigma}_0\succ\mathbf{0}$, which is a contradiction. Since $\mathbf{\Sigma}_\infty\succ\mathbf{0}$ and the function is continuously differentiable, we have $\frac{\partial L^{\text{Wiesel}}(\mathbf{\Sigma})}{\partial\mathbf{\Sigma}}\Big|_{\mathbf{\Sigma}_\infty}=\mathbf{0}$. Since $\mathrm{Tr}\left(\mathbf{\Sigma}_\infty\right)=1$, $\mathbf{\Sigma}_\infty=\hat{\mathbf{\Sigma}}$.

The set $\mathcal{X}^0=\{\mathbf{\Sigma}|L\left(\mathbf{\Sigma}\right)\leq L\left(\mathbf{\Sigma}_0\right)\}\cap\mathcal{S}$ is a compact set, and $\{\mathbf{\Sigma}_t\}$ lies in this set, hence $\{\mathbf{\Sigma}_t\}$ converges to $\hat{\mathbf{\Sigma}}$. □

Algorithm 4.2 KL divergence penalized shrinkage estimator

1. Initialize $\mathbf{\Sigma}_0$ as an arbitrary positive definite matrix.
2. Iterate
$$\mathbf{\Sigma}_{t+1} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T}$$
until convergence.

4.3.2 Regularization via Kullback-Leibler Penalty

Following the same approach, for the KL divergence penalty problem:

$$\begin{array}{ll}\underset{\mathbf{\Sigma}}{\text{minimize}} & \log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) + \alpha_0\left(\text{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) + \log\det(\mathbf{\Sigma})\right) \\ \text{subject to} & \mathbf{\Sigma} \succeq \mathbf{0}\end{array} \tag{4.3.8}$$

We can majorize $L^{\text{KL}}(\mathbf{\Sigma})$ at $\mathbf{\Sigma}_t$ by function

$$(1+\alpha_0)\log\det(\mathbf{\Sigma}) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \alpha_0\text{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{T}\right) \tag{4.3.9}$$

the stationary condition leads to the iteration

$$\mathbf{\Sigma}_{t+1} = \frac{1}{1+\alpha_0}\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{\Sigma}_t^{-1}\mathbf{x}_i} + \frac{\alpha_0}{1+\alpha_0}\mathbf{T}. \tag{4.3.10}$$

Algorithm 4.2 summarizes the procedure for KL shrinkage estimator.

Proposition 4.5. *The sequence* $\{\mathbf{\Sigma}_t\}$ *generated by Algorithm 4.2 converges to the global minimizer of problem* (4.3.8).

Proof. We verify the assumptions required for the convergence of algorithm [53], namely (4.3.1) and the compactness of initial level set $\mathcal{X}^0 = \left\{\mathbf{\Sigma} | L^{\text{KL}}(\mathbf{\Sigma}) \leq L^{\text{KL}}(\mathbf{\Sigma}_0),\ \mathbf{\Sigma} \succ \mathbf{0}\right\}$.

The first condition in (4.3.1) is satisfied by construction. To verify the second condition, we see that the gradient of the surrogate function $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ has a unique zero. Since $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$ is a global upperbound for $L^{\text{KL}}(\mathbf{\Sigma})$, $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t) \to +\infty$ as $\mathbf{\Sigma}$ goes to the boundary of $\mathbb{S}_+^K$. By the continuity of $g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$, a minimizer $\mathbf{\Sigma}^\star \succ \mathbf{0}$ exists and has to satisfy $\frac{\partial g}{\partial \mathbf{\Sigma}} = \mathbf{0}$. Therefore the unique zero has to be the global minimum, i.e., $\mathbf{\Sigma}_{t+1} = \arg\min_{\mathbf{\Sigma}\succeq\mathbf{0}} g(\mathbf{\Sigma}|\mathbf{\Sigma}_t)$. The last condition is satisfied since $L^{\text{KL}}(\mathbf{\Sigma})$ is continuously differentiable on $\mathbb{S}_{++}^K$.

It is proved in Theorems 4.4 and 4.5 that on set $\mathbb{S}_{++}^K$, $L^{\mathrm{KL}}(\mathbf{\Sigma})$ has a unique stationary point and it is the global minimum. Furthermore, the conditions in Theorem 4.5 ensures $L^{\mathrm{KL}}(\mathbf{\Sigma}) \to +\infty$ when $\mathbf{\Sigma}$ goes to the boundary of $\mathbb{S}_{+}^K$. The initial level set

$$\mathcal{X}^0 = \left\{ \mathbf{\Sigma} | L^{\mathrm{KL}}(\mathbf{\Sigma}) \leq L^{\mathrm{KL}}(\mathbf{\Sigma}_0),\ \mathbf{\Sigma} \succ \mathbf{0} \right\}$$

is compact follows easily.

Therefore the sequence $\{\mathbf{\Sigma}_t\}$ converges to the set of stationary points, hence the global minimum of problem (4.3.8). □

4.3.3 Parameter Tuning

A crucial issue in regularized covariance estimator is to choose the penalty parameter α_0. We have shown that if the population distribution is continuous, for both Wiesel's penalty and KL divergence penalty, we require $\alpha_0 > \frac{K}{N} - 1$ to guarantee the existence of the regularized estimator. A commonly adopted approach is to select α_0 by cross-validation [20]. However, in the context of low sample support and outlier contamination, the performance of traditional cross-validation method is questionable if the risk obtained on validation set cannot well approximate the true risk. A method based on random matrix theory has also been investigated in a recent work [100].

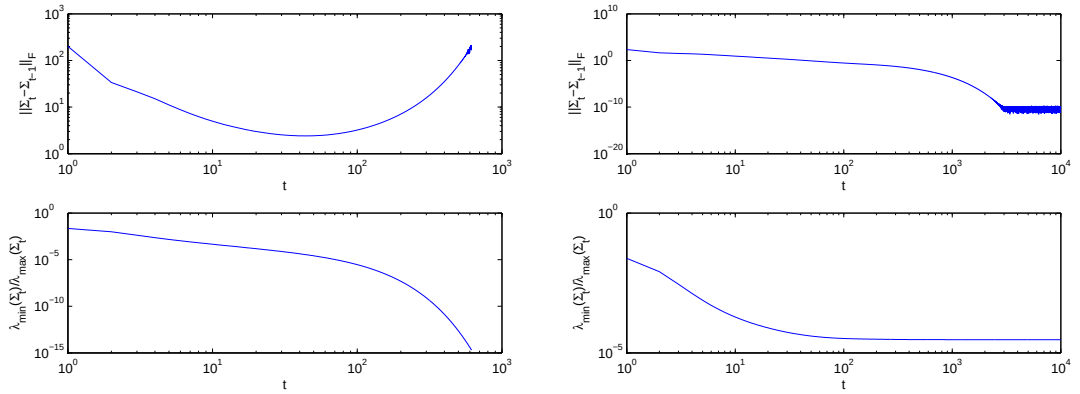
4.4 Numerical Results

In all of the simulations, the estimator performance is evaluated according to the criteria in [20], namely, the normalized mean-square error

$$\mathrm{NMSE} = \frac{\mathbb{E}\left(\left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^{\mathrm{true}} \right\|_F^2 \right)}{\left\| \mathbf{\Sigma}^{\mathrm{true}} \right\|_F^2}$$

where all matrices $\mathbf{\Sigma}$ are all normalized by their trace. The expected value is approximated by 100 times Monte-Carlo simulations.

The first two simulations aims at verifying the existence conditions for both Wiesel's shrinkage estimator and KL shrinkage estimator. We choose $N = 8$ and $K = 10$ with the samples drawn a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ and shrinkage target $\mathbf{T}$ are some



(a) (b)

Figure 4.1: Algorithm convergence of Wiesel's shrinkage estimator: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.24$, and (b) when the existence conditions are satisfied with $\alpha_0 = 0.26$.

arbitrary positive definite matrix[3]. According to the result in Section 4.2, $\alpha_0 > \frac{K}{N} - 1$, i.e., $\alpha_0 > 0.25$, is the necessary and sufficient condition for the existence of a positive definite estimator. We simulate two scenarios with $\alpha_0 = 0.24$ and 0.26. Fig. 4.1 plots $\|\mathbf{\Sigma}_t - \mathbf{\Sigma}_{t-1}\|_F$ and the inverse of the condition number, namely $\frac{\lambda_{\min}(\mathbf{\Sigma}_t)}{\lambda_{\max}(\mathbf{\Sigma}_t)}$, as a function of the number of iterations in log-scale for Wiesel's shrinkage estimator and with $\alpha_0 = 0.24$ (left) and $\alpha_0 = 0.26$ (right) respectively. Fig. 4.1 shows that for Wiesel's shrinkage estimator, when $\alpha_0 = 0.24$ $\mathbf{\Sigma}_t$ diverges, and when $\alpha_0 = 0.26$ $\mathbf{\Sigma}_t$ converges to a nonsingular limit. Fig. 4.2 shows similar situation happens for KL shrinkage estimator.

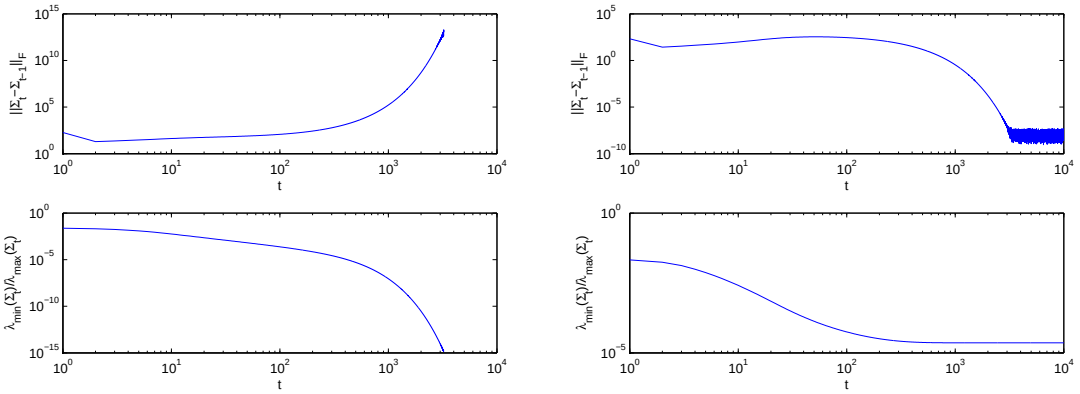
For the rest of the simulations, the shrinkage parameter α_0 is selected by grid search. That is, we define $\rho = \frac{1}{1+\alpha_0}$ and enumerate ρ uniformly on interval $(0, 1]$, and select the ρ (equivalently α_0) that gives the smallest error.

Fig. 4.3 demonstrates the performance of shrinkage Tyler's estimator in the sample deficient case. The tuning parameter is selected to be the one that yields the smallest NMSE for each estimator as proposed in [20]. We choose the example

$$\mathbf{\Sigma}\left(\beta\right)_{ij} = \beta^{|i-j|}$$

with $K = 30$. In this simulation, the underlying distribution is chosen to be a Student's t-distribution with parameters $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}\left(0.8\right)$, and $\nu = 3$, and the shrinkage target is

[3]We generate a positive definite matrix $\mathbf{M}$ by first generating a random matrix $\mathbf{F}$ with *i.i.d.* Gaussian entries of size $K \times L$ with $L > K$, then obtain $\mathbf{M}$ by $\mathbf{M} = \mathbf{F}\mathbf{F}^T$.



(a) (b)

Figure 4.2: Algorithm convergence of KL shrinkage estimator: (a) when the existence conditions are not satisfied with $\alpha_0 = 0.24$, and (b) when the existence conditions are satisfied with $\alpha_0 = 0.26$.

set to be an identity matrix. The number of samples N starts from 11 to 211. The estimators compared are: sample covariance matrix (SCM), normalized sample covariance matrix (NSCM) defined as $\hat{\mathbf{\Sigma}}_{NSCM} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\|\mathbf{x}_i\|^2}$, shrinkage NSCM defined as $\hat{\mathbf{\Sigma}}_{NSCM} + \rho\mathbf{T}$[4], Tyler estimator, and three types of shrinkage Tyler's estimators (Chen's estimator given by (4.2.2), Wiesel's estimator given by (4.3.5), and KL estimator given by (4.3.10)). The curve corresponding to Tyler's estimator starts at $N = 31$ since the condition for Tyler's estimator to exist is $N > K$, i.e., $N > 30$ in this case.

The figure illustrates that all estimators outperform the sample covariance matrix when they exist. Further, when the number of samples is small ($N \in [11, 91]$), NSCM outperforms Tyler's estimator, and the shrinkage estimators achieves almost the same estimation error. And when the number of samples is relatively large ($N > 91$), Tyler's estimator outperforms NSCM, and shrinkage Tyler's estimators outperforms shrinkage NSCM.

Figs. 4.4 and 4.5 compare the performance of different shrinkage Tyler's estimators following roughly one of the simulation set-up in [20] for a fair comparison. The samples are drawn from a Student's t-distribution with parameters $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}(0.8)$ and $\nu = 3$. The number of samples N varies from 11 to 91. Fig. 4.4 shows the estimation error when setting $\mathbf{T} = \mathbf{I}$ and Fig. 4.5 shows that when setting $\mathbf{T} = \mathbf{\Sigma}(0.7)$, the searching step size of ρ is set to be 0.01. The result indicates that estimation accuracy is increased due to shrinkage when the number of sample is not enough. Wiesel's shrinkage estimator and KL shrinkage

[4]Parameter ρ is chosen as the one that minimizes the NMSE. In the simulation this is obtained by increasing ρ from 0 with a step-size 0.01 until the NMSE does not decrease any more

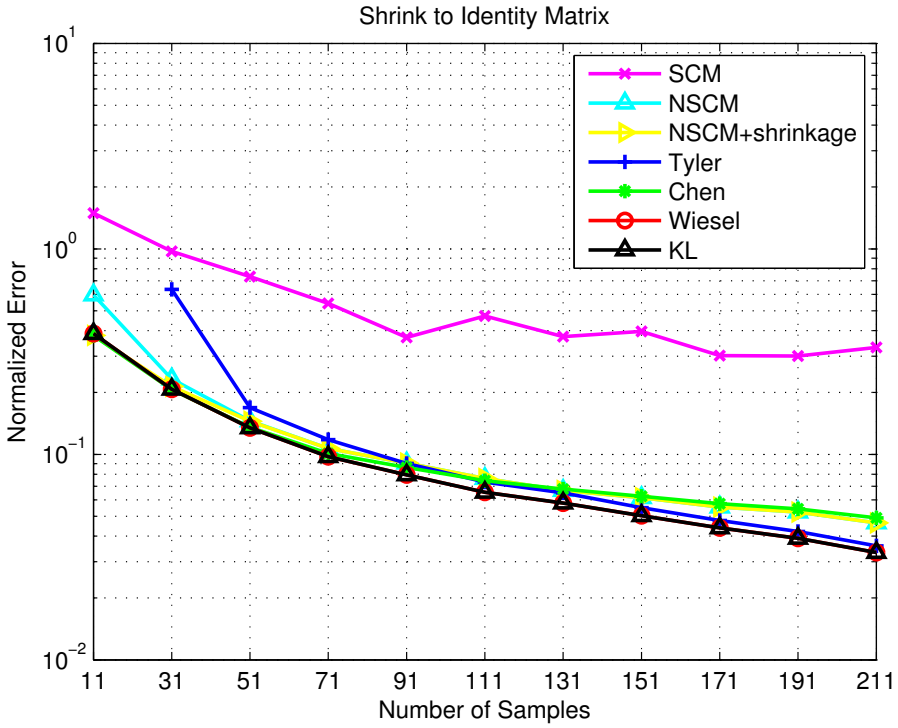


Figure 4.3: Illustration of the benefit of shrinkage estimators with $K = 30$ and shrinkage target matrix $\mathbf{I}$.

estimator yield the same NMSE. Interestingly, Chen's shrinkage estimator gives roughly the same NMSE, although with a different shrinkage parameter α_0. Chen's and KL shrinkage estimator thus find their advantage in practice since an easier way of choosing α_0 rather than cross-validation has been investigated in the literature [2, 100], a detailed comparison of them from random matrix theory perspective has also been provided in [100].

Finally, the performance of Tyler's estimator is tested on a real financial data set. We choose daily close prices p_t from Jan 1, 2008 to July 31, 2011, 720 days in total, of $K = 45$ stocks from the Hang Seng Index provided by Yahoo Finance. The samples are constructed as $r_t = \log p_t - \log p_{t-1}$, i.e., the daily log-returns. The process r_t is assumed to be stationary. The vector $\mathbf{r}_t$ is constructed by stacking the log-returns of all K stocks. $\mathbf{r}_t$ that is close to $\mathbf{0}$ (all elements are less than 10^{-6}) is discarded. We compare the performance of different covariance estimators in the minimum variance portfolio set up, that is, we allocate the portfolio weights to minimize the overall risk. The problem can be formulated formally as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} \quad & \mathbf{1}^T \mathbf{w} = 1 \end{aligned} \tag{4.4.1}$$

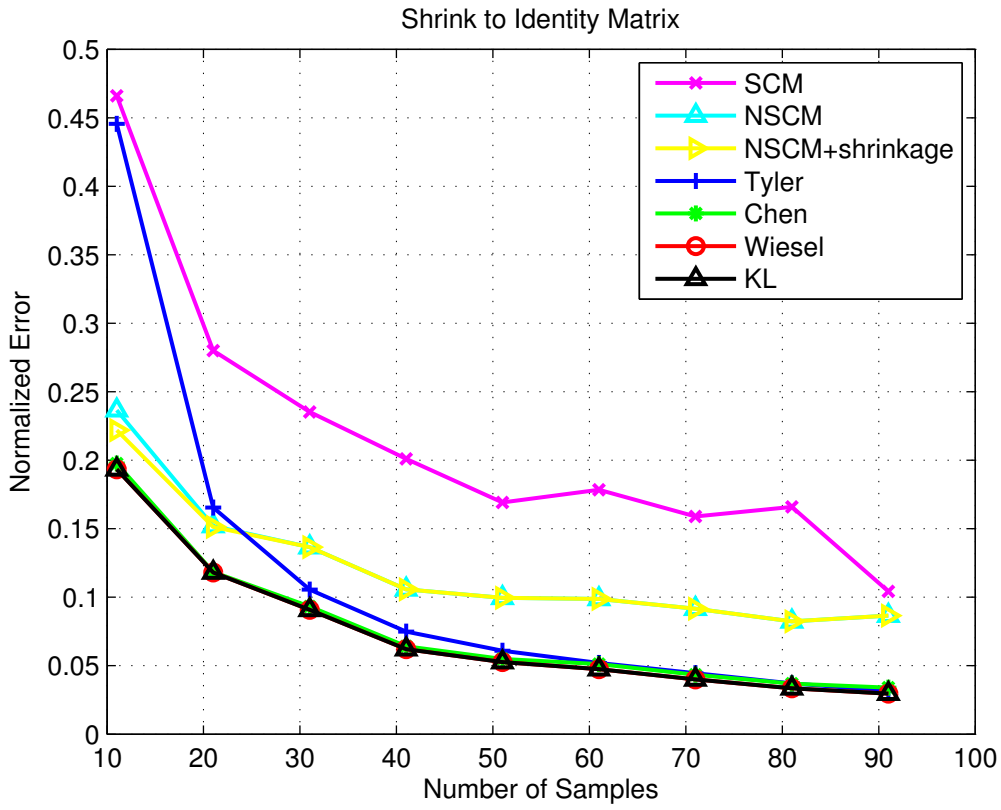


Figure 4.4: Illustration of the benefit of shrinkage estimators with $K = 10$ and shrinkage target matrix $\mathbf{I}$.

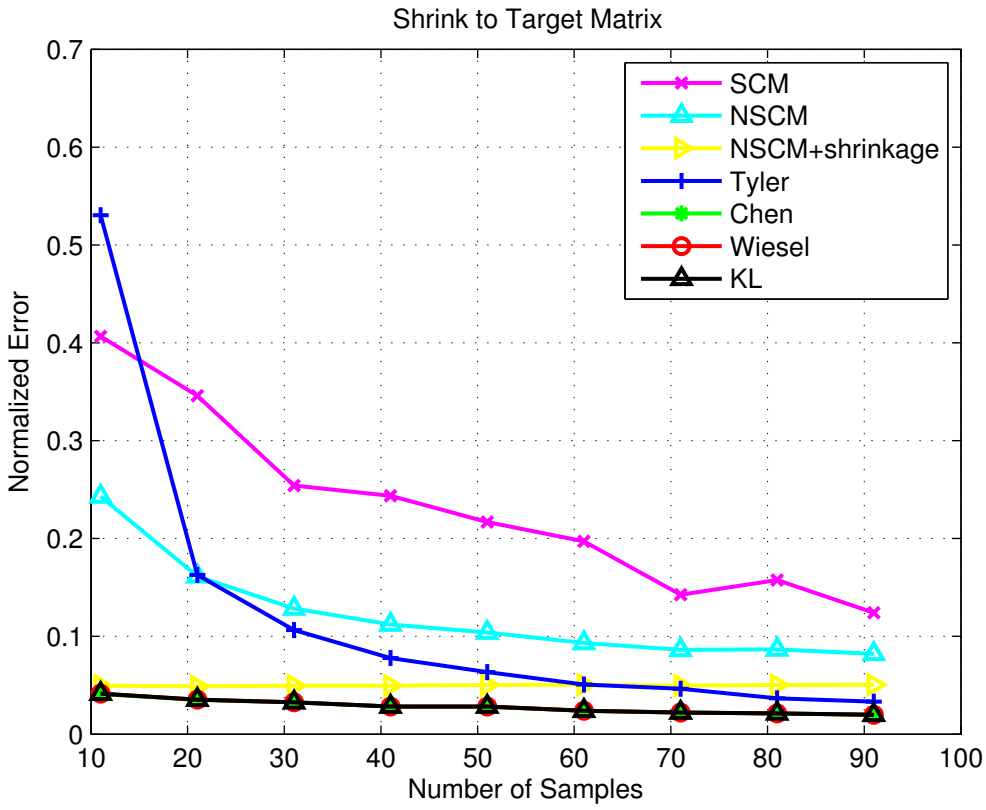


Figure 4.5: Illustration of the benefit of shrinkage estimators with $K = 10$ and a knowledge-aided shrinkage target matrix $\mathbf{T}$.

with $\mathbf{\Sigma}$ being the covariance matrix of $\mathbf{r}_t$. Clearly the scaling of $\mathbf{\Sigma}$ does not affect the solution to this problem.

The simulation takes the following procedure. At day t, we use the previous N $\mathbf{r}_t$'s as samples to estimate the covariance matrix. The training samples are split into two parts as $N = N^{\text{train}} + N^{\text{val}}$. For nonshrinkage estimators, at day t, we take the $\mathbf{r}_i$'s with $i \in \left[t - N^{\text{train}} - N^{\text{val}}, t-1\right]$ as samples to estimate the normalized covariance matrix $\mathbf{\Sigma}$. For a particular shrinkage estimator, the target matrix is set to be $\mathbf{I}$ and the tuning parameter ρ is chosen as follows: for each value of $\rho \in \{0.01, 0.02, \ldots, 1\}$, we calculate the shrinkage estimator $\mathbf{\Sigma}^{\rho}$ with samples $\mathbf{r}_i$, $i \in \left[t - N^{\text{train}} - N^{\text{val}}, t - N^{\text{val}} - 1\right]$ and the corresponding $\mathbf{w}^{\rho}$ by solving (4.4.1). We then take the $\mathbf{r}_i$'s with $i \in \left[t - N^{\text{val}}, t-1\right]$ as validation data and evaluate the variance of portfolio series $\left\{\left(\mathbf{w}^{\rho}\right)^T \mathbf{r}_i\right\}$ in this period, the best $\rho^{\star}$ is chosen to be the one that yields the smallest variance. Finally the shrinkage estimator is obtained using samples $\mathbf{r}_i$ with $i \in \left[t - N^{\text{train}} - N^{\text{val}}, t-1\right]$ and tuning parameter $\rho^{\star}$. With the allocation strategy $\mathbf{w}$ for each of the estimators as the solution to (4.4.1), we construct portfolio for the next N^{test} days and collect the returns. The procedure is repeated every N^{test} days till the end and the variance of the portfolio constructed based on different estimators is calculated.

In the simulation, we choose $N^{\text{val}} = N^{\text{test}} = 10$ and vary N^{train} from 70 to 100. Fig. 4.6 compares the variance (risk) of portfolio constructed based on different estimators, with one additional baseline portfolio constructed by equal investment in each asset. From the figure we can see shrinkage estimators achieves relatively better performance than the nonshrinkage ones.

4.5 Conclusion

In this work, we have given a rigorous proof for the existence and uniqueness of the regularized Tyler's estimator prov c v posed in [20], and a iterative diagonal loading shrinkage estimator with the KL divergence. Under the condition that samples are reasonably spread out, i.e., $P_N(S) < \frac{(1+\alpha_0)\dim(S)}{K}$, or $N > \frac{K}{1+\alpha_0}$ if the underlying distribution is continuous, the estimators have been shown to exist and be unique (up to a positive scale factor for Wiesel's estimator). Algorithms based on the majorization-minimization framework have also been

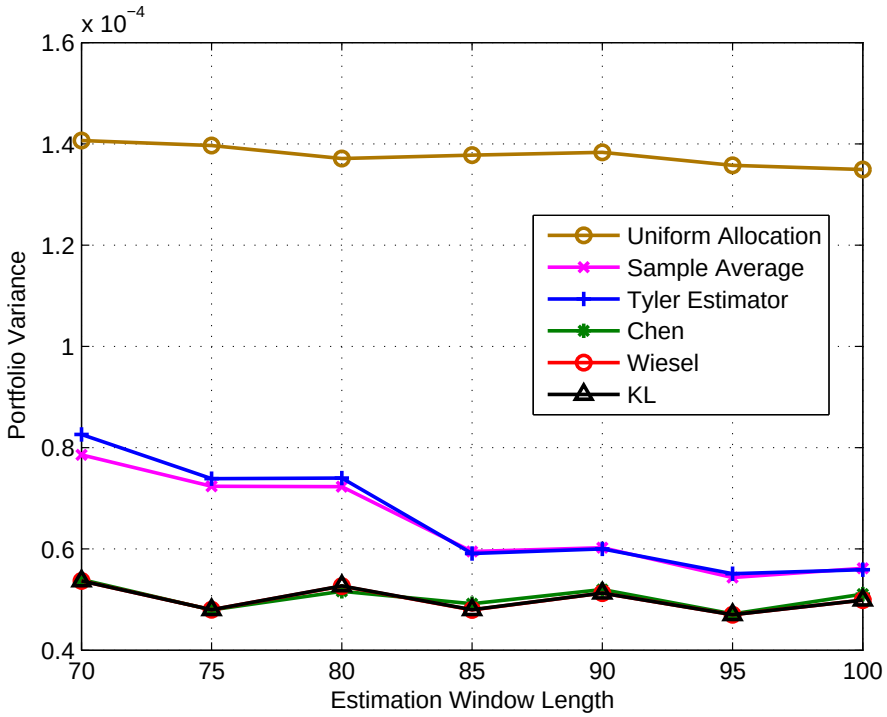


Figure 4.6: Comparison of portfolio risk constructed based on different covariance estimators.

provided with guaranteed convergence. Finally we have discussed structure constrained estimation and have shown via simulation that imposing such constraint helps improving estimation accuracy.

4.6 Appendix

4.6.1 Proof for Theorem 4.1

For the loss function

$$\tilde{L}\left(\mathbf{\Sigma}\right)=\frac{N}{2}\log\det\left(\mathbf{\Sigma}\right)+\sum_{i=1}^{N}\rho\left(\mathbf{x}_{i}^{T}\mathbf{\Sigma}^{-1}\mathbf{x}_{i}\right)+\left(\alpha\log\det\left(\mathbf{\Sigma}\right)+\sum_{l=1}^{L}\alpha_{l}h_{l}\left(\mathrm{Tr}\left(\mathbf{A}_{l}^{T}\mathbf{\Sigma}^{-1}\mathbf{A}_{l}\right)\right)\right)$$

where the regularization term is written in general as $\alpha\log\det\left(\mathbf{\Sigma}\right)+\sum_{l=1}^{L}\alpha_{l}h_{l}\left(\mathrm{Tr}\left(\mathbf{A}_{l}^{T}\mathbf{\Sigma}^{-1}\mathbf{A}_{l}\right)\right)$. Define a_{ρ}, $a_{\rho}^{'}$ for $\rho\left(s\right)$ and a_{l}, $a_{l}^{'}$ for $\alpha_{l}h_{l}$ as in Definition 4.1.

Define function

$$\begin{aligned} G\left(\mathbf{\Sigma}\right) &= \exp\left\{-\tilde{L}\left(\mathbf{\Sigma}\right)\right\} \\ &= \det\left(\mathbf{\Sigma}\right)^{-\frac{N}{2}-\alpha} \prod_i g_\rho\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) \cdot \prod g_l\left(\sum_j \lambda_j^{-1}\left\|\tilde{\mathbf{a}}_{lj}\right\|^2\right) \end{aligned}$$

where $\tilde{\mathbf{a}}_{lj}$ is defined as the jth row of $\tilde{\mathbf{A}}_l = \mathbf{U}^T\mathbf{A}_l$ with $\mathbf{U}$ being the unitary matrix such that $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{\Sigma}$, $\mathbf{\Lambda} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, and $g_\rho\left(s\right) = \exp\left\{-\rho\left(s\right)\right\}$, $g_l\left(s\right) = \exp\left\{-\alpha_l h_l\left(s\right)\right\}$. The eigenvalues λ_j is arranged in descending order, i.e., $\lambda_1 \geq \cdots \geq \lambda_K$, and denote the inverse of λ as φ, hence $\varphi_1 \leq \cdots \leq \varphi_K$.

Denote the eigenvectors corresponding to λ_j as $\mathbf{u}_j$, the subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ as S_j and $D_j = S_j \backslash S_{j-1} = \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x} \in S_j, \mathbf{x} \notin S_{j-1}\right\}$ with $S_0 = \{0\}$ and $D_0 = \{0\}$. By definition, $D_j,\ j = 0, \ldots, K$ partition the whole $\mathbb{R}^K$ space. Notice that $P_N\left\{S_0\right\} = 0$ by the assumption that no $\mathbf{x}_i$ lies on the origin, we have $\sum_{j=1}^m P_N\left(D_j\right) = P_N\left(S_m\right)$ and $\sum_{j=m}^K P_N\left(D_j\right) = 1 - P_N\left(S_{m-1}\right)$.

Partition the samples $\mathbf{x}_i$ according to D_j, $j = 0$ is excluded hereafter, define function

$$G_j \ = \begin{cases} \lambda_j^{-\frac{N}{2}-\alpha} \prod\limits_{\mathbf{x}_i \in D_j} g_\rho\left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right), & \text{if } \exists \mathbf{x}_i \in D_j \\ \lambda_j^{-\frac{N}{2}-\alpha}, & \text{if } \nexists \mathbf{x}_i \in D_j \end{cases}$$

and we have $G\left(\mathbf{\Sigma}\right) = \prod_{j=1}^K G_j\left(\mathbf{\Sigma}\right) \prod g_l\left(\sum_j \lambda_j^{-1}\left\|\tilde{\mathbf{a}}_{lj}\right\|^2\right)$.

For the $\mathbf{A}_l$'s, denote $\mathbf{A}_l = [\mathbf{a}_{l1}, \mathbf{a}_{l2}, \ldots, \mathbf{a}_{lp}]$. For each $\mathbf{a}_l$, there exists some D_j such that $\mathbf{a}_l \in D_j$, since the D_j's partition the whole space. Define q_l to be the maximum index of D_j that the $\mathbf{a}_l$'s belongs to. Therefore we have $\|\tilde{\mathbf{a}}_{lq_l}\| \neq 0$ and $\|\tilde{\mathbf{a}}_{lj}\| = 0$ for $j > q_l$.

We analyze the behavior of $G\left(\mathbf{\Sigma}\right)$ at the boundary of its feasible set $\mathbb{S}_{++}^K$, by Lemma 4.1, we only need to ensure $G\left(\mathbf{\Sigma}\right) \to 0$, then there exists $\tilde{L}\left(\hat{\mathbf{\Sigma}}\right) \leq \tilde{L}\left(\mathbf{\Sigma}\right)\ \forall \mathbf{\Sigma} \succ \mathbf{0}$, and $\hat{\mathbf{\Sigma}} \succ \mathbf{0}$.

Consider the general case that some of the λ_j's go to zero, some remains bounded away from both 0 and positive infinity, and the rests tend to positive infinity. Formally, define two integers r and s that $1 \leq r \leq s \leq K$, such that $\lambda_j \to +\infty$ for $j \in [1, r]$, λ_j is bounded for $j \in (r, s]$ and $\lambda_j \to 0$ for $j \in (s, K]$. Denote some arbitrary small positive quantity by ϵ.

First we analyze the terms G_j with $\lambda_j \to 0$. Consider the samples $\mathbf{x}_i \in D_h$ for some $h \in (s, K]$, then $\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i = \sum_{j=1}^h \lambda_j^{-1}\left\|\mathbf{u}_j^T\mathbf{x}_i\right\|^2 \geq \lambda_h^{-1}\left\|\mathbf{u}_h^T\mathbf{x}_i\right\|^2$, which is $+\infty > \left(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i\right)\lambda_h >$

0. Since $\lambda_h \to 0$, we have $\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \to +\infty$. By definition

$$\begin{aligned} & \lim_{\lambda_h \to 0} g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right)^{(a_\rho - \epsilon)/2} \\ = & \lim_{\lambda_h \to 0} \left\{ g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \lambda_h^{-(a_\rho - \epsilon)/2} \right\} \cdot \left\{ \left(\left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \lambda_h \right)^{(a_\rho - \epsilon)/2} \right\} \\ = & \ 0 \end{aligned}$$

which implies $\lim_{\lambda_h \to 0} \left\{ g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \lambda_h^{-(a_\rho - \epsilon)/2} \right\} = 0$, i.e., $g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) = o\left(\lambda_h^{\frac{a_\rho - \epsilon}{2}} \right)$. Therefore, if $\mathbf{x}_i \in D_j$, $g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) = o\left(\lambda_j^{\frac{a_\rho - \epsilon}{2}} \right)$. For each G_j we have

$$G_j = o\left(\lambda_j^{\frac{a_\rho - \epsilon}{2} \cdot NP_n(D_K) - \frac{N}{2} - \alpha - \epsilon} \right) \ \forall j \geq s+1.$$

In the second step, we analyze the terms G_j with $\lambda_j \to +\infty$. Consider the samples $\mathbf{x}_i \in D_h$ for some $h \in [1, r]$, we have shown that $0 < \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \lambda_h < +\infty$. Since $\lambda_h \to +\infty$, $\left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \to 0$. Given that $a_\rho^{'} > -\infty$, $g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) = o\left(\varphi_h^{-\frac{a_\rho^{'} + \epsilon}{2}} \right)$ by

$$\lim_{\varphi_h \to 0} g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right)^{\left(a_\rho^{'} + \epsilon \right)/2} = 0.$$

Therefore for each G_j we have

$$G_j = o\left(\varphi_j^{\frac{N}{2} + \alpha - \frac{a_\rho^{'} + \epsilon}{2} NP_N(D_j) - \epsilon} \right) \ \forall j \leq r.$$

For the G_j with λ_j being some constant, it is easy to see that $g_\rho \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i \right) = O(1)$, which does not affect the order of $G(\mathbf{\Sigma})$.

Now we have characterized the g_ρ's, we move to the g_l's. Since $\|\tilde{\mathbf{a}}_{lq_l}\| \neq 0$ and $\|\tilde{\mathbf{a}}_{lj}\| = 0$ for $j > q_l$, by the same reasoning above, $g_l = o\left(\varphi_{q_l}^{-\frac{a_l^{'} + \epsilon}{2}} \right)$ if $q_l \leq r$ and $g_l = o\left(\lambda_{q_l}^{\frac{a_l - \epsilon}{2}} \right)$ if

$q_l \geq s+1$. Therefore

$$\begin{aligned} G(\mathbf{\Sigma}) &= \prod_{j=1}^{K} G_j(\mathbf{\Sigma}) \prod g_l \left(\sum_j \lambda_j^{-1} \left\| \tilde{\mathbf{a}}_{lj} \right\|^2 \right) \\ &= \prod_{j=1}^{r} o\left(\varphi_j^{\frac{N}{2}+\alpha-\frac{a_\rho^{'}+\epsilon}{2} NP_N(D_j)-\epsilon} \right) \prod_{j=s+1}^{K} o\left(\lambda_j^{\frac{a_\rho-\epsilon}{2}\cdot NP_N(D_j)-\frac{N}{2}-\alpha-\epsilon} \right) \times \\ &\quad \prod_{\{l|q_l \geq s+1\}} o\left(\lambda_{q_l}^{\frac{a_l-\epsilon}{2}} \right) \prod_{\{l|q_l \leq r\}} o\left(\varphi_{q_l}^{-\frac{a_l^{'}+\epsilon}{2}} \right) \end{aligned}$$

with $\prod_{\{l|q_l \geq s+1\}}$ defined to be 1 if the set $\{l|q_l \geq s+1\}$ is empty, and the same for $\prod_{\{l|q_l \leq r\}}$.

We make the following assumption:

$$\begin{aligned} &\left(\tfrac{N}{2}+\alpha-\epsilon\right) m - \tfrac{a_\rho^{'}+\epsilon}{2} N \textstyle\sum_{j=1}^{m} P_N(D_j) - \sum_{q_l \leq m} \tfrac{a_l^{'}+\epsilon}{2} \geq 0,\ \forall 1 \leq m \leq r \\ &\tfrac{a_\rho}{2} \cdot N \textstyle\sum_{j=m}^{K} P_N(D_j) - \left(\tfrac{N}{2}+\alpha+\epsilon\right)(K-m+1) + \sum_{q_l \geq m} \tfrac{a_l-\epsilon}{2} \geq 0,\ \forall K \geq m \geq s+1 \end{aligned} \tag{4.6.1}$$

by the order $\lambda_1 \geq \cdots \geq \lambda_K$ hence $\varphi_1 \leq \cdots \leq \varphi_K$, and base on the fact that

$$\begin{aligned} o(\lambda_1^{\alpha_1}) o(\lambda_2^{\alpha_2}) &= o\left(\lambda_1^{\alpha_1+\alpha_2}\right)\ if\ \alpha_2 \geq 0 \\ o(\varphi_1^{\alpha_1}) o(\varphi_2^{\alpha_2}) &= o\left(\varphi_2^{\alpha_1+\alpha_2}\right)\ if\ \alpha_1 \geq 0 \end{aligned}$$

the order of $G(\mathbf{\Sigma})$ is

$$\begin{aligned} G(\mathbf{\Sigma}) = \quad & o\left(\varphi_r^{\left(\frac{N}{2}+\alpha-\epsilon\right)r - \frac{a_\rho^{'}+\epsilon}{2} N \sum_{j=1}^{r} P_N(D_j) - \sum_{q_l \leq r} \frac{a_l^{'}+\epsilon}{2}} \right) \cdot \\ & o\left(\lambda_{s+1}^{\frac{a_\rho-\epsilon}{2}\cdot N \sum_{j=s+1}^{K} P_N(D_j) - \left(\frac{N}{2}+\alpha+\epsilon\right)(K-s) + \sum_{q_l \geq s+1} \frac{a_l-\epsilon}{2}} \right) \end{aligned}$$

and it goes to zero.

Now we simplify the assumption (4.6.1). Since $\sum_{j=1}^{m} P_N(D_j) = P_N(S_m)$ and $\sum_{j=m}^{K} P_N(D_j) = 1 - P_N(S_{m-1})$, and r, s can take any value that satisfies $1 \leq r \leq s < K$, we end up with the following condition:

$$\begin{aligned} &\left(\tfrac{N}{2}+\alpha-\epsilon\right) d - \tfrac{a_\rho^{'}+\epsilon}{2} NP_N(S_d) - \textstyle\sum_{q_l \leq d} \tfrac{a_l^{'}+\epsilon}{2} \geq 0 \\ &\tfrac{a_\rho-\epsilon}{2} \cdot N\left(1-P_N(S_d)\right) - \left(\tfrac{N}{2}+\alpha+\epsilon\right)(K-d) + \textstyle\sum_{q_l \geq d+1} \tfrac{a_l-\epsilon}{2} \geq 0 \end{aligned}$$

for all $1 \leq d \leq K-1$.

Define sets $\omega = \{l|q_l \leq d\}$ and $\upsilon = \{l|q_l > d\}$, consider when $l \in \omega$, which means

$q_l \leq d$, by the definition of q_l, is equivalent to $\mathrm{range}\,(\mathbf{A}_l) \subseteq S_d$, similarly for $l \in \upsilon$, which means $q_l > d$, is equivalent to $\mathrm{range}\,(\mathbf{A}_l) \nsubseteq S_d$.

The condition should be valid for any $\mathbf{U}$ and $1 \leq d \leq K-1$, tidy up the expression and let $\epsilon \to 0$ results in: for any proper subspace S

$$P_N(S) \quad < \min\left\{1 - \frac{(N+2\alpha)(K-\dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}, \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}\right\}$$

where sets ω and υ are defined as $\omega = \{l|\mathrm{range}\,(\mathbf{A}_l) \subseteq S\}$, $\upsilon = \{l|\mathrm{range}\,(\mathbf{A}_l) \nsubseteq S\}$.

For the case $r = 0$, $1 \leq s < K$, which means no $\lambda \to +\infty$ and some, not all $\lambda \to 0$, following the same reasoning gives condition

$$P_N(S) < 1 - \frac{(N+2\alpha)(K - \dim(S)) - \sum_{l\in\upsilon} a_l}{a_\rho N}$$

and for $s = K$, $1 \leq r < K$, which means no $\lambda \to 0$ and some, not all $\lambda \to +\infty$, gives condition

$$P_N(S) < \frac{(N+2\alpha)\dim(S) - \sum_{l\in\omega} a_l^{'}}{a_\rho^{'} N}.$$

Notice that the above two conditions are included in the first one.

And finally under the scenario that all $\lambda \to +\infty$, it's easy to see

$$G(\mathbf{\Sigma}) = o\left(\varphi_K^{\left(\frac{N}{2}+\alpha-\epsilon\right)K - \frac{a_\rho^{'}}{2}N - \frac{1}{2}\sum_l\left(a_l^{'}+\epsilon\right)}\right)$$

goes to zero if $\left(-\frac{N}{2}-\alpha\right)K + \frac{a_\rho^{'}}{2}N + \frac{1}{2}\sum_l a_l^{'} < 0$, and under the case that all $\lambda \to 0$, $G(\mathbf{\Sigma}) = o\left(\lambda_1^{\frac{a_\rho}{2}\cdot N - \left(\frac{N}{2}+\alpha+\epsilon\right)K + \frac{1}{2}\sum_l(a_l-\epsilon)}\right)$ goes to zero if $\frac{a_\rho}{2}\cdot N - \left(\frac{N}{2}+\alpha\right)K + \frac{1}{2}\sum_l a_l > 0$.

Chapter 5

Regularized Robust Estimation of Mean and Covariance Matrix under Heavy-Tailed Distributions

In Chapter 4, we have pointed out two issues in the covariance estimation problem assuming a known mean, i.e., the existence of abnormal samples and high dimensionality. The shrinkage Tyler's estimator turned out to be a promising solution, since it is robust to outliers and is well defined when the number of samples is less than the dimension. In this chapter, we are going to consider the situation where the mean is unknown and the problem of how to estimate the mean and covariance matrix robustly in the high dimension regime.

5.1 Introduction

In the [15], Tyler proposed an M-estimator that estimates the normalized covariance matrix for samples drawn from an elliptical distribution with a known mean. To enjoy its good statistical properties such as the minimax robustness in the high dimension regime, the estimator has been adapted by shrinking it towards a given target matrix. Various versions of shrinkage Tyler's estimator have been proposed, and have been demonstrated to work effectively when the number of samples is relatively small compared to the dimension of the problem [2, 3, 19, 20, 101, 102].

The (shrinkage) Tyler's estimator assumes a known mean. In practice, there are scenarios in which both the mean and the covariance are required to be estimated from the samples. In

this case, to apply the estimator the mean has to be estimated in the first step. Nevertheless, ignoring the correlation of each component when estimating the mean may result in treating an outlier as a normal sample [7]. Moreover, the estimation error of the mean propagates to the covariance estimation since the samples will be centered at a wrong location. A way to design a robust estimator that estimates the mean and covariance jointly and performs well in the small sample scenario remains unclear.

This chapter focuses on addressing the problem of joint mean and covariance estimation, assuming *i.i.d.* samples from a heavy-tailed elliptical distribution when the sample size is small relative to the problem dimension. A regularized mean-covariance estimator is proposed defined as the minimizer of a penalized or regularized loss function. The loss function is chosen to be the negative log-likelihood function of the Cauchy distribution for its heavy-tail property that is capable of modeling abnormal observations as well as the tractability of analysis (note that the samples are not assumed to be drawn from a Cauchy distribution). The proposed estimator shrinks the mean and covariance matrix towards a prior target. Theoretical results including the existence and uniqueness of the estimator are proved under certain regularity conditions on the samples. The conditions indicate that the shrinkage estimator overcomes the drawback of the Cauchy maximum likelihood estimator without shrinkage, as it exists even when the number of samples is smaller than the problem dimension. Different numerical algorithms are provided and compared for the proposed shrinkage estimator based on the majorization-minimization framework with provable convergence.

This chapter is organized as follows: In Section 5.2, we propose a regularized robust estimator for the joint mean-covariance estimation problem with small sample size, and establish the conditions for the existence and uniqueness of the shrinkage estimator. Algorithms for the proposed estimator based on the majorization-minimization framework are derived in Section 5.3. Simulation studies on both the estimator performance and the algorithm convergence are conducted in Section 5.5. We conclude in Section 5.5.

5.2 Regularized Robust Estimator of Mean and Covariance Matrix

Assuming the samples are drawn independently from a continuous distribution f, then the condition $N > K+1$ guarantees the existence of the Cauchy MLE with probability one [88],

and a reliable estimation requires even more samples. However, in some applications the number of samples is relatively small compared to the number of parameters being estimated. In this case, the algorithm designed for the estimator may fail to converge. Motivated by the idea of [20], we regularize the Cauchy MLE by shrinking it to a prior target $(\mathbf{t}, \mathbf{T})$. The advantage of a shrinkage estimator is twofold: it provides a way to incorporate prior information into the estimator, and it helps stabilizing the estimator in the small sample situation.

We devise the following penalty function:

$$h\left(\boldsymbol{\mu}, \mathbf{R}\right) = \alpha\left(K \log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right) + \log\det\left(\mathbf{R}\right)\right) + \gamma \log\left(1 + \left(\boldsymbol{\mu} - \mathbf{t}\right)^T \mathbf{R}^{-1}\left(\boldsymbol{\mu} - \mathbf{t}\right)\right) \tag{5.2.1}$$

for some finite-valued nonnegative parameters α and γ. The following proposition shows that $(\mathbf{t}, r\mathbf{T})$ minimizes $h\left(\boldsymbol{\mu}, \mathbf{R}\right)$ with $r > 0$, therefore justifies that $h\left(\boldsymbol{\mu}, \mathbf{R}\right)$ is indeed a proper penalty function.

Proposition 5.1. *The minimizer of (5.2.1) on the set $\mathbb{R}^K \times \mathbb{S}_{++}^K$ is given by $(\mathbf{t}, r\mathbf{T})$ for $r > 0$.*

Proof. See Appendix 5.6.1. □

The scale-invariant property of the minimizers of $h\left(\boldsymbol{\mu}, \mathbf{R}\right)$ is important due to the following reason. Since asymptotically $(\boldsymbol{\mu}_0, c\mathbf{R}_0)$ minimizes $L\left(\boldsymbol{\mu}, \mathbf{R}\right)$ with c depending on the unknown f, a way of setting a shrinkage target $\mathbf{T}$ for $\mathbf{R}_0$ (or $\mathbf{R}_0/\mathrm{Tr}\left(\mathbf{R}_0\right)$) is by adding a penalty term $h\left(\boldsymbol{\mu}, \mathbf{R}\right)$ that is minimized for $\mathbf{R}$ proportional to $\mathbf{T}$ (by passing the value of c).

The regularized estimation problem is stated below with a shrinkage target $(\mathbf{t}, \mathbf{T})$ for $(\boldsymbol{\mu}, \mathbf{R})$:

$$\begin{aligned} \underset{\boldsymbol{\mu}, \mathbf{R} \succ \mathbf{0}}{\text{minimize}} \quad & \frac{(K+1)}{2} \sum_{i=1}^{N} \log\left(1 + \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T \mathbf{R}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right) \\ & + \alpha\left(K \log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right) + \log\det\left(\mathbf{R}\right)\right) \\ & + \gamma \log\left(1 + \left(\boldsymbol{\mu} - \mathbf{t}\right)^T \mathbf{R}^{-1}\left(\boldsymbol{\mu} - \mathbf{t}\right)\right) + \frac{N}{2} \log\det\left(\mathbf{R}\right). \end{aligned} \tag{5.2.2}$$

The shrinkage estimator $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$ with $\hat{\mathbf{R}} \succ \mathbf{0}$, which is defined as the solution of problem

(5.2.2), has to satisfy the following fixed-point equations:

$$\begin{aligned}\mathbf{R} =& \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}\\ &+\frac{2\gamma}{N+2\alpha}\frac{(\boldsymbol{\mu}-\mathbf{t})(\boldsymbol{\mu}-\mathbf{t})^T}{1+(\boldsymbol{\mu}-\mathbf{t})\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})}+\frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}(\mathbf{R}^{-1}\mathbf{T})}\\ \mathbf{0} =& (K+1)\sum_{i=1}^{N}\frac{(\mathbf{x}_i-\boldsymbol{\mu})}{1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}\\ &+2\gamma\frac{\mathbf{t}-\boldsymbol{\mu}}{1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})},\end{aligned} \tag{5.2.3}$$

which are derived by setting the gradient of the objective function, denoted by $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R})$, to zero[1]. Note that since α and γ are finite-valued, the effect of the penalty term on $\left(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}}\right)$ vanishes as $N\to+\infty$. As a result, $\left(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}}/\mathrm{Tr}\left(\hat{\mathbf{R}}\right)\right)\to(\boldsymbol{\mu}_0,\mathbf{R}_0/\mathrm{Tr}(\mathbf{R}_0))$ in probability asymptotically.

By defining $\bar{\mathbf{x}}_i=[\mathbf{x}_i;1]$, $\bar{\mathbf{t}}=[\mathbf{t};1]$, and using the reparametrization (3.2.4), problem (5.2.2) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{\Sigma}\succ\mathbf{0}}{\text{minimize}}\quad & \left(\frac{N}{2}+\alpha\right)\log\det(\boldsymbol{\Sigma})+\frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)\\ &+\alpha K\log\left(\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)\right)+\gamma\log\left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right)\\ \text{subject to}\quad & \Sigma_{K+1,K+1}=1\end{aligned} \tag{5.2.4}$$

with $\mathbf{S}$ being a selection matrix defined as $\mathbf{S}=\begin{bmatrix}\mathbf{I}_K\\ \mathbf{0}_{1\times K}\end{bmatrix}$. Denote the objective function by $L^{\text{shrink}}(\boldsymbol{\Sigma})$. A sufficient condition for the existence of the estimator $\left(\hat{\boldsymbol{\mu}},\hat{\mathbf{R}}\right)$ is stated in the following theorem.

Theorem 5.1. *For the loss function*

$$\begin{aligned}L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R}) =& \frac{N}{2}\log\det(\mathbf{R})+\frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(1+(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)\\ &+\alpha\left(K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det(\mathbf{R})\right)+\gamma\log\left(1+(\boldsymbol{\mu}-\mathbf{t})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{t})\right)\end{aligned} \tag{5.2.5}$$

defined on $\mathbb{R}^K\times\mathbb{S}_{++}^K$*, under the assumption that* $\mathbf{T}$ *is full rank, a minimum of* $L^{\text{shrink}}(\boldsymbol{\mu},\mathbf{R})$

[1]The solutions of equation (5.2.3) are stationary points of problem (5.2.2), which contains the global minimum

exists under the following condition:

for any hyperplane $H \subset \mathbb{R}^K$ *with dimension* $0 \leq \dim(H) < K$*, if* H *contains* $\mathbf{t}$*,*

$$P_N(H) \quad < \quad \frac{(2\alpha + N)\dim(H) + N}{(K+1)N},$$

and if H *does not contain* $\mathbf{t}$*,*

$$P_N(H) \quad < \quad \frac{(2\alpha + N)\dim(H) + N + 2\gamma}{(K+1)N}.$$

Proof. See Appendix 5.6.2. □

Notice that setting α and γ to zero recovers the condition without regularization, which is $P_N(H) < \frac{\dim(H)+1}{K+1}$ [88].

To make the condition practically computable, we provide the following corollary.

Corollary 5.1. *Assume that the population distribution* f *is continuous, then the estimator* $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$ *exists for* $N > 1$ *if* $\alpha, \gamma \geq 0$ *and either of the following conditions is satisfied:*

(i) if $\gamma > \gamma_1$*, then* $\alpha > \alpha_1$*,*

(ii) if $\gamma_2 < \gamma \leq \gamma_1$*, then* $\alpha > \alpha_2(\gamma)$*,*

where

$$\alpha_1 = \frac{1}{2}(K - N),$$
$$\alpha_2(\gamma) = \frac{1}{2}\left(K+1-N-\frac{2\gamma + N - K - 1}{N-1}\right),$$

and $\gamma_1 = \frac{1}{2}K$*,* $\gamma_2 = \frac{1}{2}(K + 1 - N)$*.*

Proof. Notice that since the distribution is continuous, then with probability one, a d-dimensional hyperplane at most touches $d + 1$ points. The condition in Theorem 5.1 can be simplified as follows:

$$\begin{cases} \dfrac{\min\{d, N\}}{N} < \dfrac{(2\alpha + N)d + N}{(K+1)N} \\ \dfrac{\min\{d+1, N\}}{N} < \dfrac{(2\alpha + N)d + 2\gamma + N}{(K+1)N} \end{cases}, \forall\, 0 \leq d \leq K-1,$$

which is equivalent to

$$\begin{cases} (K + 1 - 2\alpha - N)d < N, \forall\, 0 \leq d \leq \min\{K-1, N\}, \\ (K + 1 - 2\alpha - N)d < 2\gamma + N - K - 1, \forall\, 0 \leq d \leq \min\{K-1, N-1\}. \end{cases} \tag{5.2.6}$$

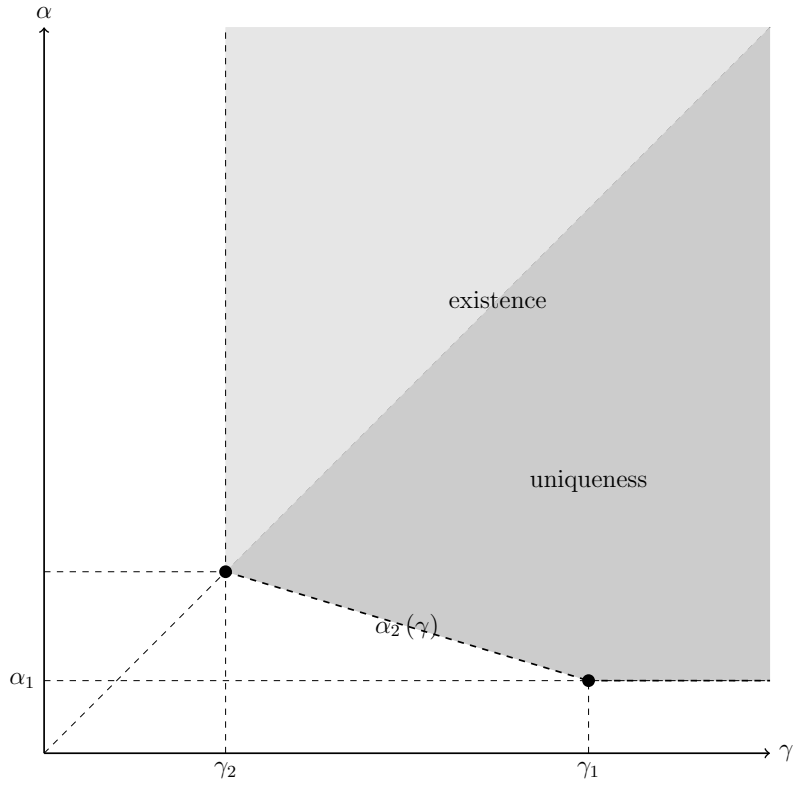


Figure 5.1: Values that the regularization parameters α and γ can take for the existence and uniqueness of the shrinkage estimator.

Since both α and γ are nonnegative, the conditions above is satisfied if $N \geq K + 1$. Under the case that $N \leq K$, some algebraic calculation reveals that (5.2.6) is equivalent to the statement of the corollary. □

The feasible region of (α, γ) is shown pictorially in Fig. 1. The condition in Corollary 5.1 implies the tradeoff between the regularization parameters α and γ. Specifically, since $\alpha_2(\gamma)$ is decreasing in γ, the condition indicates that when the confidence on the prior information of $\boldsymbol{\mu}$ gets weaker, which corresponds to a smaller value of γ, the regularization for $\mathbf{R}$ should be stronger. For the special case that $\alpha = \gamma$, the condition reduces to $\alpha > \frac{1}{2}(K + 1 - N)$, or equivalently $N > K + 1 - 2\alpha$. The lower bound on the number of samples is decreased by 2α as a result of regularization.

Now that we have established the existence condition for the shrinkage estimator $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$, in the following theorem, we are going to show that the estimator is unique when $\gamma \geq \alpha$.

Theorem 5.2. *Under the regularity condition in Theorem 5.1,* $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$ *is unique when* $\gamma \geq \alpha$.

Proof. See Appendix 5.6.3. □

5.3 Algorithms

In this section, we derive four algorithms for the regularized estimator based on the (block) MM framework, and in Section 5.4 we study the convergence speed of the proposed algorithms numerically. With a slight abuse of notation, $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ refers to the initial point of the algorithm in this section.

Recall that the optimization problem takes the form

$$\begin{aligned}\underset{\boldsymbol{\mu},\mathbf{R}\succ\mathbf{0}}{\text{minimize}} \quad & \frac{N}{2}\log\det\left(\mathbf{R}\right) \\ & + \frac{(K+1)}{2}\sum_{i=1}^{N}\log\left(1+\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right) \\ & + \alpha\left(K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{R}\right)\right) \\ & + \gamma\log\left(1+\left(\boldsymbol{\mu}-\mathbf{t}\right)^T\mathbf{R}^{-1}\left(\boldsymbol{\mu}-\mathbf{t}\right)\right).\end{aligned} \tag{5.3.1}$$

Before introducing the algorithms, we first state the following lemma, which is needed for proving the convergence of the algorithms.

Lemma 5.1. *Given any initial point $(\boldsymbol{\mu}_0, \mathbf{R}_0)$ with $\mathbf{R}_0 \succ \mathbf{0}$, the level set $\mathcal{X}^0 = \{(\boldsymbol{\mu}, \mathbf{R}) \,|L^{\textit{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \leq L^{\textit{shrink}}(\boldsymbol{\mu}_0, \mathbf{R}_0)\}$ is compact.*

Proof. Under the conditions stated in Theorem 5.1, a unique minimizer $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$ of $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ exists. Observe that

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{\Sigma}\right) &= \sup_{\|\mathbf{y}\|=1}\mathbf{y}^T\boldsymbol{\Sigma}\mathbf{y} \\ &= \sup_{\|\tilde{\mathbf{y}}\|^2+y^2=1}\begin{bmatrix}\tilde{\mathbf{y}}^T & y\end{bmatrix}\begin{bmatrix}\mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{y}} \\ y\end{bmatrix} \\ &= \sup_{\|\tilde{\mathbf{y}}\|^2+y^2=1}\left\{\tilde{\mathbf{y}}^T\mathbf{R}\tilde{\mathbf{y}}+\left(\tilde{\mathbf{y}}^T\boldsymbol{\mu}+y\right)^2\right\} \\ &\geq \sup_{\|\tilde{\mathbf{y}}\|^2=1}\left\{\tilde{\mathbf{y}}^T\left(\mathbf{R}+\boldsymbol{\mu}\boldsymbol{\mu}^T\right)\tilde{\mathbf{y}}\right\}.\end{aligned}$$

Suppose $\lambda_{\max}(\mathbf{R}) \to +\infty$, then $\lambda_{\max}(\boldsymbol{\Sigma}) \to +\infty$. By Theorem 5.1, it is known that $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) = L^{\text{shrink}}(\boldsymbol{\Sigma}) \to +\infty$ as $\lambda_{\max}(\boldsymbol{\Sigma}) \to +\infty$, which contradicts the fact that on $\mathcal{X}^0$, $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ is bounded above. Therefore $\mathbf{R}$ is bounded. Now suppose $\boldsymbol{\mu}$ is unbounded, set $\tilde{\mathbf{y}} = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}$ and $y = 0$ we have $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \|\boldsymbol{\mu}\|^2 \to +\infty$. Therefore $\mathcal{X}^0$ is bounded. The continuity of $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ implies $\mathcal{X}^0$ is closed. Hence $\mathcal{X}^0$ is compact. □

Algorithm 5.1 Majorization-Minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.
2) Iterate

$$\begin{aligned}\boldsymbol{\mu}_{t+1} =& \frac{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\mathbf{x}_i + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\mathbf{t}}{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right) + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)} \\ \mathbf{R}_{t+1} =& \frac{K+1}{N+2\alpha}\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)^T \\ &+ \frac{2\gamma}{N+2\alpha} w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{t} - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{t} - \boldsymbol{\mu}_{t+1}\right)^T \\ &+ \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\end{aligned} \tag{5.3.4}$$

with $w_i\left(\boldsymbol{\mu}, \mathbf{R}\right)$ and $w_{\mathbf{t}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ given in (5.3.3) until convergence.

5.3.1 Majorization-Minimization

By the concavity of $\log\left(\cdot\right)$, at point $\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ function (5.2.5) is majorized by

$$\begin{aligned}& L\left(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right) \\ =& \frac{K+1}{2}\sum w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right) \\ &+ \gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{t} - \boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{t} - \boldsymbol{\mu}\right) \\ &+ \left(\frac{N}{2} + \alpha\right)\log\det\left(\mathbf{R}\right) + \alpha K\frac{\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} + \mathrm{const.}\end{aligned} \tag{5.3.2}$$

with weights

$$\begin{aligned}w_i\left(\boldsymbol{\mu}, \mathbf{R}\right) =& \frac{1}{1 + \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)} \\ w_{\mathbf{t}}\left(\boldsymbol{\mu}, \mathbf{R}\right) =& \frac{1}{1 + \left(\mathbf{t} - \boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{t} - \boldsymbol{\mu}\right)}.\end{aligned} \tag{5.3.3}$$

Setting the gradient of $L\left(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ to zero leads to the update equations in Algorithm 5.1.

Lemma 5.2. *The pair* $\left(\boldsymbol{\mu}_{t+1}, \mathbf{R}_{t+1}\right)$ *given by equation (5.3.4) uniquely minimizes the surrogate function (5.3.2).*

Proof. Since the surrogate function $L\left(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ upper bounds the cost function $L^{\text{shrink}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ globally, the minimum of $L\left(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ exists with $\mathbf{R} \succ \mathbf{0}$ and $\boldsymbol{\mu}$ being finite. Observe that $L\left(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ is the negative log-likelihood function of a Gaussian distribution and has a unique stationary point $\left(\boldsymbol{\mu}_{t+1}, \mathbf{R}_{t+1}\right)$ on $\mathbb{R}^K \times \mathbb{S}_+^K$, it has to be the minimum. □

Proposition 5.2. *The sequence* $\{\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\}$ *generated by Algorithm 5.1 converges to*

(i) the set of stationary points of problem (5.3.1) if $\alpha > \gamma$;

(ii) the global minimizer of problem (5.3.1) if $\alpha \leq \gamma$.

Proof. By Lemma 5.1 we have that the initial level set $\mathcal{X}^0$ is compact. Furthermore, by Lemma 5.2, $\left(\boldsymbol{\mu}_{t+1}, \mathbf{R}_{t+1}\right)$ uniquely minimizes the surrogate function $L\left(\boldsymbol{\mu}, \mathbf{R} | \boldsymbol{\mu}_t, \mathbf{R}_t\right)$. Hence the sequence $\{(\boldsymbol{\mu}_t, \mathbf{R}_t)\}$ converges to the set of stationary points of $L^{\text{shrink}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ [53]. It has been proved in Theorem 5.2 that the stationary point of problem (5.3.1) is unique, and it is the global minimum when $\alpha \leq \gamma$. Therefore in this case $\{(\boldsymbol{\mu}_t, \mathbf{R}_t)\}$ converges to the global minimizer of problem (5.3.1). □

5.3.2 Block Majorization-Minimization

Instead of upperbounding the whole function at point $(\boldsymbol{\mu}_t, \mathbf{R}_t)$, majorization can also be applied blockwise. Specifically, an upperbound for $L^{\text{shrink}}\left(\boldsymbol{\mu}_t, \mathbf{R}\right)$ can be obtained as:

$$\begin{aligned} & L\left(\mathbf{R} | \boldsymbol{\mu}_t, \mathbf{R}_t\right) \\ =& \frac{K+1}{2} \sum w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_t\right)^T \mathbf{R}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_t\right) \\ & + \alpha K \frac{\operatorname{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)}{\operatorname{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} + \gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{t} - \boldsymbol{\mu}_t\right)^T \mathbf{R}^{-1}\left(\mathbf{t} - \boldsymbol{\mu}_t\right) \\ & + \left(\frac{N}{2} + \alpha\right) \log\det\left(\mathbf{R}\right) + \text{const.} \end{aligned} \tag{5.3.5}$$

with the value of $\boldsymbol{\mu}$ fixed as $\boldsymbol{\mu}_t$, which leads to the update equation (5.3.7) in Algorithm 5.2 for $\mathbf{R}$.

Then we fix the value of $\mathbf{R}$ as $\mathbf{R}_{t+1}$ and get an upperbound for $L^{\text{shrink}}\left(\boldsymbol{\mu}, \mathbf{R}_{t+1}\right)$ as follows:

$$\begin{aligned} & L\left(\boldsymbol{\mu} | \boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right) \\ =& \frac{K+1}{2} \sum w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_t\right)^T \mathbf{R}_{t+1}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_t\right) \\ & + \gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)\left(\mathbf{t} - \boldsymbol{\mu}_t\right)^T \mathbf{R}_{t+1}^{-1}\left(\mathbf{t} - \boldsymbol{\mu}_t\right) + \text{const.} \end{aligned} \tag{5.3.6}$$

which leads to the update equation (5.3.8) in Algorithm 5.2 for $\boldsymbol{\mu}$.

Notice that the update order can be reversed, i.e., first fix $\mathbf{R}_t$ and get $\boldsymbol{\mu}_{t+1}$, then fix $\boldsymbol{\mu}_{t+1}$ and get $\mathbf{R}_{t+1}$. This leads to similar iterations to Algorithm 5.2.

Proposition 5.3. *The sequences $\{(\boldsymbol{\mu}_t, \mathbf{R}_t)\}$ generated by Algorithm 5.2 converge to*

(i) the set of stationary points of problem (5.3.1) if $\alpha > \gamma$;

Algorithm 5.2 Block majorization-minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.
2) Iterate

$$\begin{aligned}\mathbf{R}_{t+1} =& \frac{K+1}{N+2\alpha}\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_t\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_t\right)^T \\ &+ \frac{2\gamma}{N+2\alpha} w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{t}-\boldsymbol{\mu}_{\mathbf{t}}\right)\left(\mathbf{t}-\boldsymbol{\mu}_{\mathbf{t}}\right)^T \\ &+ \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\operatorname{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\end{aligned} \tag{5.3.7}$$

$$\boldsymbol{\mu}_{t+1} = \frac{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)\mathbf{x}_i + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)\mathbf{t}}{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right) + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)} \tag{5.3.8}$$

with $w_i\left(\boldsymbol{\mu}, \mathbf{R}\right)$ and $w_{\mathbf{t}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ given in (5.3.3) until convergence.

(ii) the global minimizer of problem (5.3.1) if $\alpha \leq \gamma$.

Proof. We verify the sufficient condition (2) in Theorem 2.2 for block MM presented in Chapter 2. First, the level set $\mathcal{X}^0$ is compact by Lemma 5.1. Second, with fixed $\boldsymbol{\mu}_t$, the surrogate function $L\left(\mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$ upperbounds $L^{\text{shrink}}\left(\boldsymbol{\mu}_t, \mathbf{R}\right)$, therefore $L\left(\mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right) \to +\infty$ when $\mathbf{R}$ goes to the boundary of $\mathbb{S}_{++}^K$. This implies that $\mathbf{R}_{t+1}$ given by (5.3.7), which satisfies the stationary condition of $L\left(\mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$, is the unique minimizer of $L\left(\mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t\right)$. Similarly we can prove that $\boldsymbol{\mu}_{t+1}$ given by (5.3.8) is the unique minimizer of $L\left(\boldsymbol{\mu}|\boldsymbol{\mu}_t, \mathbf{R}_{t+1}\right)$. Therefore $\{(\boldsymbol{\mu}_t, \mathbf{R}_t)\}$ converges to the set of stationary points of problem (5.3.1), which satisfies the fixed-point equations (5.2.3). In the case that $\alpha \leq \gamma$, equation (5.2.3) has a unique solution (cf. Th. 5.2), therefore, the unique stationary point has to be the global minimum. □

5.3.3 Special Case for $\alpha = \gamma$

In this subsection we provide an algorithm for the special case $\alpha = \gamma$, which is simpler than the previously described ones. It has been proved in Theorem 5.2 that when $\alpha = \gamma$, the objective function $L^{\text{shrink}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ is scale-invariant and has a unique minimizer on $\mathbb{R}^K \times \mathbb{S}_+^K$ up to a positive scale factor. The $\boldsymbol{\Sigma}$ reparametrization (3.2.4) yields the following equivalent

Algorithm 5.3 Algorithm for $\alpha = \gamma$

1) Initialize $\mathbf{\Sigma}_0$ as an arbitrary positive definite matrix.
2) Iterate

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{t+1} =& \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{\bar{\mathbf{x}}_i^T\mathbf{\Sigma}_t^{-1}\bar{\mathbf{x}}_i} \\ &+ \frac{2\alpha}{N+2\alpha}\left(\frac{K\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}^T\mathbf{\Sigma}_t^{-1}\mathbf{S}\mathbf{T}\right)} + \frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{\bar{\mathbf{t}}^T\mathbf{\Sigma}_t^{-1}\bar{\mathbf{t}}}\right) \\ \mathbf{\Sigma}_{t+1} =& \tilde{\mathbf{\Sigma}}_{t+1}/\left(\tilde{\mathbf{\Sigma}}_{t+1}\right)_{K+1,K+1}\end{aligned}$$

until convergence.

optimization problem:

$$\begin{aligned}\underset{\mathbf{\Sigma}\succcurlyeq\mathbf{0}}{\text{minimize}} \quad & \left(\frac{N}{2}+\alpha\right)\log\det\left(\mathbf{\Sigma}\right) + \frac{K+1}{2}\sum_{i=1}^{N}\log\left(\bar{\mathbf{x}}_i^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}_i\right) \\ & + \alpha K\log\left(\mathrm{Tr}\left(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)\right) + \alpha\log\left(\bar{\mathbf{t}}^T\mathbf{\Sigma}^{-1}\bar{\mathbf{t}}\right) \\ \text{subject to} \quad & \Sigma_{K+1,K+1} = 1.\end{aligned} \tag{5.3.9}$$

For this special case, in addition to Algorithm 5.1 and 5.2, we propose a more efficient one as described in Algorithm 5.3. The convergence follows the same reasoning as Proposition 17 in [102].

5.3.4 Accelerated Majorization-Minimization

Recall it is proved in Theorem 5.2 that the stationary condition (5.2.3) can be embedded in (5.6.4), which is equivalent to the equation system (5.6.5)-(5.6.7).

Since ζ can be solved as $\zeta = \frac{N+2\gamma}{N+2\alpha}$, we can substitute the value of ζ into (5.6.6) and eliminate (5.6.7). This leads to the iterations described in Algorithm 5.4. Compared to Algorithm 5.1, the update equation of $\boldsymbol{\mu}_{t+1}$ in Algorithm 5.4 is unchanged, while that of $\mathbf{R}_{t+1}$ is multiplied by a factor of β_t. In the case that $\alpha = \gamma$, Algorithm 5.4 and Algorithm 5.3 are identical.

Algorithm 5.4 Accelerated majorization-minimization

1) Initialize $\mathbf{R}_0$ as an arbitrary positive definite matrix, and $\boldsymbol{\mu}_0$ as an arbitrary vector.
2) Iterate

$$\begin{aligned}\boldsymbol{\mu}_{t+1} =& \frac{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\mathbf{x}_i + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\mathbf{t}}{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right) + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)} \\ \mathbf{R}_{t+1} =& \beta_t \left\{ \frac{K+1}{N+2\alpha}\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)^T \right. \\ &+ \frac{2\gamma}{N+2\alpha} w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)\left(\mathbf{t} - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{t} - \boldsymbol{\mu}_{t+1}\right)^T \\ &\left. + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)} \right\}\end{aligned} \tag{5.3.10}$$

with $w_i\left(\boldsymbol{\mu}, \mathbf{R}\right)$ and $w_{\mathbf{t}}\left(\boldsymbol{\mu}, \mathbf{R}\right)$ given in (5.3.3) and

$$\beta_t = \frac{N+2\gamma}{(K+1)\sum_{i=1}^{N} w_i\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right) + 2\gamma w_{\mathbf{t}}\left(\boldsymbol{\mu}_t, \mathbf{R}_t\right)},$$

until convergence.

5.4 Numerical Results

In this section, we conduct a simulation study of the performance of the proposed shrinkage estimator and the convergence of the numerical algorithms presented in Sec. IV. The overall estimation error of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$ is measured by the symmetrized KL divergence

$$\begin{aligned}\mathrm{err}\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right) = \frac{1}{2}\mathbb{E}\left\{ D_{KL}\left(\mathcal{N}\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right) \| \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{R}_0\right)\right)\right. \\ \left. + D_{KL}\left(\mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{R}_0\right) \middle\| \mathcal{N}\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)\right)\right\},\end{aligned}$$

where all covariance matrices are normalized by their traces. The expected error of the estimator is approximated by averaging 200 independent simulations with randomly generated data sets following the same underlying distribution. The measure is chosen to account for the estimation error of $\boldsymbol{\mu}$ and $\mathbf{R}$ simultaneously. Compared to adding together the NMSEs of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$, defined as $\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|/\|\boldsymbol{\mu}_0\|$ and $\mathbb{E}\|\hat{\mathbf{R}} - \mathbf{R}_0\|/\|\mathbf{R}_0\|$, respectively, the KL divergence error has a geometric interpretation. In practice, the error measure should be chosen based on the risk to be minimized. In all the following simulations, if not specified, the scatter parameter $\mathbf{R}_0\left(\beta\right)$ is set to be a Toeplitz matrix of the form

$$\left(\mathbf{R}_0\right)_{ij} = \beta^{|i-j|}$$

with $K = 100$, $\beta = 0.8$, and the distribution center $\boldsymbol{\mu}_0$ is fixed to be $\mathbf{1}$.

The first simulation compares the estimation error of the sample average estimator, the Cauchy MLE, MLE, and two plug-in estimators with samples drawn from a Student's t-distribution with degree of freedom parameter ν, denoted as $t_\nu(\boldsymbol{\mu}_0, \mathbf{R}_0)$. The plug-in estimator with legend "median" is obtained by first estimating $\boldsymbol{\mu}$ by sample median and then estimating $\mathbf{R}$ by NSCM; and that with legend "20% trimmed mean" is obtained by replacing the sample median by 20% trimmed mean. In this simulation, ν varies from 1 to 30 and the number of samples N is set to be 120. Fig. 5.2 plots the NMSE of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$, where the expected value is obtained by averaging 200 independent sample realizations. We see that In the case that the tail of the underlying distribution is not as heavy as the Cauchy distribution, which corresponds to a large value of ν, the estimation error incurred by fitting the samples to a Cauchy distribution is small compared to the MLE that assumes perfect knowledge of ν. For this reason, the negative log-likelihood function of Cauchy distribution is an acceptable choice as the loss function of estimating the parameters $(\boldsymbol{\mu}, \mathbf{R})$ of an elliptical distribution. Compared to plug-in estimators, the Cauchy MLE achieves a smaller NMSE for the estimation of $\boldsymbol{\mu}$, but a larger NMSE for the estimation of $\mathbf{R}$. This observation is consistent with Fig 4.3, where NSCM also outperform Tyler's estimator in the small sample regime.

The second simulation shows the performance of the proposed shrinkage estimator with a small sample size. The estimators that we consider are the sample average, the Cauchy MLE, the plug-in estimator (first estimate the location by sample mean or sample median then estimate covariance by shrinkage Tyler's estimator [102] with the estimated mean), and the proposed shrinkage Cauchy MLE. As for the tuning parameter of the shrinkage estimators, we define $\rho(\alpha) = \frac{N}{N+2\alpha}$ and search for the $\rho^\star$ on the grid $\{0.1, 0.2. \ldots, 1\}$ that yields the shrinkage estimator with the smallest estimation error as proposed in [20], so as to eliminate the effect of parameter tuning that is important but not the focus of this thesis. The shrinkage target $\mathbf{T}$ is set to be $\mathbf{I}/K$ motivated by the idea of diagonal loading [93] and $\mathbf{t}$ is set to be the sample mean and the sample median. Notice that in both cases the shrinkage target does not depend on any prior knowledge of the true parameter. Fig. 5.2 shows that for Gaussian distributed samples, heavy-tailed samples $\mathbf{x}_i \sim t_3(\boldsymbol{\mu}_0, \mathbf{R}_0)$ and elliptically distributed samples $\mathbf{x}_i \sim \boldsymbol{\mu} + \sqrt{\tau}\mathbf{u}$, where $\tau \sim \chi^2$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, the proposed shrinkage estimator with the shrinkage target being the sample mean achieves the smallest estimation error. The proposed shrinkage estimator is robust to the class of outliers that are distributed far away from the

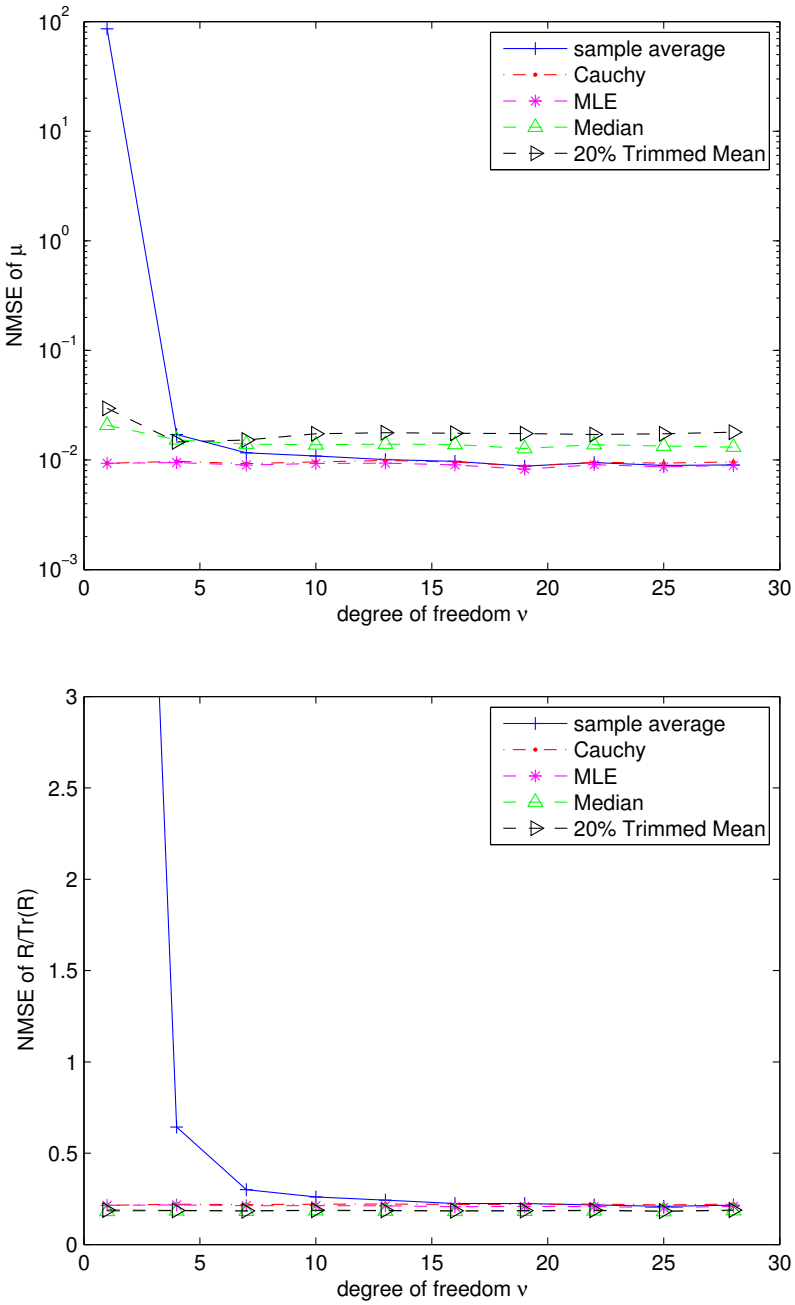


Figure 5.2: NMSE of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{R}}$ with $N = 120$ 100-dimensional samples drawn from a Student's t-distribution.

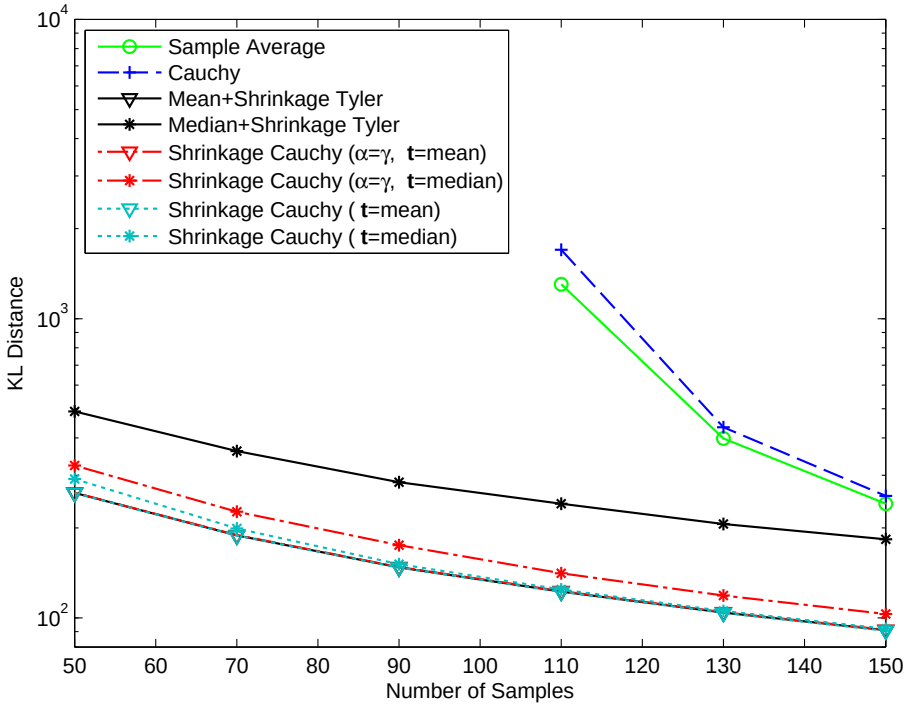
	$\beta = 0.1$		$\beta = 0.3$		$\beta = 0.5$		$\beta = 0.7$	
	$N = 80$	$N = 150$	$N = 80$	$N = 150$	$N = 80$	$N = 150$	$N = 80$	$N = 150$
$t = 0.1$	153.47	86.01	138.48	80.56	125.42	73.76	110.48	63.24
$t = 0.3$	148.87	84.25	133.98	79.00	121.18	72.22	105.22	61.58
$t = 0.5$	138.67	80.36	124.42	75.46	112.06	68.72	94.33	57.92
$t = 0.7$	111.14	69.47	100.57	64.90	87.86	58.54	68.90	47.63
$t = 0.9$	52.78	37.01	45.20	32.47	35.42	27.85	20.09	17.66

Table 5.1: Sensitivity analysis: averaged estimation error of the proposed shrinkage estimator for different values of $(\mathbf{t}, \mathbf{T})$.

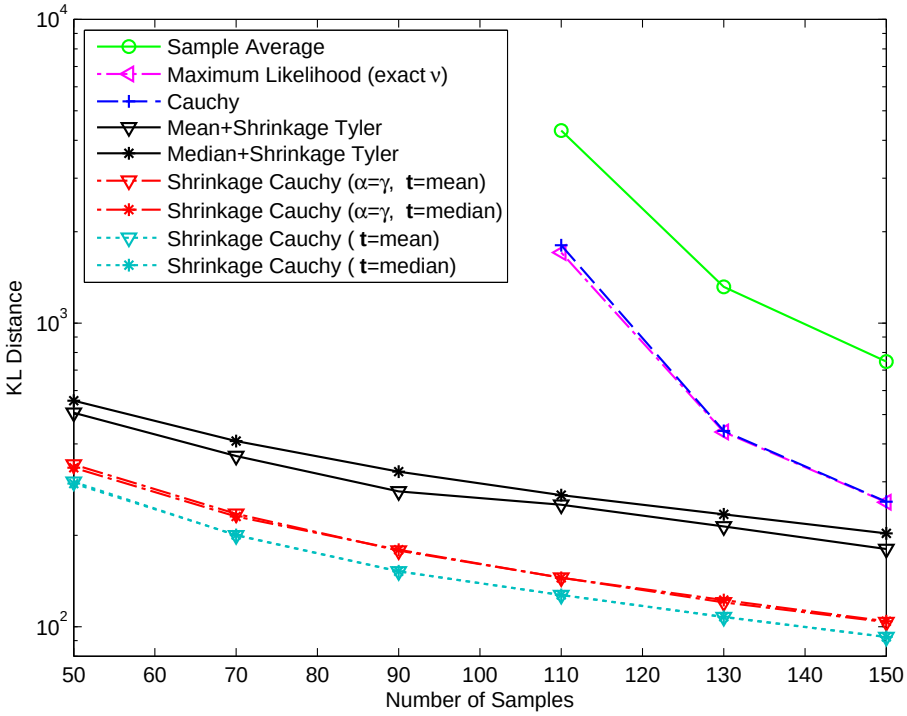
"good" samples. A plot in Fig. 5.2 (d) shows the estimation error versus the percentage of outliers with normal samples $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{R}_0)$ and outliers $\mathbf{x}_{\text{outlier}} \sim \boldsymbol{\mu}_0 + r\mathbf{s}$, where $\mathbf{s}$ is uniformly distributed on a $(K-1)$-dimensional sphere and r is uniformly distributed on the interval $[2l, 2l+1]$, l is set to be $l \triangleq \max\{\|\mathbf{x}_i\|_2\}$. The total number of samples is $N = 120$.

The third simulation analyzes the sensitivity of the shrinkage estimator to the shrinkage target $(\mathbf{t}, \mathbf{T})$. The underlying distribution is chosen to be $t_3(\boldsymbol{\mu}_0, \mathbf{R}_0)$. The estimation error of the shrinkage estimator with $\mathbf{t} = t\mathbf{1}$, $t \in \{0.1, 0.3, \ldots, 0.9\}$, and $\mathbf{T}$ being a Toeplitz matrix $\mathbf{R}_0(\beta)$, $\beta \in \{0.1, 0.3, 0.5, 0.7\}$ is listed in Table 5.1 for $N = 80$ and 150 respectively. The table indicates that the estimation error decreases as $(\mathbf{t}, \mathbf{T})$ gets closer to the true parameter $(\boldsymbol{\mu}_0, \mathbf{R}_0)$. For $N = 150$, the estimation error of the MLE is 260.84, which turns out to be much larger than the maximum error for the $N = 150$ case in Table 5.1. The reason is that although $\mathbf{t} = 0.1 \times \mathbf{1}$ is far away from $\boldsymbol{\mu}_0$, the regularization parameter γ is small so that $\boldsymbol{\mu}$ is estimated majorly based on the samples, and $\mathbf{T}$ is close to the identity matrix $\mathbf{I}$ when $\beta = 0.1$, which still helps in improving the estimation accuracy by shrinking the eigenvalues of $\hat{\mathbf{R}}$ towards the center in the small sample regime. To conclude, one can expect that a more informative prior ($(\mathbf{t}, \mathbf{T})$ is closer to $(\boldsymbol{\mu}_0, \mathbf{R}_0)$) leading to a more accurate estimator $\left(\hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$. Even if the prior on the parameters is completely wrong, the shrinkage estimator performs no worse than the Cauchy MLE provided that the tuning parameters α and γ are properly selected.

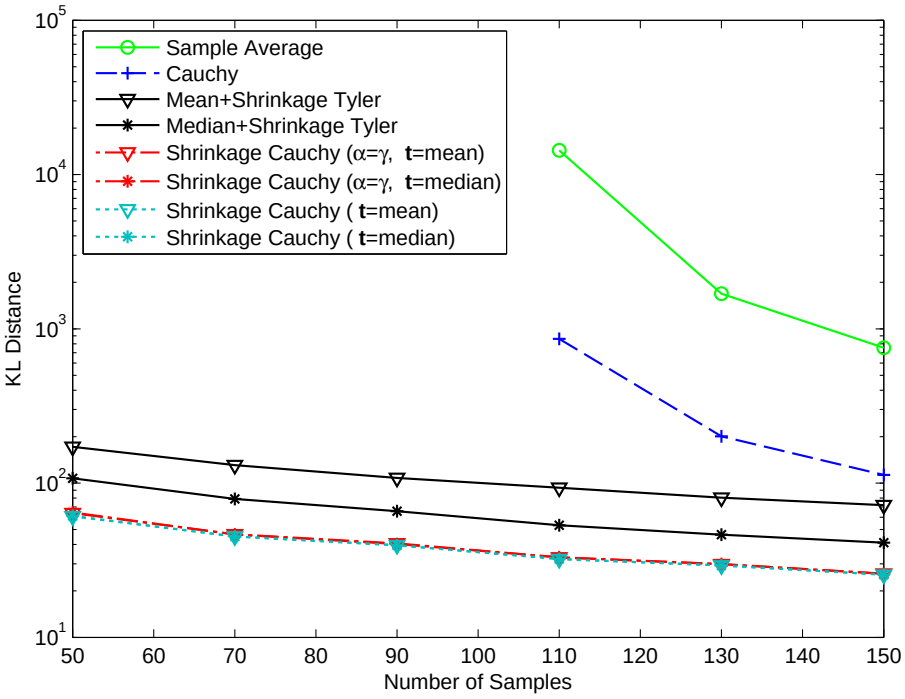
Fig. 5.3 and Table 5.2 demonstrate the convergence of the algorithms provided in Section 5.3. The convergence criterion is set to be $\|\mathbf{R}_t - \mathbf{R}_{t+1}\|_F < 10^{-5}$ and $\left\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1}\right\|_F < 10^{-5}$. Fig. 5.3 plots the evolution curve of the objective value versus the number of iterations. The parameters are set to be $N = 80$, $K = 100$, $\mathbf{t} = 0.9 \times \mathbf{1}$, $\mathbf{T} = \mathbf{I}/K$, $\gamma = 360$ and $\alpha = 160$.



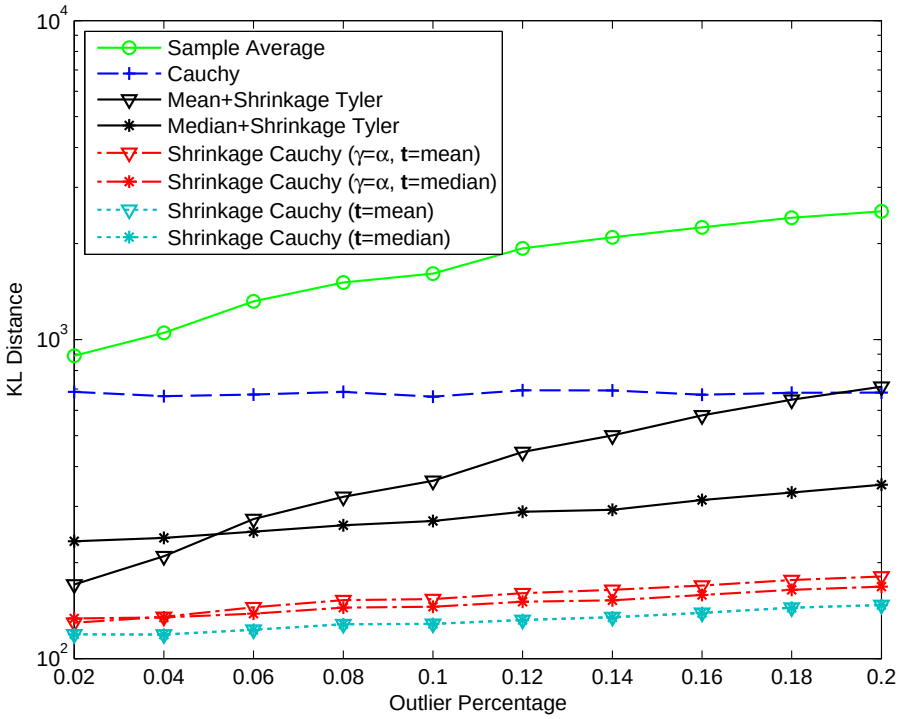
(a) Gaussian distributed samples



(b) Student's t-distributed samples ($\nu = 3$)



(c) Elliptically distributed samples ($\mathbf{x}_i \sim \sqrt{\tau}\mathbf{u}$)



(d) Gaussian distributed samples with outlier contamination

Figure 5.2: Performance comparison for different estimators.

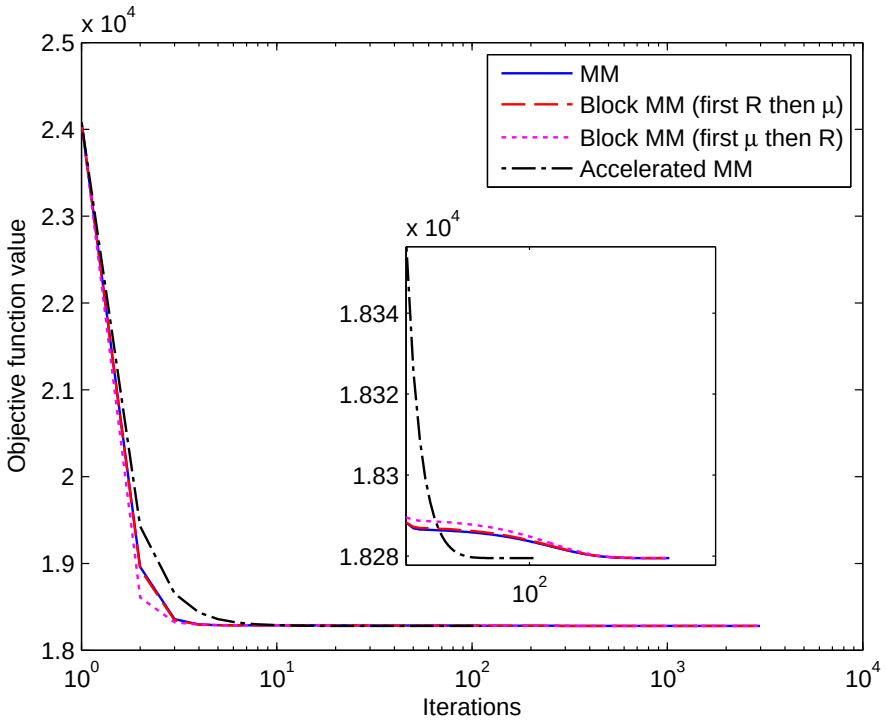


Figure 5.3: Convergence comparison for algorithms in Sec. IV.

While the computational cost per iteration of all the algorithms is roughly the same, it can be seen that the accelerated algorithm requires far fewer iterations than the MM algorithms (MM and block MM). Table 5.2 lists the average number of iterations that each of the algorithms require to converge as α and K changes. In this simulation, α and γ are set to be equal, consequently Algorithm 5.3 and Algorithm 5.4 turn out to be the same. The ratio N/K is fixed to be 1.2. The data indicates that the accelerated MM algorithm not only converges much faster, but also is not very sensitive to the problem dimension. For these reasons, the accelerated MM is recommended for practical applications.

	MM			Block MM (first update $\mathbf{R}$ then $\boldsymbol{\mu}$)			Block MM (first update $\boldsymbol{\mu}$ then $\mathbf{R}$)			Accelerated MM		
	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$	$K = 30$	$K = 50$	$K = 100$
$\rho = 0.2$	1015.07	1596.16	2944.42	1021.06	1604.055	2955.325	1028.85	1622.74	3007.825	55.215	56.215	57.93
$\rho = 0.5$	419.505	666.975	1250.175	421.045	668.91	1252.775	426.615	680.485	1281.28	23.995	23.94	24.06
$\rho = 0.8$	309.73	484.775	907.175	311.96	487.66	911.045	321.175	507.835	963.06	40.51	37.4	34.83

Table 5.2: Average number of iterations required for algorithms, i.e., MM, block MM and accelerated MM, to converge.

Finally, we test the performance of the shrinkage estimators on a real financial data set for minimum variance portfolio problem.

Data construction: We choose weekly close prices p_t from Jan 1, 2010 to June 8, 2014, 230 weeks in total, of $K = 40$ stocks selected from the S&P 500 index components provided by Yahoo Finance. The samples are constructed as $r_t = \log p_t - \log p_{t-1}$, i.e., the weekly log-returns. The process r_t is assumed stationary. The vector $\mathbf{r}_t$ is constructed by stacking the log-returns of all K stocks.

Problem set up: We compare estimator performance in the minimum variance portfolio set up, that is, we allocate the portfolio weights to minimize the overall variance. Recall that the problem is formulated as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \\ \text{subject to} \quad & \mathbf{1}^T \mathbf{w} = 1 \end{aligned} \tag{5.4.1}$$

with $\mathbf{\Sigma}$ being the covariance matrix of $\mathbf{r}_t$. It can be seen that the scale of $\mathbf{\Sigma}$ does not affect the solution to this problem.

Training: To estimate $\mathbf{\Sigma}$, we use a rolling window approach with window size N. In particular, for the nonshrinkage estimators, at week t we use the log-returns of the previous N weeks to estimate the normalized covariance $\mathbf{\Sigma}$ and find the optimal portfolio allocation $\mathbf{w}_t^\star$ according to problem (5.4.1). For the shrinkage estimators, we further divide the N weeks returns into two parts with the first $N^{\text{train}} = N - N^{\text{val}}$ weeks for estimating $\mathbf{\Sigma}$ for different regularization parameters (α, γ) and find the allocation strategy $\mathbf{w}$, and the remaining N^{val} weeks as validation data for selecting the best $(\alpha^\star, \gamma^\star)$, which is the one that yields the smallest empirical portfolio variance $\text{Var}\left(\left\{\mathbf{w}^T \mathbf{r}_i\right\}_{i=t-N^{\text{val}},\ldots,t-1}\right)$. The normalized covariance $\mathbf{\Sigma}$ is then re-estimated with the overall N weeks log-return and tuning parameter $(\alpha^\star, \gamma^\star)$.

Testing: Having obtained $\mathbf{w}_t^\star$ in the training process, we then use it to invest for N^{test} weeks and collect the returns $v_t = \left(\mathbf{w}_t^\star\right)^T \mathbf{r}_t$. Every N^{test} weeks we rebalance the portfolio based on this procedure.

Fig. 5.4 compares the variance (risk) of the portfolio constructed based on different estimators. The parameters are set to be $N^{\text{val}} = 24$ and $N^{\text{test}} = 12$, the shrinkage targets are $\mathbf{t} = \mathbf{0}$ since the log-returns are close to zero, and $\mathbf{T} = \mathbf{I}/K$ so that $\text{Tr}\,(\mathbf{T}) = 1$, and the step size of the searching grid for (α, γ) is set to be 0.2. The size of the rolling window N is varied from 72 to 124. We can see that shrinkage estimators always yield the lowest risk. The performance gain of shrinkage Cauchy estimator against the two plug-in shrinkage estimators,

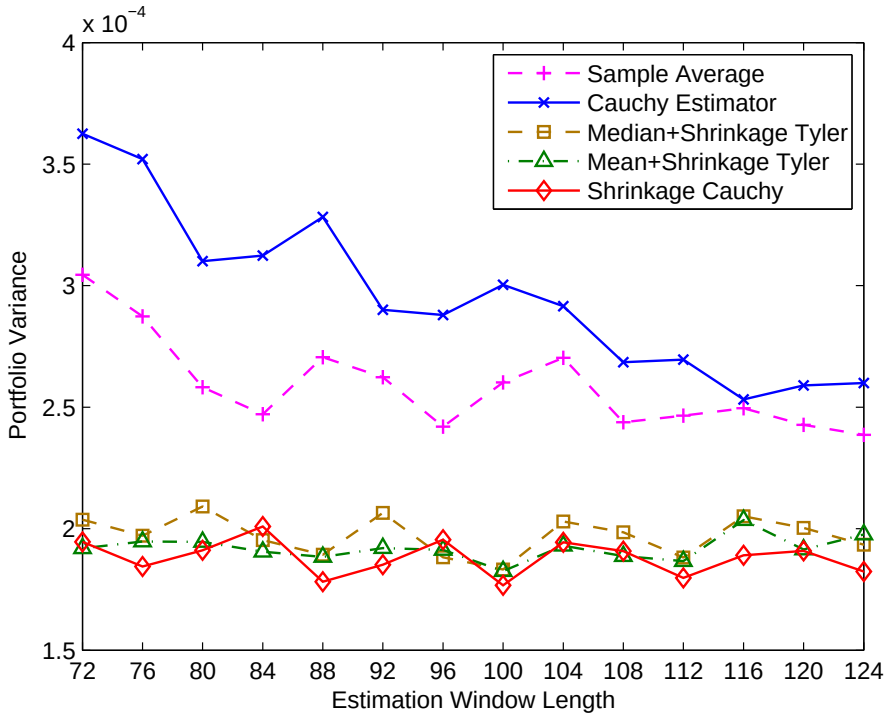


Figure 5.4: Risk (variance) comparison of portfolio constructed based on different covariance estimators.

however, is not significant enough, as opposed to the previous synthetic data simulation. This can be caused by the non-stationarity of stock data, so that a good in-sample performance does not lead to a good out-of-sample performance.

5.5 Conclusion

In this chapter, we have considered the robust mean-covariance estimation with samples drawn from an elliptical distribution that is capable of modeling heavy tails and outliers. In particular, we have proposed a robust shrinkage estimator by adding a penalty term to the Cauchy likelihood function, and established the existence and uniqueness result of the shrinkage estimator under certain regularity conditions. Efficient numerical algorithms have been provided based on the majorization-minimization framework with provable convergence and simulation results have shown that the proposed estimator works better in the small sample scenario with the presence of erroneous observations.

5.6 Appendix

5.6.1 Proof for Proposition 5.1

We analyze the following nested optimization problem:

$$\min_{\mathbf{R}\succ\mathbf{0}}\min_{\boldsymbol{\mu}} \quad \alpha\left(K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{R}\right)\right) \\ +\gamma\log\left(1+\left(\boldsymbol{\mu}-\mathbf{t}\right)^{T}\mathbf{R}^{-1}\left(\boldsymbol{\mu}-\mathbf{t}\right)\right).$$

For any fixed value of $\mathbf{R}\succ\mathbf{0}$, it is easy to see that the optimal $\boldsymbol{\mu}$ is $\boldsymbol{\mu}^{\star}=\mathbf{t}$. Substituting the optimal $\boldsymbol{\mu}$ into the problem we have

$$\underset{\mathbf{R}\succ\mathbf{0}}{\text{minimize}} \quad K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{R}\right).$$

Setting the gradient to zero yields the stationary condition

$$\mathbf{R}=\frac{K\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)}$$

with solution $\mathbf{R}=r\mathbf{T}$ for any $r>0$. We now show that the stationary points $\mathbf{R}=r\mathbf{T}$ are actually the minimizers.

We claim that the objective function $h\left(\mathbf{t},\mathbf{R}\right)=K\log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right)+\log\det\left(\mathbf{R}\right)$ goes to positive infinity when $\mathbf{R}$ tends to a singular matrix. To see this, first notice that $h\left(\mathbf{t},\mathbf{R}\right)$ is scale-invariant, i.e., $h\left(\mathbf{t},\mathbf{R}\right)=h\left(\mathbf{t},r\mathbf{R}\right)$, therefore we add the constraint $\mathrm{Tr}\left(\mathbf{R}\right)=1$ to remove the scale ambiguity. Eigendecompose $\mathbf{R}$ as $\mathbf{R}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{T}$, and define $\tilde{\mathbf{T}}=\mathbf{U}^{T}\mathbf{T}\mathbf{U}$. The diagonal components of $\tilde{\mathbf{T}}$ are all positive since $\mathbf{T}\succ\mathbf{0}$. Therefore $h\left(\mathbf{t},\mathbf{R}\right)=K\log\left(\sum_{i=1}^{K}\lambda_{i}^{-1}\tilde{t}_{ii}\right)+\sum_{i=1}^{K}\log\lambda_{i}$ with λ_{i} being the i-th diagonal component of $\boldsymbol{\Lambda}$ and $\tilde{t}_{ii}$ being the i-th diagonal component of $\tilde{\mathbf{T}}$. Without loss of generality we can assume the ordering $\lambda_{1}\geq\cdots\geq\lambda_{K}$. Consider the case $\lambda_{j}\to0$ for some $1<j\leq K$, then we have

$$\begin{aligned}h\left(\mathbf{t},\mathbf{R}\right)&\geq K\log\left(\sum_{i=j}^{K}\lambda_{i}^{-1}\tilde{t}_{ii}\right)+\sum_{i=j}^{K}\log\lambda_{i}+\text{const.}\\&\geq\frac{K\sum_{i=j}^{K}\log\lambda_{i}^{-1}}{K-j+1}+\sum_{i=j}^{K}\log\lambda_{i}+\text{const.}\to+\infty.\end{aligned}$$

Therefore, a minimizer of $h\left(\mathbf{t},\mathbf{R}\right)$ exists on $\mathbb{S}_{++}^{K}$ and has to satisfy the stationary condition.

The scale-invariant property of $h(\mathbf{t}, \mathbf{R})$ implies $r\mathbf{T}$ must be a global minima.

5.6.2 Proof for Theorem 5.1

Notice the fact that $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \to +\infty$ on the boundary of the feasible set $\mathbb{R}^K \times \mathbb{S}_{++}^K$ implies the minimum exists, we therefore seek for the condition that guarantees $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) \to +\infty$ on the boundary.

Define $\bar{\mathbf{x}}_i = [\mathbf{x}_i; 1]$, $\bar{\mathbf{t}} = [\mathbf{t}; 1]$ and matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{R} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix},$$

we have the following identities

$$\begin{aligned} \bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i &= 1 + (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ \bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}} &= 1 + (\mathbf{t} - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{t} - \boldsymbol{\mu}) \\ \mathbf{S}^T \boldsymbol{\Sigma}^{-1} \mathbf{S} &= \mathbf{R}^{-1} \end{aligned} \tag{5.6.1}$$

with $\mathbf{S} = \begin{bmatrix} \mathbf{I}_K \\ \mathbf{0}_{1\times K} \end{bmatrix}$. The loss function $L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R})$ can be equivalently written in $\boldsymbol{\Sigma}$ as

$$\begin{aligned} & L^{\text{shrink}}(\boldsymbol{\mu}, \mathbf{R}) = L^{\text{shrink}}(\boldsymbol{\Sigma}) \\ = & \left(\alpha + \frac{N}{2}\right) \log\det(\boldsymbol{\Sigma}) + \frac{K+1}{2} \sum_i \log\left(\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i\right) \\ & + \gamma \log\left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right) + \alpha K \log\left(\mathrm{Tr}\left(\mathbf{S}^T \boldsymbol{\Sigma}^{-1} \mathbf{S}\mathbf{T}\right)\right). \end{aligned}$$

Define the feasible set of $\boldsymbol{\Sigma}$ as $\mathcal{S} = \{\boldsymbol{\Sigma} | \boldsymbol{\Sigma} \succ \mathbf{0},\ \Sigma_{K+1,K+1} = 1\}$. In the rest of the proof, we are going to find the condition that ensures $L^{\text{shrink}}(\boldsymbol{\Sigma}) \to +\infty$ as $\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})} \to +\infty$ for all $\boldsymbol{\Sigma} \in \mathcal{S}$, which implies that a minimum exists on $\mathcal{S}$. Denote the eigenvalues of $\boldsymbol{\Sigma}$ as $\lambda_1 \geq \cdots \geq \lambda_{K+1}$, on the set $\mathcal{S}$ we have $\lambda_1 \geq 1$ and $\lambda_{K+1} \leq 1$.

The quantities $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i$ and $\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}$ are all greater than or equal to 1 by the identities (5.6.1), therefore the corresponding terms in $L^{\text{shrink}}(\boldsymbol{\Sigma})$ are nonnegative. For the term

$\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)$ we have

$$\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right) = \mathrm{Tr}\left(\begin{bmatrix} \mathbf{T}^{\frac{1}{2}} & \mathbf{0} \end{bmatrix}\boldsymbol{\Sigma}^{-1}\begin{bmatrix} \mathbf{T}^{\frac{1}{2}} \\ \mathbf{0} \end{bmatrix}\right).$$

Eigendecompose $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ with $\boldsymbol{\Lambda} \triangleq \mathrm{diag}\left(\lambda_1, \ldots, \lambda_{K+1}\right)$, and denote the eigenvector corresponding to λ_j as $\mathbf{u}_j$, we can express $\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right)$ as $\mathrm{Tr}\left(\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{T}\right) = \sum_j \lambda_j^{-1}\left\|\tilde{\mathbf{t}}_j\right\|^2$, where $\tilde{\mathbf{t}}_j = \begin{bmatrix} \mathbf{T}^{\frac{1}{2}} & \mathbf{0} \end{bmatrix}\mathbf{u}_j$.

Now define the function

$$\begin{aligned} G\left(\boldsymbol{\Sigma}\right) &= \exp\left(-L^{\text{shrink}}\left(\boldsymbol{\Sigma}\right)\right) \\ &= \det\left(\boldsymbol{\Sigma}\right)^{-\frac{N}{2}-\alpha}\prod_i\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)^{-\frac{K+1}{2}} \cdot \\ &\quad \left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right)^{-\gamma}\left(\sum_{j=1}^{K+1}\lambda_j^{-1}\left\|\tilde{\mathbf{t}}_j\right\|^2\right)^{-\alpha K}. \end{aligned}$$

The function $L^{\text{shrink}}\left(\boldsymbol{\Sigma}\right) \to +\infty$ if and only if $G\left(\boldsymbol{\Sigma}\right) \to 0$.

Denote the subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ as S_j and define $D_j = S_j \backslash S_{j-1} = \{\mathbf{x} \in \mathbb{R}^{K+1} | \mathbf{x} \in S_j, \mathbf{x} \notin S_{j-1}\}$ with $S_0 = \{0\}$ and $D_0 = \{0\}$. The D_j's partition the whole $\mathbb{R}^{K+1}$ space. Notice that $P_N\left\{S_0\right\} = 0$ since the last element of the augmented sample $\bar{\mathbf{x}}_i$ is 1, therefore $\sum_{j=1}^m P_N\left(D_j\right) = P_N\left(S_m\right)$ and $\sum_{j=m}^{K+1} P_N\left(D_j\right) = 1 - P_N\left(S_{m-1}\right)$.

Partition the samples $\bar{\mathbf{x}}_i$ according to D_j's and define

$$G_j = \lambda_j^{-\frac{N}{2}-\alpha}\prod_{\bar{\mathbf{x}}_i \in D_j}\left(\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i\right)^{-\frac{K+1}{2}}$$

therefore the objective function can be written as

$$G\left(\boldsymbol{\Sigma}\right) = \prod_{j=1}^{K+1} G_j \cdot \left(\bar{\mathbf{t}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{t}}\right)^{-\gamma}\left(\sum_{j=1}^{K+1}\lambda_j^{-1}\left\|\tilde{\mathbf{t}}_j\right\|^2\right)^{-\alpha K}.$$

If $\bar{\mathbf{x}}_i \in D_h$, we have the following fact:

$$\bar{\mathbf{x}}_i^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_i = \sum_{j=1}^{h}\lambda_j^{-1}\left\|\bar{\mathbf{x}}_i^T\mathbf{u}_j\right\|^2 \geq \lambda_h^{-1}\left\|\bar{\mathbf{x}}_i^T\mathbf{u}_h\right\|^2 > 0.$$

Define two integers r and s with $0 \leq r \leq K$, $1 \leq s \leq K+1$, and $r \leq s$, such that $\lambda_h \to +\infty$ for $h \in [1, r]$, λ_h is bounded for $h \in (r, s]$, and $\lambda_h \to 0$ for $h \in (s, K+1]$.

First consider the G_h's with $h \in [1, r]$. If $D_h \neq \emptyset$, then $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i = \sum_{j=1}^{h} \lambda_j^{-1} \left\| \bar{\mathbf{x}}_i^T \mathbf{u}_j \right\|^2$. Since for all the λ_j's with $j \leq r$ all goes to infinity, the quantity $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i$ must go to zero. This contradicts the fact that $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i \geq 1$. Therefore no $\bar{\mathbf{x}}_i$ lies in D_h for $h \in [1, r]$, and $G_h = O\left(\lambda_h^{-\frac{N}{2}-\alpha}\right)$.

Next, for the G_h's with $h \in (r, s]$, clearly G_h is bounded away from both 0 and $+\infty$, thus does not have any effect on the order of $G(\boldsymbol{\Sigma})$. Finally, consider the G_h's with $h \in (s, K]$, since $\lambda_h \to 0$, we have $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i \to +\infty$. Since

$$+\infty > \lim_{\lambda_h \to 0} \left(\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i \right) \lambda_h > 0,$$

we have $\bar{\mathbf{x}}_i^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i = O\left(\lambda_h^{-1}\right)$, and $G_h = O\left(\lambda_h^{-\frac{N}{2}-\alpha+\frac{K+1}{2}NP_N(D_h)}\right)$.

Having finished the terms associated with $\mathbf{x}_i$, we then analyze the terms contributed by regularization.

For the term $\left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right)^{-\gamma}$, since the D_j's partition the whole space, there must exist some D_h that $\bar{\mathbf{t}} \in D_h$. By the previous analysis, we have $h > r$. If $h \leq s$ then $0 < \left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right)^{-\gamma} < +\infty$, which is $O(1)$, and if $h \in (s, K]$, $\left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right)^{-\gamma} = O\left(\lambda_h^{\gamma}\right)$. In short, $\left(\bar{\mathbf{t}}^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{t}}\right)^{-\gamma} = O\left(\lambda_h^{\gamma 1_{\{h>s\}}}\right)$.

Finally we analyze the last term $\left(\sum_{j=1}^{K+1} \lambda_j^{-1} \left\| \tilde{\mathbf{t}}_j \right\|^2\right)^{-\alpha K}$. Recall the definition $\tilde{\mathbf{t}}_j = \begin{bmatrix} \mathbf{T}^{\frac{1}{2}} & \mathbf{0} \end{bmatrix} \mathbf{u}_j$, and denote each column of $\mathbf{T}^{\frac{1}{2}}$ as $\mathbf{t}_i \in \mathbb{R}^K$. Then for each vector $[\mathbf{t}_i; 0] \in \mathbb{R}^{K+1}$ there must exist some D_j to which it belongs. Denote the largest index of such D_j as q, we have $\left\| \tilde{\mathbf{t}}_q \right\| \neq 0$ and $\left\| \tilde{\mathbf{t}}_j \right\| = 0$ for all $j > q$. Repeating the previous reasoning we conclude that if $q \leq r$ or $q > s$, $\left(\sum_{j=1}^{K+1} \lambda_j^{-1} \left\| \tilde{\mathbf{t}}_j \right\|^2\right)^{-\alpha K} = O\left(\lambda_q^{\alpha K}\right)$ and if $q \in (r, s]$, it is some constant. In short, $\left(\sum_{j=1}^{K+1} \lambda_j^{-1} \left\| \tilde{\mathbf{t}}_j \right\|^2\right)^{-\alpha K} = O\left(\lambda_q^{\alpha K 1_{\{q \leq r\} \cup \{q > s\}}}\right)$.

Combining the three terms above, denote the partition that $\bar{\mathbf{t}}$ belongs to as D_h, i.e., $\bar{\mathbf{t}} \in$

D_h, we have

$$\begin{aligned} G(\mathbf{\Sigma}) &= \prod_{j=1}^{K+1} G_j \cdot \left(\bar{\mathbf{t}}^T \mathbf{\Sigma}^{-1} \bar{\mathbf{t}}\right)^{-\gamma} \left(\sum_{j=1}^{K+1} \lambda_j^{-1} \left\|\tilde{\mathbf{t}}_j\right\|^2\right)^{-\alpha K} \\ &= \prod_{j=1}^{r} O\left(\lambda_j^{-\frac{N}{2}-\alpha}\right) \prod_{j=s+1}^{K+1} O\left(\lambda_j^{-\frac{N}{2}-\alpha+\frac{K+1}{2}NP_N(D_j)}\right) \cdot \\ &\quad O\left(\lambda_h^{\gamma 1_{\{h>s\}}}\right) O\left(\lambda_q^{\alpha K 1_{\{q\leq r\}\cup\{q>s\}}}\right) \end{aligned}$$

with $\prod_{j=a}^{b} \triangleq 1,$ if $a > b$. By the ordering $\lambda_1 \geq \cdots \geq \lambda_{K+1}$, to guarantee $G \to 0$ we impose the conditions

$$\left(-\frac{N}{2}-\alpha\right) m + \alpha K 1_{\{q\leq m\}} < 0, \ \forall 1 \leq m \leq r \tag{5.6.2}$$

and

$$\begin{aligned} &\left(-\frac{N}{2}-\alpha\right)(K+2-m) + \frac{K+1}{2} N \sum_{j=m}^{K+1} P_N(D_j) \\ &+ \gamma 1_{\{m\leq h\}} + \alpha K 1_{\{m\leq q\}} > 0, \ \forall K+1 \geq m \geq s+1, \end{aligned} \tag{5.6.3}$$

where the first one forces the terms in the product corresponding to $\lambda \to +\infty$ to go to zero, and the second one forces the terms in the product corresponding to $\lambda \to 0$ to go to zero. Before simplifying the conditions, we first establish the following lemma.

Lemma 5.3. *$q \leq m$ if and only if $S_m \supseteq \mathbb{R}^K$. If $S_m = \mathbb{R}^K$, the corresponding $\mathbf{\Sigma}$ takes the form $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, which implies $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda_{K+1} = 1$.*

Proof. Recall the definition of q, which is the largest index j of D_j that $[\mathbf{t}_i; 0]$ belongs to. Therefore $q \leq m$ if and only if $[\mathbf{t}_i; 0] \in S_m$ for all i. Under the assumption that $\mathbf{T}$ is full rank, we have $S_m \supseteq \mathbb{R}^K$. If $S_m = \mathbb{R}^K$, since S_m is a K-dimensional subspace spanned by $[\mathbf{u}_1, \ldots, \mathbf{u}_K]$, which are the eignenvectors of $\mathbf{\Sigma}$, we have $\mathbf{\Sigma} = \sum_{j=1}^{K} \lambda_j \mathbf{u}_j \mathbf{u}_j^T + \lambda_{K+1} \mathbf{e}_{K+1} \mathbf{e}_{K+1}^T$, where $\mathbf{e}_{K+1} \triangleq [\mathbf{0}_{K\times 1}; 1]$. Clearly $\mathbf{\Sigma}$ must take the form $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, which implies that $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda_{K+1} = 1$. □

Consider the condition (5.6.2). r can take an arbitrary integer value from 0 to K. When $m \leq K-1$, or $m = K$ and $S_m \neq \mathbb{R}^K$, the condition is satisfied automatically since $\left(-\frac{N}{2}-\alpha\right) m$ is always negative and $q > m$. When $S_m = \mathbb{R}^K$, $\left(-\frac{N}{2}-\alpha\right) K + \alpha K < 0$. We therefore have proved that condition (5.6.2) is satisfied for any proper subspace of $\mathbb{R}^{K+1}$.

Now we move to condition (5.6.3). Since $\sum_{j=m}^{K+1} P_N(D_j) = 1 - P_N(S_{m-1})$ and condition (5.6.3) should be valid for all $s \in [1, K+1]$. Substituting $d = m-1$ yields the condition that for any proper subspace S_d with $d \in [1, K]$

$$\left(-\frac{N}{2} - \alpha\right)(K+1-d) + \frac{K+1}{2} N\left(1 - P_N(S_d)\right) + \gamma 1_{\{d+1 \le h\}} + \alpha K 1_{\{d+1 \le q\}} > 0$$

We have $h \le d$ if $\bar{\mathbf{t}} \in S_d$ and $q \le d$ if $S_d = \mathbb{R}^K$ by the previous argument, therefore condition (5.6.3) is equivalent to: for any proper subspace $S \subseteq \mathbb{R}^{K+1}$, if $\bar{\mathbf{t}} \notin S$ and $S \neq \mathbb{R}^K$, we need

$$P_N(S) < \frac{\dim(S)(N+2\alpha) + 2\gamma - 2\alpha}{(K+1)N}$$

and if $\bar{\mathbf{t}} \in S$, which implies $S \neq \mathbb{R}^K$ since the last entry of $\bar{\mathbf{t}}$ is 1, we need

$$P_N(S) < \frac{\dim(S)(N+2\alpha) - 2\alpha}{(K+1)N}$$

and finally if $S = \mathbb{R}^K$, which implies $\bar{\mathbf{t}} \notin S$, which corresponds to the situation when $\lambda_K \to +\infty$ and $\lambda_{K+1} = 1$ according to Lemma 5.3, we actually do not have condition (5.6.3) since no $\lambda \to 0$.

Notice that $\bar{\mathbf{x}}$ and $\bar{\mathbf{t}}$ belong to the same subspace S if and only if $\mathbf{x}$ and $\mathbf{t}$ belong to the same hyperplane, and $\bar{\mathbf{x}} \in S$ if and only if $\mathbf{x} \in H$ with $\dim(H) = \dim(S) - 1$. Finally we arrive at the condition on samples: for any hyperplane $H \subset \mathbb{R}^K$ with dimension $0 \le \dim(H) < K$, if H contains $\mathbf{t}$,

$$P_N(H) \;<\; \frac{(2\alpha+N)\dim(H) + N}{(K+1)N}$$

if H does not contain $\mathbf{t}$,

$$P_N(H) \;<\; \frac{(2\alpha+N)\dim(H) + N + 2\gamma}{(K+1)N}.$$

5.6.3 Proof for Theorem 5.2

Define matrix

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} \zeta^{-1}\mathbf{R} + \boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}$$

and assume that the following equality is satisfied:

$$\begin{aligned} \tilde{\mathbf{\Sigma}} &= \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\bar{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{x}}_i} \\ &\quad +\frac{2\gamma}{N+2\alpha}\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\bar{\mathbf{t}}^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{t}}} \\ &\quad +\frac{2\alpha K}{N+2\alpha}\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}^T\tilde{\mathbf{\Sigma}}^{-1}\mathbf{S}\mathbf{T}\right)}. \end{aligned} \tag{5.6.4}$$

Rewriting the equality above in terms of the original variables yields the system of equations as follows:

$$\boldsymbol{\mu} = \frac{(K+1)\sum w_i\mathbf{x}_i + 2\gamma w_{\mathbf{t}}\mathbf{t}}{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}} \tag{5.6.5}$$

$$\begin{aligned} \mathbf{R} &= \frac{\zeta}{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}} \\ &\quad \cdot\left\{(K+1)\sum_{i=1}^{N} w_i\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\right. \\ &\quad \left.+2\gamma w_{\mathbf{t}}\left(\mathbf{t}-\boldsymbol{\mu}\right)\left(\mathbf{t}-\boldsymbol{\mu}\right)^T + \frac{2\alpha\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)}\right\} \end{aligned} \tag{5.6.6}$$

$$\zeta = \frac{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}}{N+2\alpha}, \tag{5.6.7}$$

where

$$w_i = \frac{1}{1+\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)}$$

and

$$w_{\mathbf{t}} = \frac{1}{1+\left(\mathbf{t}-\boldsymbol{\mu}\right)^T\mathbf{R}^{-1}\left(\mathbf{t}-\boldsymbol{\mu}\right)}.$$

By substituting (5.6.7) into (5.6.6) we get exactly the condition (5.2.3) that a stationary point of problem (5.2.2) should satisfy. Namely, if $(\boldsymbol{\mu}, \mathbf{R})$ solves (5.2.3) then there exists a corresponding $\tilde{\mathbf{\Sigma}}$ that solves (5.6.4).

Multiplying both sides of (5.6.4) by $\tilde{\mathbf{\Sigma}}$ and taking the trace gives the identity

$$\frac{2\gamma - 2\alpha}{N + 2\alpha} = \frac{\zeta - 1}{\zeta} \cdot \frac{(K+1)\sum w_i + 2\gamma w_{\mathbf{t}}}{N + 2\alpha},$$

combined with (5.6.7), ζ can be solved as $\zeta = \frac{N+2\gamma}{N+2\alpha} > 0$.

Under the existence condition provided in Theorem 5.1, the global minimal of (5.2.2) is a solution of (5.2.3) with $\mathbf{R}^\star \succ \mathbf{0}$, which implies (5.6.4) has at least one symmetric positive definite solution $\tilde{\mathbf{\Sigma}}^\star$. It suffices to prove that the solution is unique on $\mathbb{S}_{++}^{K+1}$ when $\gamma \geq \alpha$.

First consider the case $\gamma > \alpha$, which implies $\zeta - 1 > 0$. Without loss of generality we can assume $\tilde{\mathbf{\Sigma}} = \mathbf{I}$ is a solution, since if $\mathbf{\Sigma}$ is a solution we can always define $\tilde{\bar{\mathbf{x}}}_i = \mathbf{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{x}}_i$, $\tilde{\bar{\boldsymbol{\mu}}} = \mathbf{\Sigma}^{-\frac{1}{2}}\bar{\boldsymbol{\mu}}$ and $\tilde{\mathbf{S}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}$. Now assume there is another solution $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_1$ and its largest eigenvalue $\lambda_1 > 1$, then

$$\begin{aligned}
\tilde{\mathbf{\Sigma}} \leq & \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\lambda_1^{-1}\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i} + \frac{2\gamma}{N+2\alpha}\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\lambda_1^{-1}\bar{\mathbf{t}}^T\bar{\mathbf{t}}} + \frac{2\alpha K}{N+2\alpha}\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\lambda_1^{-1}\mathrm{Tr}\left(\mathbf{S}\mathbf{T}\mathbf{S}^T\right)} \\
< & \frac{K+1}{N+2\alpha}\sum_{i=1}^{N}\frac{\lambda_1\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{(\zeta-1)+\bar{\mathbf{x}}_i^T\bar{\mathbf{x}}_i} + \frac{2\alpha K}{N+2\alpha}\frac{\lambda_1\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\mathbf{S}\mathbf{T}\mathbf{S}^T\right)} + \frac{2\gamma}{N+2\alpha}\frac{\lambda_1\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{(\zeta-1)+\bar{\mathbf{t}}^T\bar{\mathbf{t}}} \\
= & \lambda_1\mathbf{I},
\end{aligned}$$

where the first inequality follows from the fact that for any positive semidefinite matrix $\mathbf{A}$, $\mathrm{Tr}\left(\mathbf{\Sigma}_1^{-1}\mathbf{A}\right) \geq \mathrm{Tr}\left(\lambda_1^{-1}\mathbf{A}\right)$, the second strict inequality follows from the assumption that $\lambda_1 > 1$ and the last equality follows from the assumption that $\mathbf{I}$ is a solution to the equation (5.6.4). We have a contradiction $\lambda_1 < \lambda_1$, hence $\lambda_1 \leq 1$. Similarly we can prove $\lambda_{K+1} \geq 1$, therefore $\mathbf{\Sigma}_1 = \mathbf{I}$.

Next consider the case $\gamma = \alpha$, which implies $\zeta - 1 = 0$. The equation simplifies to

$$\left(\frac{N}{2} + \alpha\right)\tilde{\mathbf{\Sigma}} = \frac{K+1}{2}\sum_{i=1}^{N}\frac{\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T}{\bar{\mathbf{x}}_i^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{x}}_i} + \alpha K\frac{\mathbf{S}\mathbf{T}\mathbf{S}^T}{\mathrm{Tr}\left(\tilde{\mathbf{\Sigma}}^{-1}\mathbf{S}\mathbf{T}\mathbf{S}^T\right)} + \alpha\frac{\bar{\mathbf{t}}\bar{\mathbf{t}}^T}{\bar{\mathbf{t}}^T\tilde{\mathbf{\Sigma}}^{-1}\bar{\mathbf{t}}}.$$

By the same reasoning as Theorem 6 in [102] and the fact that $\tilde{\mathbf{\Sigma}}_{K+1,K+1} = 1$, we conclude that the solution to the above equation is unique.

Chapter 6

Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions

The previous two chapters discuss the problem of estimating a covariance matrix and jointly estimating the mean and covariance matrix in the high dimension regime. The basic methodology is to shrink the estimates to some prior target to reduce estimation variance. In this chapter, we study the robust covariance estimation problem in the context that the true parameter is structured. For presentation clarity, we focus on the covariance estimation part, but the developed theory can be generalized to the estimation of mean and a structured covariance jointly.

6.1 Introduction

Covariance matrix in some applications naturally possesses some special structure. Exploiting the structure information in the estimation process usually implies a reduction in the number of parameters to be estimated, and thus is beneficial to improving the estimation accuracy [23]. Various types of structures have been studied in the literature, including the Toeplitz, sparsity, sparse inverse, banded, group symmetry, spiked, and Kronecker structures [1, 23–30].

While the previously mentioned works have shown that enforcing a prior structure on the covariance estimator improves its performance in many applications, most of them either

assume that the samples follow a Gaussian distribution or attempt to regularize the sample covariance matrix. Not surprisingly, the resulting estimator will give a large estimation error is the population distribution is heavy-tails or there exists outlying observations.

A way to address this problem is to find a robust structured covariance matrix estimator that performs well even if the underlying distribution deviates from the Gaussian assumption. One approach is to refer to the minimax principle and seek the "best" estimate of the covariance for the worst case noise. To be precise, the underlying probability distribution of the samples f is assumed to belong to an uncertainty set of functions $\mathcal{F}$ that contains the Gaussian distribution, and the desired minimax robust estimator is the one whose maximum asymptotic variance over the set $\mathcal{F}$ is less than that of any other estimator. Two types of uncertainty sets $\mathcal{F}$, namely the ε-contamination and the Kolmogorov class, were considered in [103], where a structured maximum likelihood type estimate (M-estimate) was derived as the solution of a constrained optimization problem.

With the population distribution being the family of elliptically symmetric distributions, we have introduced the Tyler's estimator in Chapter 4 as a minimax and distribution free robust estimator in the real field. A benefit of employing the Tyler's estimator, apart from its robustness, is that it can be viewed as the MLE obtained by fitting the samples normalized by their length to their joint density function regardless of the parametric form of the population distribution. Due to these advantages, Tyler's estimator has attracted a particular attention compared to other M-estimators.

The problem of obtaining a structured Tyler's estimator was investigated in the recent works [104] and [105]. In particular, the authors of [104] focused on the group symmetry structure and proved that it is geodesically convex. As the Tyler's estimator can be defined alternatively as the minimizer of a cost function that is also geodesically convex, it is concluded that any local minimum of the cost function on a group symmetry constraint set is a global minimum. A numerical algorithm was also proposed to solve the constrained minimization problem. In [105], a convex structural constraint set was studied and a generalized method of moments type covariance estimator, COCA, was proposed. A numerical algorithm was also provided based on semidefinite relaxation. It was proved that COCA is an asymptotically consistent estimator. However, the algorithm suffers from the drawback that the computational cost increases as either N or K grows.

In this chapter, we formulate the structured covariance estimation problem as the minimization of Tyler's cost function under the structural constraint. Our work generalizes [104] by considering a much larger family of structures, which includes the group symmetry structure. Instead of attempting to obtain a global optimal solution, which is a challenging task due to the non-convexity of the objective function, we focus on devising algorithms that converge to a stationary point of the problem. We first work out an algorithm framework for the general convex structural constraint based on the MM framework, where a sequence of convex programming is required to be solved. Then we consider several special cases that appear frequently in practical applications. By exploiting specific problem structures, the algorithm is particularized, significantly reducing the computational load. We further discuss in the end two types of widely studied non-convex structures that turn out to be computationally tractable under the MM framework; one of them being the Kronecker structure and the other one being the spiked covariance structure. Under the assumption that the objective function goes to infinity whenever the variable tends to a singular limit, which is guaranteed when the samples are drawn from a continuous distribution and the number of samples N is larger than its dimension K, the sequences generated by the algorithms converges to a stationary point. It is worth mentioning that the Tyler's cost function is shown to be geodesically convex in [106], therefore the algorithm converges to a global minimizer when the constraint set is also geodesically convex [104, 107].

This chapter is organized as follows. The robust covariance estimation problem is formulated in Section 6.2. In Section 6.3, we derive a majorization-minimization based algorithm framework for the general convex structure. Several special cases are considered in Section 6.4, where the algorithm is particularized obtaining higher efficiency by considering the specific form of the structure. Section 6.5 discusses the Kronecker structure and the spiked covariance structure, which are non-convex but algorithmically tractable. Numerical results are presented in Section 6.6 and we conclude in Section 6.7.

6.2 Tyler's Estimator with Structural Constraint

In this chapter, we assume the samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathbb{C}^K$ are *complex-valued* with pdf as follows:

$$f(\mathbf{x}) = c \det(\mathbf{R}_0)^{-1} g\left((\mathbf{x}-\boldsymbol{\mu}_0)^H \mathbf{R}_0^{-1} (\mathbf{x}-\boldsymbol{\mu}_0)\right). \tag{6.2.1}$$

In the rests of this chapter, we assume $\boldsymbol{\mu}_0$ is known and equals to a zero vector without loss of generality, since otherwise we can work with the centered data $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}_0$.

Analogous to the real-valued sample case, Tyler's estimator for $\mathbf{R}_0$ in the complex field is defined as the solution to the following fixed-point equation:

$$\mathbf{R} = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^H}{\mathbf{x}_i^H \mathbf{R}^{-1} \mathbf{x}_i}. \tag{6.2.2}$$

Most of its properties in the real field extend to the complex field. For instance, the normalized random variable $\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ follows a complex angular central Gaussian distribution with pdf

$$f(\mathbf{s}) \propto \det(\mathbf{R}_0)^{-1} \left(\mathbf{s}^H \mathbf{R}_0^{-1} \mathbf{s}\right)^{-K}. \tag{6.2.3}$$

Tyler's estimator coincides with the maximum likelihood estimator (MLE) of $\mathbf{R}_0$ by fitting the normalized samples $\{\mathbf{s}_i\}$ to $f(\mathbf{s})$. In other words, the estimator $\hat{\mathbf{R}}$ is the minimizer of the following cost function

$$L(\mathbf{R}) = \log\det(\mathbf{R}) + \frac{K}{N} \sum_{i=1}^{N} \log\left(\mathbf{x}_i^H \mathbf{R}^{-1} \mathbf{x}_i\right) \tag{6.2.4}$$

on the positive definite cone $\mathbb{S}_{++}^K$.

It has been noticed that in some applications, the covariance matrix possesses a certain structure and taking account this information into the estimation yields a better estimate of $\mathbf{R}_0$ [28–30, 103]. Motivated by this idea, we focus on the problem of including a prior structure information into the Tyler's estimator to improve its estimation accuracy. To formulate the problem, we assume that $\mathbf{R}_0$ is constrained in a non-empty set $\mathcal{S}$ that is the intersection of a closed set, which characterizes the covariance structure, and the positive semidefinite cone $\mathbb{S}_{+}^K$, and then proceed to solve the optimization problem:

$$\begin{aligned} \underset{\mathbf{R}}{\text{minimize}} \quad & \log\det(\mathbf{R}) + \frac{K}{N} \sum_{i=1}^{N} \log\left(\mathbf{x}_i^H \mathbf{R}^{-1} \mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R} \in \mathcal{S}. \end{aligned} \tag{6.2.5}$$

The minimizer $\hat{\mathbf{R}}$ of the above problem is the one in the structural set $\mathcal{S}$ that maximizes the likelihood of the normalized samples $\{\mathbf{s}_i\}$.

Throughout the chapter, we make the following assumption.

Assumption 1: The cost function $L(\mathbf{R}_t) \to +\infty$ when the sequence $\{\mathbf{R}_t\}$ tends to a singular limit point of the constraint set $\mathcal{S}$.

Under this assumption, the case that $\mathbf{R}$ is singular can be excluded in the analysis of the algorithms hereafter.

A sufficient condition for Assumption 1 to hold is stated as follows.

Assumption 2: f is a continuous probability distribution, and $N > K$.

Assumption 2 implies $L(\mathbf{R}_t) \to +\infty$ whenever $\mathbf{R}_t$ tends to the boundary of the positive semidefinite cone $\mathbb{S}_+^K$ with probability one [108]. Therefore it also implies Assumption 1 as $\mathcal{S} \subseteq \mathbb{S}_+^K$.

Problem (6.2.5) is difficult to solve for two reasons. First, the constraint set $\mathcal{S}$ is too general to tackle. Second, even if $\mathcal{S}$ possesses a nice property such as convexity, the objective function is still non-convex. Instead of trying to find the global minimizer, which appears to be too ambitious for the reasons pointed out above, we aim at devising efficient algorithms that are capable of finding a stationary point of (6.2.5). We rely on the MM framework to derive the algorithms.

In the rest of this chapter, we are going to derive the specific form of the surrogate function $g(\mathbf{R}|\mathbf{R}_t)$ based on a detailed characterization of various kinds of $\mathcal{S}$. In addition, we are going to show how the algorithm can be particularized at a lower computational cost with a finer structure of $\mathcal{S}$ available. Before moving to the algorithmic part, we first compare our formulation with several related works in the literature.

6.2.1 Related Works

In [103], the authors derived a minimax robust covariance estimator assuming that $f(\mathbf{x})$ is a corrupted Gaussian distribution with noise that belongs to the ε-contamination class and the Kolmogorov class. The estimator is defined as the solution of a constrained optimization problem similar to (6.2.5), but with a different cost function. Apart from the distinction that the family of distributions we consider is the set of elliptical distributions, the focus of our work, which completely differs from [103], is on developing efficient numerical algorithms for different types of structural constraint set $\mathcal{S}$.

Two other closely related works are [104] and [105]. In [104], the authors have investigated a special case of (6.2.5), where $\mathcal{S}$ is the set of all positive semidefinite matrices with

group symmetry structure. It has been shown that both $L(\mathbf{R})$ and the group symmetry constraint are geodesically convex, therefore any local minimizer of (6.2.5) is global. Several examples, including the circulant and persymmetry structure, have been proven to be a special case of the group symmetry constraint. A numerical algorithm has also been provided that decreases the cost function monotonically. Our work includes the group symmetry structure as a special case since the constraint is linear, and provides an alternative algorithm to solve the problem.

In [105], the authors have considered imposing convex constraint on Tyler's estimator. A generalized method of moment type estimator based on semidefinite relaxation defined as the solution of the following problem:

$$\begin{aligned} \underset{\mathbf{R}\in\mathcal{S},d_i}{\text{minimize}} \quad & \left\| \mathbf{R} - \frac{1}{N}\sum_{i=1}^{N} d_i \mathbf{x}_i \mathbf{x}_i^H \right\| \\ \text{subject to} \quad & \mathbf{R} \succeq \frac{1}{K} d_i \mathbf{x}_i \mathbf{x}_i^H,\ \forall i = 1,\ldots,N, \\ & d_i > 0,\ \forall i = 1,\ldots,N, \end{aligned} \tag{6.2.6}$$

was proposed and proved to be asymptotically consistent. Nevertheless, the number of constraints grows linearly in N and as it was pointed out in the paper, the algorithm becomes computationally demanding either when the problem dimension K or the number of samples N is large. On the contrary, our algorithm based on formulation (6.2.5) is less affected by the number of samples N and is therefore more computationally tractable.

6.3 Tyler's Estimator with Convex Structural Constraint

In this section, we are going to derive a general algorithm for problem (6.2.5) with $\mathcal{S}$ being a closed convex subset of $\mathbb{S}_+^K$, which enjoys a wide range of applications. For instance, the Toeplitz structure can be imposed on the covariance matrix of the received signal in DOA problems. Banding is also considered as a way of regularizing a covariance matrix whose entries decay fast as they get far away from the main diagonal.

Since $\mathcal{S}$ is closed and convex, constructing a convex surrogate function $g(\mathbf{R}|\mathbf{R}_t)$ for $L(\mathbf{R})$ turns out to be a natural idea since then $\mathbf{R}_{t+1}$ can be found via

$$\mathbf{R}_{t+1} = \arg\min_{\mathbf{R}\in\mathcal{S}} \ g(\mathbf{R}|\mathbf{R}_t), \tag{6.3.1}$$

which is a convex programming.

Proposition 6.1. *At any $\mathbf{R}_t \succ \mathbf{0}$, the objective function $L(\mathbf{R})$ can be upperbounded by the convex surrogate function*

$$g(\mathbf{R}|\mathbf{R}_t) = Tr\left(\mathbf{R}_t^{-1}\mathbf{R}\right) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i}{\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i} + const. \tag{6.3.2}$$

with equality achieved at $\mathbf{R}_t$.

Proof. Since $\log\det(\cdot)$ is concave, $\log\det(\mathbf{R})$ can be upperbounded by its first order Taylor expansion at $\mathbf{R}_t$:

$$\log\det(\mathbf{R}) \leq \log\det(\mathbf{R}_t) + \mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{R}\right) - K \tag{6.3.3}$$

with equality achieved at $\mathbf{R}_t$.

Also, by the concavity of the $\log(\cdot)$ function we have

$$\log(x) \leq \log a + \frac{x}{a} - 1,\ \forall a > 0, \tag{6.3.4}$$

which leads to the bound

$$\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right) \leq \frac{\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i}{\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i} + \log\left(\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i\right) - 1$$

with equality achieved at $\mathbf{R}_t$. □

The variable $\mathbf{R}$ then can be updated as (6.3.1) with surrogate function (6.3.2).

By the convergence result of the MM algorithm, it can be concluded that every limit point of the sequence $\{\mathbf{R}_t\}$ is a stationary point of problem (6.2.5). Note that for all of the structural constraints that we are going to consider in this work, the set $\mathcal{S}$ possesses the property that

$$\mathbf{R} \in \mathcal{S} \text{ iff } r\mathbf{R} \in \mathcal{S},\ \forall r > 0. \tag{6.3.5}$$

In other words, $\mathcal{S}$ is a cone. Since the cost function $L(\mathbf{R})$ is scale-invariant in the sense that $L(\mathbf{R}) = L(r\mathbf{R})$, we can add a trace normalization step after the update of $\mathbf{R}_t$ without affecting the value of the objective function. The algorithm for a general convex structural set is summarized in Algorithm 6.1.

Algorithm 6.1 Robust covariance estimation under convex structure

1: Set $t = 0$, initialize $\mathbf{R}_t$ to be any positive definite matrix.
2: **repeat**
3: Compute $\mathbf{M}_t = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^H}{\mathbf{x}_i^H \mathbf{R}_t^{-1} \mathbf{x}_i}$.
4: Update $\mathbf{R}_{t+1}$ as

$$\tilde{\mathbf{R}}_{t+1} = \arg\min_{\mathbf{R} \in \mathcal{S}} \operatorname{Tr}\left(\mathbf{R}_t^{-1}\mathbf{R}\right) + \operatorname{Tr}\left(\mathbf{M}_t \mathbf{R}^{-1}\right) \tag{6.3.6}$$

$$\mathbf{R}_{t+1} = \tilde{\mathbf{R}}_{t+1} / \operatorname{Tr}\left(\tilde{\mathbf{R}}_{t+1}\right). \tag{6.3.7}$$

5: $t \leftarrow t + 1$.
6: **until** Some convergence criterion is met

Proposition 6.2. *If the set $\mathcal{S}$ satisfies (6.3.5), then the sequence $\{\mathbf{R}_t\}$ generated by Algorithm 6.1 satisfies*

$$\lim_{t \to \infty} d\left(\mathbf{R}_t, \mathcal{S}^\star\right) = 0, \tag{6.3.8}$$

where $\mathcal{S}^\star$ is the set of stationary points of problem (6.2.5).

Proof. Since the objective function $L(\mathbf{R})$ is scale-invariant, and the constraint set satisfies (6.3.5), solving (6.2.5) is equivalent to solving

$$\begin{aligned} \underset{\mathbf{R} \in \mathcal{S}}{\text{minimize}} \quad & \log\det(\mathbf{R}) + \frac{K}{N} \sum_{i=1}^{N} \log\left(\mathbf{x}_i^H \mathbf{R}^{-1} \mathbf{x}_i\right) \\ \text{subject to} \quad & \operatorname{Tr}(\mathbf{R}) = 1. \end{aligned}$$

The conclusion follows by a similar argument to Proposition 17 in [102]. □

6.3.1 General Linear Structure

In this subsection we further assume that the set $\mathcal{S}$ is the intersection of $\mathbb{S}_+^K$ and an affine set $\mathcal{A}$. The following lemma shows that in this case, the update of $\mathbf{R}$ (eqn. (6.3.6)) can be recast as a semidefinite programming (SDP).

Lemma 6.1. *Problem* (6.3.6) *is equivalent to*

$$\begin{aligned} \underset{\mathbf{S}, \mathbf{R} \in \mathcal{S}}{\textit{minimize}} \quad & \textit{Tr}\left(\mathbf{R}_t^{-1}\mathbf{R}\right) + \textit{Tr}\left(\mathbf{M}_t \mathbf{S}\right) \\ \textit{subject to} \quad & \begin{bmatrix} \mathbf{S} & \mathbf{I} \\ \mathbf{I} & \mathbf{R} \end{bmatrix} \succeq \mathbf{0}, \end{aligned} \tag{6.3.9}$$

in the sense that if $(\mathbf{S}^\star, \mathbf{R}^\star)$ *solves (6.3.9), then* $\mathbf{R}^\star$ *solves* (6.3.6)*.*

Proof. Problem (6.3.6) can be written equivalently as

$$\begin{aligned} \underset{\mathbf{S},\mathbf{R}\in\mathcal{S}}{\text{minimize}} \quad & \text{Tr}\left(\mathbf{R}_t^{-1}\mathbf{R}\right) + \text{Tr}\left(\mathbf{M}_t\mathbf{S}\right) \\ \text{subject to} \quad & \mathbf{S} = \mathbf{R}^{-1}. \end{aligned}$$

Now we relax the constraint $\mathbf{S} = \mathbf{R}^{-1}$ as $\mathbf{S} \succeq \mathbf{R}^{-1}$. By the Schur complement lemma for a positive semidefinite matrix, if $\mathbf{R} \succ \mathbf{0}$, then $\mathbf{S} \succeq \mathbf{R}^{-1}$ is equivalent to

$$\begin{bmatrix} \mathbf{S} & \mathbf{I} \\ \mathbf{I} & \mathbf{R} \end{bmatrix} \succeq \mathbf{0}.$$

Therefore (6.3.9) is a convex relaxation of (6.3.6).

The relaxation is tight since $\text{Tr}\left(\mathbf{M}_t\mathbf{S}\right) \geq \text{Tr}\left(\mathbf{M}_t\mathbf{R}^{-1}\right)$ if $\mathbf{M}_t \succeq \mathbf{0}$ and $\mathbf{S} \succeq \mathbf{R}^{-1}$. □

Lemma 6.1 reveals that for linear structural constraint, Algorithm 6.1 can be particularized as solving a sequence of SDPs.

An application is the case that $\mathbf{R}$ can be parameterized as

$$\mathbf{R} = \sum_{j=1}^{L} a_j \mathbf{B}_j \tag{6.3.10}$$

with $a_j \in \mathbb{C}$ being the variable and $\mathbf{B}_j \in \mathbb{C}^{K\times K}$ being the corresponding given basis matrix, and $\mathbf{R}$ is constrained to be in $\mathbb{S}_+^K$. Using expression (6.3.10), the minimization problem (6.3.9) can be simplified as

$$\begin{aligned} \underset{\mathbf{S},\{a_j\}}{\text{minimize}} \quad & \sum_{j=1}^{L} a_j \text{Tr}\left(\mathbf{R}_t^{-1}\mathbf{B}_j\right) + \text{Tr}\left(\mathbf{M}_t\mathbf{S}\right) \\ \text{subject to} \quad & \begin{bmatrix} \mathbf{S} & \mathbf{I} \\ \mathbf{I} & \sum_{j=1}^{L} a_j\mathbf{B}_j \end{bmatrix} \succeq \mathbf{0}. \end{aligned} \tag{6.3.11}$$

6.4 Tyler's Estimator with Special Convex Structures

Having introduced the general algorithm framework for a convex structure in the previous section, we are going to discuss in detail some convex structures that arise frequently in signal

processing related fields, and show that by exploiting the problem structure the algorithm can be particularized with a significant reduction in the computational load.

6.4.1 Sum of Rank-One Matrices Structure

The structure set $\mathcal{S}$ that we study in this part is

$$\mathcal{S} = \left\{ \mathbf{R} | \mathbf{R} = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H,\ p_j \geq 0 \right\}, \tag{6.4.1}$$

where the $\mathbf{a}_j$'s are known vectors in $\mathbb{C}^K$. The matrix $\mathbf{R}$ can be interpreted as a weighted sum of given matrices $\mathbf{a}_j \mathbf{a}_j^H$.

As an example application where structure (6.4.1) appears, consider the following signal model

$$\mathbf{x} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6.4.2}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_L]$. Assuming that the signal $\boldsymbol{\beta}$ and noise $\boldsymbol{\varepsilon}$ are zero-mean random variables and any two elements of them are uncorrelated, then the covariance matrix of $\mathbf{x}$ takes the form

$$\mathrm{Cov}\left(\mathbf{x}\right) = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H + \boldsymbol{\Sigma}, \tag{6.4.3}$$

where $p_j = \mathrm{Var}\left(\beta_j\right)$ is the signal variance and $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1, \ldots, \sigma_K\right)$ is the noise covariance matrix.

Define $\mathbf{p} = \left[p_1, \ldots p_L\right]^H$ and $\mathbf{P} = \mathrm{diag}\left(\mathbf{p}\right)$, then $\mathbf{R}$ can be written compactly as $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \boldsymbol{\Sigma}$. Further define

$$\begin{aligned} \tilde{\mathbf{P}} &= \mathrm{diag}\left(p_1, \ldots, p_L, \sigma_1, \ldots, \sigma_K\right) \\ \tilde{\mathbf{A}} &= \left[\mathbf{A}, \mathbf{I}\right] \end{aligned} \tag{6.4.4}$$

then $\mathbf{R} = \tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H$. Therefore, without loss of generality, we can focus on the expression $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H$, assuming that every K columns of $\mathbf{A}$ are linearly independent and $L > K$.

One application of the sum of rank-one matrices structure is the direction of arrival estimation problem. The signal model is expressed as:

$$\mathbf{x}\left(t\right) = \mathbf{A}\left(\boldsymbol{\theta}_s\right)\mathbf{s}\left(t\right) + \mathbf{n}\left(t\right),$$

where $\boldsymbol{\theta}_s = [\theta_1, \ldots, \theta_m]$ is a vector with elements representing the arriving directions of signal $\mathbf{s}(t)$,

$$\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_m)] \tag{6.4.5}$$

is the steering matrix and $\mathbf{n}$ is zero mean additive noise. We study the simple case of an ideal uniform linear array (ULA) with half-wavelength inter-element spacing, where

$$\mathbf{a}(\theta) = \left[1, e^{-j\pi\sin(\theta)}, \ldots, e^{-j\pi(K-1)\sin(\theta)}\right]^T. \tag{6.4.6}$$

Assuming that the signal $\mathbf{s}(t)$ is a wide-sense stationary random process with zero mean, the covariance of $\mathbf{x}(t)$ is

$$\mathbf{R} = \mathbf{A}(\boldsymbol{\theta}_s)\,\mathrm{Cov}(\mathbf{s})\,\mathbf{A}(\boldsymbol{\theta}_s)^H + \mathrm{Cov}(\mathbf{n}).$$

Further assume that the signals arriving from different directions are uncorrelated and that the noise is spatially white, i.e., $\mathrm{Cov}(\mathbf{s}) = \mathrm{diag}(p_1^s, \ldots, p_m^s) \triangleq \mathbf{P}_s$ and $\mathrm{Cov}(\mathbf{n}) = \sigma^2\mathbf{I}$, the covariance model simplifies to be

$$\mathbf{R} = \mathbf{A}(\boldsymbol{\theta}_s)\,\mathbf{P}_s\mathbf{A}(\boldsymbol{\theta}_s)^H + \sigma^2\mathbf{I}. \tag{6.4.7}$$

Now let us uniformly grid the interval $\left[-90^\circ, 90^\circ\right]$ with spacing $\Delta\theta$ and create vector $\boldsymbol{\theta}_e$ of length L. We assume that all the elements of $\boldsymbol{\theta}_s$ are on the grid. Then expression (6.4.7) can be rewritten as

$$\mathbf{R} = \mathbf{A}(\boldsymbol{\theta}_e)\,\mathbf{P}_e\mathbf{A}(\boldsymbol{\theta}_e)^H + \sigma^2\mathbf{I}, \tag{6.4.8}$$

with $p_i^e = p_j^s$ if the i-th column of $\mathbf{A}(\boldsymbol{\theta}_e)$ and the j-th column of $\mathbf{A}(\boldsymbol{\theta}_s)$ are equal, and $p_i^e = 0$ otherwise.

Recall that the problem to be solved takes the form

$$\begin{aligned} \underset{\mathbf{R},\mathbf{P}\succeq\mathbf{0}}{\text{minimize}} \quad & \log\det(\mathbf{R}) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H. \end{aligned} \tag{6.4.9}$$

Since $\mathbf{R}$ is linear in the p_j's, Algorithm 6.1 can be applied. In the following, we are going to provide a more efficient algorithm by substituting $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H$ into the objective function

$L(\mathbf{R})$ and applying the MM procedure with $\mathbf{P}$ being the variable.

Proposition 6.3. *At any $\mathbf{P}_t \succ \mathbf{0}$, the objective function*

$$L(\mathbf{P}) = \log\det\left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^H\left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}\mathbf{x}_i\right) \tag{6.4.10}$$

can be upperbounded by the surrogate function

$$g(\mathbf{P}|\mathbf{P}_t) = \mathbf{w}_t^H\mathbf{p} + \mathbf{d}_t^H\mathbf{p}^{-1} + \text{const.} \tag{6.4.11}$$

with equality achieved at $\mathbf{P} = \mathbf{P}_t$, where $\mathbf{p}^{-1}$ stands for the element-wise inverse of $\mathbf{p}$, and

$$\begin{aligned}
\mathbf{R}_t &= \mathbf{A}\mathbf{P}_t\mathbf{A}^H \\
\mathbf{M}_t &= \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^H}{\mathbf{x}_i^T\mathbf{R}_t^{-1}\mathbf{x}_i} \\
\mathbf{w}_t &= \text{diag}\left(\mathbf{A}^H\mathbf{R}_t^{-1}\mathbf{A}\right) \\
\mathbf{d}_t &= \text{diag}\left(\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1}\mathbf{M}_t\mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\right).
\end{aligned} \tag{6.4.12}$$

Proof. First, observe that inequalities (6.3.3) and (6.3.4) imply that

$$L(\mathbf{P}) \;\leq\; \mathbf{w}_t^H\mathbf{p} + \text{Tr}\left(\mathbf{M}_t\mathbf{R}^{-1}\right) + \text{const.} \tag{6.4.13}$$

with equality achieved at $\mathbf{P} = \mathbf{P}_t$.

Assume that $\mathbf{P} \succ \mathbf{0}$, from the identity

$$\begin{aligned}
\mathbf{S} &= \begin{bmatrix} \mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1} & \mathbf{I} \\ \mathbf{I} & \mathbf{A}\mathbf{P}\mathbf{A}^H \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1/2} \\ \mathbf{A}\mathbf{P}^{1/2} \end{bmatrix}\begin{bmatrix} \mathbf{P}^{-1/2}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1} & \mathbf{P}^{1/2}\mathbf{A}^H \end{bmatrix},
\end{aligned}$$

we know that $\mathbf{S} \succeq \mathbf{0}$. By the Schur complement, $\mathbf{S} \succeq \mathbf{0}$ is equivalent to

$$\mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1} \succeq \left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}. \tag{6.4.14}$$

Algorithm 6.2 Robust covariance estimation under sum of rank-one matrices structure

1: Set $t = 0$, initialize $\mathbf{p}_t$ to be any positive vector.
2: **repeat**
3: $\quad \tilde{\mathbf{R}}_t = \mathbf{A}\mathbf{P}_t\mathbf{A}^H$, $\mathbf{R}_t = \tilde{\mathbf{R}}_t / \mathrm{Tr}\left(\tilde{\mathbf{R}}_t\right)$.
4: $\quad$ Compute $\mathbf{M}_t$, $\mathbf{w}_t$, $\mathbf{d}_t$ with (6.4.12)
5: $\quad (p_j)_{t+1} = \sqrt{(d_j)_t \,/\, (w_j)_t}$
6: $\quad t \leftarrow t + 1$.
7: **until** some convergence criterion is met

Since $\mathbf{M}_t \succeq \mathbf{0}$, we have

$$\mathrm{Tr}\left(\mathbf{M}_t\mathbf{R}^{-1}\right) \leq \mathrm{Tr}\left(\mathbf{M}_t\mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1}\right) \tag{6.4.15}$$

with equality achieved at $\mathbf{P} = \mathbf{P}_t$.

Since $\mathbf{R} \succ \mathbf{0}$, the left hand side of (6.4.15) is finite. Therefore (6.4.15) is also valid for $\mathbf{P} \succeq \mathbf{0}$. Substituting (6.4.15) into (6.4.13) yields the surrogate function (6.4.11). □

Note that both $\mathbf{w}_t$ and $\mathbf{d}_t$ are real-valued, the update of $\mathbf{P}$ then can be found in closed-form as

$$(p_j)_{t+1} = \sqrt{(d_j)_t \,/\, (w_j)_t}. \tag{6.4.16}$$

The algorithm is summarized in Algorithm 6.2.

Compared to Algorithm 6.1, in which the minimization problem (6.3.1) has no closed-form solution and typically requires an iterative algorithm, the new algorithm only requires a single loop iteration in $\mathbf{p}$ and is expected to converge faster.

6.4.2 Toeplitz Structure

Consider the constraint set being the class of real-valued positive semidefinite Toeplitz matrices T_K. If $\mathbf{R} \in T_K$, then it can be completely determined by its first row $[r_0, \ldots, r_{K-1}]$[1].

In this subsection, we are going to show that based on the technique of circulant embedding, Algorithm 6.2 can be adopted to solve the Toeplitz structure constrained problem at a lower cost than applying the sequential SDP algorithm (Algorithm 6.1).

The idea of embedding a Toeplitz matrix as the upper-left part of a larger circulant matrix has been discussed in [23, 25, 109]. It was proved in [24] that any positive definite Toeplitz

[1]Following the convention, the indices for the Toeplitz structure start from 0.

matrix $\mathbf{R}$ of size $K \times K$ can be embedded in a positive definite circulant matrix $\mathbf{C}$ of larger size $L \times L$ parameterized by its first row of the form

$$[r_0, r_1, \ldots, r_{K-1}, *, \ldots, *, r_{K-1}, \ldots, r_1],$$

where $*$ denotes some real number. $\mathbf{R}$ then can be written as

$$\mathbf{R} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix} \mathbf{C} \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix}^T. \tag{6.4.17}$$

Clearly, for any fixed L, if $\mathbf{C}$ is symmetric positive semidefinite, so is $\mathbf{R}$. However, the statement is false the other way around. In other words, the set

$$T_K^L \triangleq \left\{ \mathbf{R} | \mathbf{R} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix} \mathbf{C} \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix}^T, \mathbf{C} \in C_L \right\}, \tag{6.4.18}$$

where C_L denotes the set of real-valued positive semidefinite circulant matrices of size $L \times L$, is a subset of T_K.

Instead of T_K, we restrict the feasible set to be T_K^L with $L \geq 2K - 1$. Since a circulant matrix can be diagonalized by the Fourier matrix, if $\mathbf{R} \in T_K^L$ then it can be written as

$$\mathbf{R} = \mathbf{A} \text{diag}\left(p_0, \ldots, p_{L-1}\right) \mathbf{A}^H, \tag{6.4.19}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix} \mathbf{F}_L, \tag{6.4.20}$$

with $\mathbf{F}_L$ being the normalized Fourier transform matrix of size $L \times L$ and $p_j = p_{L-j},\ \forall j = 1, \ldots, L-1$. [2]

The robust covariance estimation problem over the restricted set of Toeplitz matrices T_K^L then takes the form

$$\begin{aligned} \underset{\mathbf{R}, \mathbf{P} \succeq \mathbf{0}}{\text{minimize}} \quad & \log\det\left(\mathbf{R}\right) + \frac{K}{N} \sum_{i=1}^{N} \log\left(\mathbf{x}_i^H \mathbf{R}^{-1} \mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H \\ & p_j = p_{L-j},\ \forall j = 1, \ldots, L-1, \end{aligned} \tag{6.4.21}$$

[2] The algorithm is developed for estimating real-valued Toeplitz matrix, it can be adapted for estimating a complex-valued one by removing the constraint $p_j = p_{L-j}$.

Algorithm 6.3 Robust covariance estimation under the Toeplitz structure (Circulant Embedding)

1: Set L to be an integer such that $L \geq 2K-1$.
2: Construct matrix $\mathbf{A} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix} \mathbf{F}_L$
3: Call Algorithm 6.2 ($\mathbf{p}_t$ is initialized satisfying $p_j = p_{L-j}$, $\forall j = 1, \ldots, L-1$).

which is the same as (6.4.9) except that the last equality constraint on the p_j's.

By Proposition 6.3, the inner minimization problem takes the form

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}}{\text{minimize}} \quad & \mathbf{w}_t^T \mathbf{p} + \mathbf{d}_t^T \mathbf{p}^{-1} \\ \text{subject to} \quad & p_j = p_{L-j},\ \forall j = 1, \ldots, L-1. \end{aligned} \tag{6.4.22}$$

Note that by the property of the Fourier transform matrix, we have $\mathbf{a}_j = \bar{\mathbf{a}}_{L-j}$, $\forall j = 1, \ldots, L-1$, where the upper bar stands for element-wise complex conjugate. As a result, if $(p_j)_t = (p_{L-j})_t$, for $j = 1, \ldots, L-1$,

$$\begin{aligned} (w_j)_t &= (w_{L-j})_t \\ (d_j)_t &= (d_{L-j})_t \,, \end{aligned} \tag{6.4.23}$$

which implies that the constraint $p_j = p_{L-j}$ will be satisfied automatically.

The algorithm for the Toeplitz structure based on circulant embedding is summarized in Algorithm 6.3. Notice that Algorithm 6.3 can be generalized easily to noisy observations by the augmented representation (6.4.4).

6.4.3 Banded Toeplitz Structure

In addition to imposing the Toeplitz structure on a real-valued covariance matrix, in some applications we can further require that the Toeplitz matrix is k-banded, i.e., $r_j = 0$ if $j > k$. For example, the covariance matrix of a stationary moving average process of order k satisfies the above assumption. One may also consider banding the covariance matrix if it is known in prior that the correlation of x_t and $x_{t-\tau}$ decreases as τ increases.

Based on the circulant embedding technique introduced in the last subsection, the problem

can be formulated as

$$\begin{aligned} \underset{\mathbf{R},\mathbf{P}\succeq\mathbf{0}}{\text{minimize}} \quad & \log\det\left(\mathbf{R}\right) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H \\ & p_j = p_{L-j},\ \forall j = 1,\ldots,L-1 \\ & r_j = 0,\ \forall j = k+1,\ldots,K-1. \end{aligned} \tag{6.4.24}$$

By Proposition 6.3, the inner minimization problem becomes

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}}{\text{minimize}} \quad & \mathbf{w}_t^T\mathbf{p} + \mathbf{d}_t^T\mathbf{p}^{-1} \\ \text{subject to} \quad & p_j = p_{L-j},\ \forall j = 1,\ldots,L-1 \\ & r_j = 0,\ \forall j = k+1,\ldots,K-1, \end{aligned} \tag{6.4.25}$$

which can be rewritten compactly as

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}}{\text{minimize}} \quad & \mathbf{w}_t^T\mathbf{p} + \mathbf{d}_t^T\mathbf{p}^{-1} \\ \text{subject to} \quad & \begin{bmatrix} \mathbf{0}_{(K-k-1)\times k+1} & \mathbf{I}_{K-k-1} \end{bmatrix}\mathbf{A}\mathbf{p} = \mathbf{0} \\ & p_j = p_{L-j},\ \forall j = 1,\ldots,L-1. \end{aligned} \tag{6.4.26}$$

Recall that $\mathbf{a}_j = \bar{\mathbf{a}}_{L-j},\ \forall j = 1,\ldots,L-1$. For simplicity we assume that L is odd. The constraint $p_j = p_{L-j}$ implies that $p_j\mathbf{a}_j + p_{L-j}\bar{\mathbf{a}}_{L-j} = 2p_j\mathrm{Re}\left\{\mathbf{a}_j\right\}$. Define real-valued quantities

$$\tilde{\mathbf{A}} = \mathrm{Re}\left\{\left[\mathbf{a}_0, 2\mathbf{a}_1,\ldots,2\mathbf{a}_{\frac{L-1}{2}}\right]\right\} \tag{6.4.27}$$

$$\tilde{\mathbf{w}} = \left[w_0, 2w_1,\ldots,2w_{\frac{L-1}{2}}\right] \tag{6.4.28}$$

$$\tilde{\mathbf{d}} = \left[d_0, 2d_1,\ldots,2d_{\frac{L-1}{2}}\right], \tag{6.4.29}$$

we have the equivalent problem

$$\begin{aligned} \underset{\tilde{\mathbf{p}}\geq\mathbf{0}}{\text{minimize}} \quad & \tilde{\mathbf{w}}_t^T\tilde{\mathbf{p}} + \sum_{j=0}^{\left\lceil\frac{L-1}{2}\right\rceil}\tilde{d}_j/\tilde{p}_j \\ \text{subject to} \quad & \tilde{\mathbf{A}}\tilde{\mathbf{p}} = \mathbf{0}, \end{aligned} \tag{6.4.30}$$

Algorithm 6.4 Robust covariance estimation under the Banded Toeplitz structure (Circulant Embedding)

Set L to be an integer such that $L \geq 2K-1$.
Construct matrix $\mathbf{A} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \end{bmatrix} \mathbf{F}_L$ and $\tilde{\mathbf{A}}$ with (6.4.27).
Set $t = 0$, initialize $\mathbf{p}_t$ to be any positive vector.
repeat
    $\tilde{\mathbf{R}}_t = \mathbf{A}\mathbf{P}_t\mathbf{A}^H$, $\mathbf{R}_t = \tilde{\mathbf{R}}_t / \mathrm{Tr}\left(\tilde{\mathbf{R}}_t\right)$.
    Compute $\mathbf{M}_t$, $\mathbf{w}_t$, $\mathbf{d}_t$ with (6.4.12).
    Compute $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{d}}$ with (6.4.28) and (6.4.29), and update $\tilde{\mathbf{p}}$ as the minimizer of (6.4.32).
    Compute $\mathbf{p}$ with (6.4.31), $\mathbf{p}_t \leftarrow \mathbf{p}$
    $t \leftarrow t+1$
until some convergence criterion is met

where the variables $\tilde{\mathbf{p}}$ and $\mathbf{p}$ are related by

$$\tilde{\mathbf{p}} = \left[p_0, p_1, \ldots, p_{\frac{L-1}{2}}\right]. \tag{6.4.31}$$

Compared to (6.4.26), the equivalent problem has a lower computational cost as both the number of variables and constraints are reduced. Using the epigraph form, problem can be casted as the following second-order-cone programming (SOCP)

$$\begin{aligned} \underset{\tilde{\mathbf{p}},\mathbf{t}}{\text{minimize}} \quad & \tilde{\mathbf{w}}_t^T \tilde{\mathbf{p}} + \sum_{j=0}^{\frac{L-1}{2}} d_j t_j \\ \text{subject to} \quad & \tilde{\mathbf{A}}\tilde{\mathbf{p}} = \mathbf{0}, \\ & \left\| \begin{bmatrix} 2 \\ \tilde{p}_j - t_j \end{bmatrix} \right\| \leq \tilde{p}_j + t_j, \, \forall j. \end{aligned} \tag{6.4.32}$$

The algorithm for the banded Toeplitz structure is summarized in Algorithm 6.4.

6.4.4 Convergence Analysis

We consider Algorithm 6.2, and the argument for Algorithms 6.3, and 6.4 would be similar.

As Proposition 6.3 requires $\mathbf{P}_t \succ \mathbf{0}$, we consider the following ϵ-approximation of problem (6.4.9):

$$\begin{aligned} \underset{\mathbf{R},\mathbf{p}\geq\mathbf{0}}{\text{minimize}} \quad & \log\det\left(\mathbf{R}+\epsilon\mathbf{A}\mathbf{A}^H\right) \\ & + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^H\left(\mathbf{R}+\epsilon\mathbf{A}\mathbf{A}^H\right)^{-1}\mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R}=\mathbf{A}\mathbf{P}\mathbf{A}^H \end{aligned} \tag{6.4.33}$$

with $\varepsilon > 0$, where the upperbound derived in Proposition 6.3 can be applied for $\tilde{\mathbf{P}} \triangleq \mathbf{P}+\epsilon\mathbf{I}$. Algorithm 6.2 can be easily modified to solve problem (6.4.33), and under Assumption 1, the limit point of the sequence $\{\mathbf{p}_t^\epsilon\}$ generated by Algorithm 6.2 converges to the set of stationary points of (6.4.33).

That is, if $(\mathbf{p}^\epsilon)^\star$ is a limit point of $\{\mathbf{p}_t^\epsilon\}$, then

$$\nabla L^\epsilon\left((\mathbf{p}^\epsilon)^\star\right)^T\mathbf{d}\geq\mathbf{0} \tag{6.4.34}$$

for any feasible direction $\mathbf{d}$, where $\nabla L^\epsilon\left((\mathbf{p}^\epsilon)^\star\right)$ is the gradient of the objective function $L^\epsilon(\mathbf{p})$ at $(\mathbf{p}^\epsilon)^\star$.

Proposition 6.4. *Under Assumption 1, let ϵ_k be a positive sequence with $\lim_{k\to+\infty}\epsilon_k = 0$, then any limit point $\mathbf{p}^\star$ of the sequence $\{(\mathbf{p}^{\epsilon_k})^\star\}$ is a stationary point of problem (6.4.9).*

Proof. The conclusion follows from the continuity of $\nabla L^\epsilon\left((\mathbf{p}^\epsilon)^\star\right)$ in $(\mathbf{p}^\epsilon)^\star$ and ϵ under Assumption 2. □

In practice, as ϵ can be chosen as an arbitrarily small number, directly applying Algorithms 6.2, 6.3 and 6.4 or adapting them to solving the ϵ-approximation problem would be virtually the same.

6.5 Tyler's Estimator with Non-Convex Structure

In the previous sections we have proposed algorithms for Tyler's estimator with a general convex structural constraint and discussed in detail some special cases. For the non-convex structure, the problem is more difficult to handle. In this section, we are going to introduce two popular non-convex structures that are tractable by applying the MM algorithm, namely the spiked covariance structure and the Kronecker structure.

6.5.1 The Spiked Covariance Structure

The term "spiked covariance" was introduced in [110] and refers to the covariance matrix model

$$\mathbf{R} = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H + \sigma^2 \mathbf{I}, \tag{6.5.1}$$

where L is some integer that is less than K, and the $\mathbf{a}_j$'s are unknown orthonormal basis vectors. Note that although (6.5.1) and (6.4.3) share similar form, they differ from each other essentially since the $\mathbf{a}_j$'s in (6.4.3) are known and are not necessarily orthogonal. The model is directly related to principal component analysis, subspace estimation, and also plays an important role in sensor array applications [1, 108]. This model, referred to as factor model, is also very popular in financial time series analysis [111].

The constrained optimization problem is formulated as

$$\begin{aligned} \underset{\mathbf{R},\mathbf{a}_j,\mathbf{p}\geq\mathbf{0},\sigma}{\text{minimize}} \quad & \log\det(\mathbf{R}) + \frac{K}{N}\sum_{i=1}^{N}\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right) \\ \text{subject to} \quad & \mathbf{R} = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H + \sigma^2\mathbf{I}, \\ & \mathbf{A}^H\mathbf{A} = \mathbf{I}, \end{aligned} \tag{6.5.2}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_L]$.

Applying the upperbound (6.3.4) for the second term in the objective function yields the following inner minimization problem:

$$\begin{aligned} \underset{\mathbf{R},\mathbf{a}_j,\mathbf{p}\geq\mathbf{0},\sigma}{\text{minimize}} \quad & \log\det(\mathbf{R}) + \frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i}{\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i} \\ \text{subject to} \quad & \mathbf{R} = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H + \sigma^2\mathbf{I}, \\ & \mathbf{A}^H\mathbf{A} = \mathbf{I}. \end{aligned} \tag{6.5.3}$$

Algorithm 6.5 Robust covariance estimation under the spiked covariance structure

1: Initialize $\mathbf{R}_0$ to be an arbitrary feasible positive definite matrix.
2: **repeat**
3: $\quad \mathbf{M}_t = \frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^H}{\mathbf{x}_i^H \mathbf{R}_t^{-1} \mathbf{x}_i}$.
4: $\quad$ Eigendecompose $\mathbf{M}_t$ as $\mathbf{M}_t = \sum_{j=1}^{L} \lambda_j \mathbf{u}_j \mathbf{u}_j^T$, where $\lambda_1 \geq \cdots \geq \lambda_K$.
5: $\quad$ Compute $\sigma^\star$, $p_j^\star$, $\mathbf{a}_j^\star$ with (6.5.4)
6: $\quad \tilde{\mathbf{R}}_{t+1} = \sum_{j=1}^{L} p_j^\star \mathbf{a}_j^\star \left(\mathbf{a}_j^\star\right)^H + \sigma^\star \mathbf{I}$.
7: $\quad \mathbf{R}_{t+1} = \tilde{\mathbf{R}}_{t+1} / \mathrm{Tr}\left(\tilde{\mathbf{R}}_{t+1}\right)$.
8: $\quad t \leftarrow t+1$.
9: **until** Some convergence criterion is met.

Although the problem is non-convex, a global minimizer can be found in closed-form as

$$\begin{aligned} (\sigma^\star)^2 &= \frac{1}{K-L} \sum_{j=L+1}^{K} \lambda_j \\ p_j^\star &= \lambda_j - (\sigma^\star)^2 \\ \mathbf{a}_j^\star &= \mathbf{u}_j, \end{aligned} \tag{6.5.4}$$

where $\lambda_1 \geq \cdots \geq \lambda_K$ are the sorted eigenvalues of matrix $\frac{K}{N} \sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^H}{\mathbf{x}_i^H \mathbf{R}_t^{-1} \mathbf{x}_i}$ and the $\mathbf{u}_j$'s are the associated eigenvectors [112]. The algorithm for the spiked covariance structure is summarized in Algorithm 6.5.

As the feasible set is not convex, the convergence statement of the MM algorithm in [53] needs to be modified as follows.

Proposition 6.5. *Any limit point* $\mathbf{R}^\star$ *generated by the algorithm satisfies*

$$\mathrm{Tr}\left(\nabla L\left(\mathbf{R}^\star\right)^H \mathbf{R}\right) \geq 0, \ \forall \mathbf{R} \in \mathcal{T}_{\mathcal{S}}\left(\mathbf{R}^\star\right),$$

where $\mathcal{T}_{\mathcal{S}}\left(\mathbf{R}^\star\right)$ *stands for the tangent cone of* $\mathcal{S}$ *at* $\mathbf{R}^\star$.

Proof. The result follows by combining the standard convergence proof of the MM algorithm [53] and the necessity condition of $\mathbf{R}^\star$ being the global minimal of $g\left(\mathbf{R}|\mathbf{R}^\star\right)$ over an arbitrary set $\mathcal{S}$ (see Proposition 4.7.1 in [113]):

$$\mathrm{Tr}\left(\nabla g\left(\mathbf{R}^\star|\mathbf{R}^\star\right)^H \mathbf{R}\right) \geq 0, \ \forall \mathbf{R} \in \mathcal{T}_{\mathcal{S}}\left(\mathbf{R}^\star\right).$$

□

6.5.2 The Kronecker Structure

In this subsection we consider the covariance matrix that can be expressed as the Kronecker product of two matrices, i.e.,

$$\mathbf{R} = \mathbf{A} \otimes \mathbf{B}, \tag{6.5.5}$$

where $\mathbf{A} \in \mathbb{S}_+^p$ and $\mathbf{B} \in \mathbb{S}_+^q$.

Substituting $\mathbf{R} = \mathbf{A} \otimes \mathbf{B}$ into the objective function yields the equivalent problem:

$$\begin{aligned} \underset{\mathbf{A}\succeq\mathbf{0},\mathbf{B}\succeq\mathbf{0}}{\text{minimize}} \quad & \frac{pq}{N}\sum_{i=1}^{N}\log \text{Tr}\left(\mathbf{A}^{-1}\mathbf{M}_i^H\mathbf{B}^{-1}\mathbf{M}_i\right) \\ & + q\log\det\left(\mathbf{A}\right) + p\log\det\left(\mathbf{B}\right) \end{aligned} \tag{6.5.6}$$

where $\mathbf{M}_i \in \mathbb{C}^{q\times p}$ and $\text{vec}\left(\mathbf{M}_i\right) = \mathbf{x}_i$. Denote the objective function of (6.5.6) as $L\left(\mathbf{A},\mathbf{B}\right)$.

Note that although the objective function of the equivalent problem is still non-convex, the constraint set of the equivalent problem (6.5.6) becomes the Cartesian product of two convex sets, which is convex.

6.5.2.1 Gauss-Seidel

Since $L\left(\mathbf{R}\right)$ is scale-invariant, we can make the restriction that $\text{Tr}\left(\mathbf{A}\right) = 1$ and $\text{Tr}\left(\mathbf{B}\right) = 1$ and then problem (6.5.6) can be solved by updating $\mathbf{A}$ and $\mathbf{B}$ alternatively.

Specifically, for fixed $\mathbf{B} = \mathbf{B}_t$, we need to solve the following problem:

$$\begin{aligned} \underset{\mathbf{A}\succeq\mathbf{0}}{\text{minimize}} \quad & \log\det\left(\mathbf{A}\right) + \frac{p}{N}\sum_{i=1}^{N}\log \text{Tr}\left(\mathbf{A}^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right) \\ \text{subject to} \quad & \text{Tr}\left(\mathbf{A}\right) = 1. \end{aligned} \tag{6.5.7}$$

Setting the gradient of the objective function to zero yields the fixed-point equation

$$\mathbf{A} = \frac{p}{N}\sum_{i=1}^{N}\frac{\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i}{\text{Tr}\left(\mathbf{A}^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right)}. \tag{6.5.8}$$

As objective function of (6.5.7) is essentially the same as the Tyler's cost function (6.2.4), an argument similar to Theorem 2.1 in [15] reveals that the solution to (6.5.8) is unique up to a

Algorithm 6.6 Robust covariance estimation under the Kronecker structure (Gauss-Seidel)

1: Initialize $\mathbf{A}_0$ and $\mathbf{B}_0$ to be arbitrary positive definite matrices of size $p \times p$ and $q \times q$, respectively.
2: **repeat**
3: Update $\mathbf{A}$ with (6.5.9).
4: Update $\mathbf{B}$ with (6.5.10).
5: $t \leftarrow t + 1$.
6: **until** Some convergence criterion is met.

positive scaling factor, and under Assumption 1, the iteration

$$\begin{aligned} \tilde{\mathbf{A}} &= \frac{p}{N} \sum_{i=1}^{N} \frac{\mathbf{M}_i^H \mathbf{B}_t^{-1} \mathbf{M}_i}{\mathrm{Tr}\left(\mathbf{A}_r^{-1} \mathbf{M}_i^H \mathbf{B}_t^{-1} \mathbf{M}_i\right)} \\ \mathbf{A}_{r+1} &= \tilde{\mathbf{A}} / \mathrm{Tr}\left(\tilde{\mathbf{A}}\right) \end{aligned} \tag{6.5.9}$$

converges to the unique global minimum of (6.5.7) as $r \to +\infty$.

Assign $\mathbf{A}_{t+1} = \lim_{r \to +\infty} \mathbf{A}_r$, similarly we have the fixed-point iteration for $\mathbf{B}$ as

$$\begin{aligned} \tilde{\mathbf{B}} &= \frac{q}{N} \sum_{i=1}^{N} \frac{\mathbf{M}_i \mathbf{A}_{t+1}^{-1} \mathbf{M}_i^H}{\mathrm{Tr}\left(\mathbf{A}_{t+1}^{-1} \mathbf{M}_i^H \mathbf{B}_r^{-1} \mathbf{M}_i\right)} \\ \mathbf{B}_{r+1} &= \tilde{\mathbf{B}} / \mathrm{Tr}\left(\tilde{\mathbf{B}}\right), \end{aligned} \tag{6.5.10}$$

and $\mathbf{B}_{t+1} = \lim_{r \to +\infty} \mathbf{B}_r$.

Proposition 6.6. *Under Assumption 1, every limit point, denoted by* $(\mathbf{A}^\star, \mathbf{B}^\star)$*, of the sequence* $\{(\mathbf{A}_t, \mathbf{B}_t)\}$ *generated by Algorithm 6.6 is a global minimizer of (6.5.6).*

Proof. The convergence of block coordinate descent algorithm states that the pair $(\mathbf{A}^\star, \mathbf{B}^\star)$ is a stationary point of problem (6.5.6) (Proposition 2.7.1 in [114]). Moreover, it is proved that the Tyler's cost function (6.2.4) is geodesically convex on $\mathbb{S}_{++}$ [106]. Lemma 3 in [107] then implies that $L(\mathbf{A} \otimes \mathbf{B})$ is also geodesically convex. Finally, Corollary 3.1 in [115] the stationary point $(\mathbf{A}^\star, \mathbf{B}^\star)$ is actually a global minimizer since $L(\mathbf{A} \otimes \mathbf{B})$ is continuously differentiable. □

6.5.2.2 Block Majorization Minimization

A stationary point of $L(\mathbf{A}, \mathbf{B})$ can also be found by block majorization-minimization algorithm (Block MM). Compared to Algorithm 6.6 derived based on Gauss-Seidel update, which

is a double loop algorithm, block MM only performs a single loop iteration. Also, block MM allows one to impose additional structures on $\mathbf{A}$ and $\mathbf{B}$.

By Proposition 6.1, with the value of $\mathbf{B}_t$ fixed to be $\mathbf{B}_t$, a convex upperbound of $L(\mathbf{A}, \mathbf{B})$ on $\mathbb{S}_+^p$ at point $\mathbf{A}_t$ (ignoring a constant term and up to a scale factor of q) can be found as

$$g(\mathbf{A}|\mathbf{A}_t, \mathbf{B}_t) = \operatorname{Tr}\left(\mathbf{A}_t^{-1}\mathbf{A}\right) + \frac{p}{N}\sum_{i=1}^{N}\frac{\operatorname{Tr}\left(\mathbf{A}^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right)}{\operatorname{Tr}\left(\mathbf{A}_t^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right)}. \tag{6.5.11}$$

Lemma 6.2. *Under Assumption 1, for any $\mathbf{A}_t, \mathbf{B}_t \succ \mathbf{0}$, the matrix*

$$\mathbf{M}(\mathbf{A}_t, \mathbf{B}_t) = \frac{p}{N}\sum_{i=1}^{N}\frac{\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i}{Tr\left(\mathbf{A}_t^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right)}$$

is nonsingular.

Proof. At $(\mathbf{A}_t, \mathbf{B}_t)$ (ignoring a constant term and up to a scale factor of q) the function $L(\mathbf{A}, \mathbf{B}_t)$ can be upperbounded by

$$\tilde{g}(\mathbf{A}|\mathbf{A}_t, \mathbf{B}_t) = \log\det(\mathbf{A}) + \operatorname{Tr}\left(\mathbf{A}^{-1}\mathbf{M}(\mathbf{A}_t, \mathbf{B}_t)\right). \tag{6.5.12}$$

If $\mathbf{M}(\mathbf{A}_t, \mathbf{B}_t)$ is singular, we can eigendecompose $\mathbf{M}(\mathbf{A}_t, \mathbf{B}_t)$ as

$$\mathbf{M}(\mathbf{A}_t, \mathbf{B}_t) = \mathbf{U}\operatorname{diag}(\lambda_1, \ldots, \lambda_p)\mathbf{U}^H$$

with $\lambda_1 = 0$, and set $\mathbf{A}^{-1} = \mathbf{U}\operatorname{diag}(\sigma_1, \ldots, \sigma_p)\mathbf{U}^H$.

Letting $\sigma_1 \to 0$ would result in $\tilde{g}(\mathbf{A}|\mathbf{A}_t, \mathbf{B}_t)$ unbounded below, which implies $L(\mathbf{A}, \mathbf{B}_t)$ is also unbounded below and contradicts Assumption 1. □

An immediate implication of Lemma 6.2 is that $g(\mathbf{A}|\mathbf{A}_t, \mathbf{B}_t)$ is strictly convex on $\mathbb{S}_{++}^p$ and has a unique closed-form minimizer given by

$$\mathbf{A}_{t+1} = \mathbf{A}_t^{1/2}\left(\mathbf{A}_t^{-1/2}\mathbf{M}\mathbf{A}_t^{-1/2}\right)^{1/2}\mathbf{A}_t^{1/2}, \tag{6.5.13}$$

where $\mathbf{M} = \dfrac{p}{N}\displaystyle\sum_{i=1}^{N}\frac{\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i}{\operatorname{Tr}\left(\mathbf{A}_t^{-1}\mathbf{M}_i^H\mathbf{B}_t^{-1}\mathbf{M}_i\right)}$.

Algorithm 6.7 Robust covariance estimation under the Kronecker structure (Block Majorization Minimization)

1: Initialize $\mathbf{A}_0$ and $\mathbf{B}_0$ to be arbitrary positive definite matrices of size $p \times p$ and $q \times q$, respectively.
2: **repeat**
3: Update $\mathbf{A}$ with (6.5.13).
4: Update $\mathbf{B}$ with (6.5.14).
5: $t \leftarrow t + 1$.
6: **until** Some convergence criterion is met.

Symmetrically, we have the the update for $\mathbf{B}$ given by

$$\mathbf{B}_{t+1} = \mathbf{B}_t^{1/2} \left(\mathbf{B}_t^{-1/2} \mathbf{M} \mathbf{B}_t^{-1/2} \right)^{1/2} \mathbf{B}_t^{1/2}, \tag{6.5.14}$$

where $\mathbf{M} = \dfrac{q}{N} \displaystyle\sum_{i=1}^{N} \dfrac{\mathbf{M}_i \mathbf{A}_{t+1}^{-1} \mathbf{M}_i^H}{\mathrm{Tr}\left(\mathbf{A}_{t+1}^{-1} \mathbf{M}_i^H \mathbf{B}_t^{-1} \mathbf{M}_i \right)}$.

Proposition 6.7. *Under Assumption 1, every limit point, denoted by $(\mathbf{A}^\star, \mathbf{B}^\star)$, of the pair generated by Algorithm 6.7 is a global minimizer of problem (6.5.6).*

Proof. Theorem 2 (a) in [53] implies that $(\mathbf{A}^\star, \mathbf{B}^\star)$ is a stationary point of problem (6.5.6). The rest of the proof is the same as that of Proposition 6.6. □

Note that with the surrogate function of the form (6.5.11), we can easily impose additional convex structures on $\mathbf{A}$ and $\mathbf{B}$, and the update is found by solving the convex problem:

$$\begin{aligned} \mathbf{A}_{t+1} &= \arg\min_{\mathbf{A} \in \mathcal{A}} g\left(\mathbf{A} | \mathbf{A}_t, \mathbf{B}_t \right), \\ \mathbf{B}_{t+1} &= \arg\min_{\mathbf{B} \in \mathcal{B}} g\left(\mathbf{B} | \mathbf{A}_{t+1}, \mathbf{B}_t \right), \end{aligned} \tag{6.5.15}$$

with $\mathcal{A}$ and $\mathcal{B}$ being the convex structural constraint sets.

6.6 Numerical Results

In this section, we present numerical results that demonstrate the effect of imposing structure on the covariance estimator on reducing estimation error, and provide a comparison of the proposed estimator with some state-of-the-art estimators. The estimation error is evaluated

by the normalized mean-square error, namely

$$\text{NMSE}\left(\hat{\mathbf{R}}\right) = \frac{\mathbb{E}\left\|\hat{\mathbf{R}} - \mathbf{R}_0\right\|_F^2}{\|\mathbf{R}_0\|_F^2}, \tag{6.6.1}$$

where all of the matrices are normalized by their trace. The expected value is approximated by 100 Monte Carlo simulations. In the following, we mainly compare the performance of four estimators, namely, the SCM, unconstrained Tyler's estimator (fixed-point equation of (6.2.2)), COCA (solution to (6.2.6)), and the proposed structure constrained Tyler's estimator. The samples in all of the simulations of this section, if not otherwise specified, are *i.i.d.* following $\mathbf{x}_i \sim \sqrt{\tau}\mathbf{u}$, where $\tau \sim \chi^2$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_0)$. The dimension K is set to be 15.

All simulations were coded in MATLAB performed on a PC with a 3.20 GHz i5-3470 CPU and 8 GB RAM. Convergence criteria for all presented algorithms is set to be $|L(\mathbf{R}_{t+1}) - L(\mathbf{R}_t)|/\max(1, |L(\mathbf{R}_t)|) \leq 10^{-6}$. For all algorithms that involve a convex programming, namely, COCA (6.2.6), Algorithm 1, and Algorithm 4, we use CVX [116, 117] with solver MOSEK. In the following simulations, all of the algorithms (except for COCA) shared the same initial point, which was randomly generated each time the data set changed.

6.6.1 Toeplitz Structure

In this simulation, $\mathbf{R}_0$ is set to be a Toeplitz matrix. The parameter $\mathbf{R}_0$ is set to be $\mathbf{R}(\beta)$, whose ij-th entry is of the form

$$\left(\mathbf{R}(\beta)\right)_{ij} = \beta^{|i-j|}. \tag{6.6.2}$$

Fig. 6.1 shows the NMSE of the estimators with $\beta = 0.8$. The result indicates that the structure constrained Tyler's estimator achieves the smallest estimation error. In addition, we see that although the circulant embedding algorithm (Algorithm 6.2) with $L = 2K - 1$ approximately solves the Toeplitz structure constrained problem, it achieves virtually the same estimation error as imposing the Toeplitz structure and solving the problem via the sequential SDP algorithm (Algorithm 6.1). However, the computational cost of circulant embedding is much lower than that of sequential SDP and COCA, as shown in the average time cost plotted in Fig. 6.2.

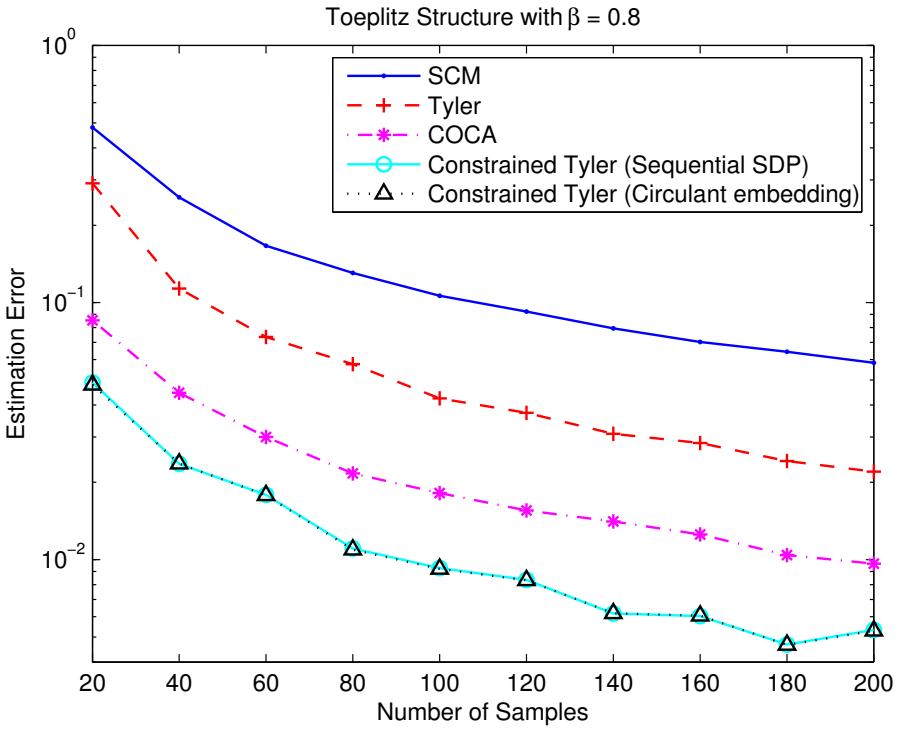


Figure 6.1: The estimation error (NMSE) of different estimators under the Toeplitz structure of the form (6.6.2).

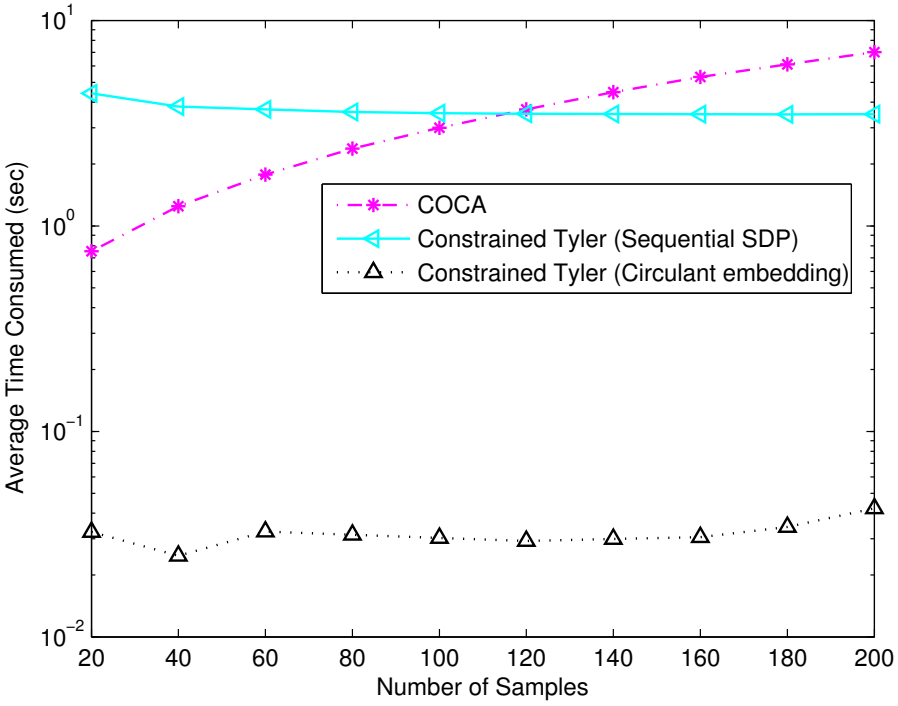


Figure 6.2: Average time (in seconds) consumed by COCA and the constrained Tyler's estimator via sequential SDP (Algorithm 1) and circulant embedding (Algorithm 2).

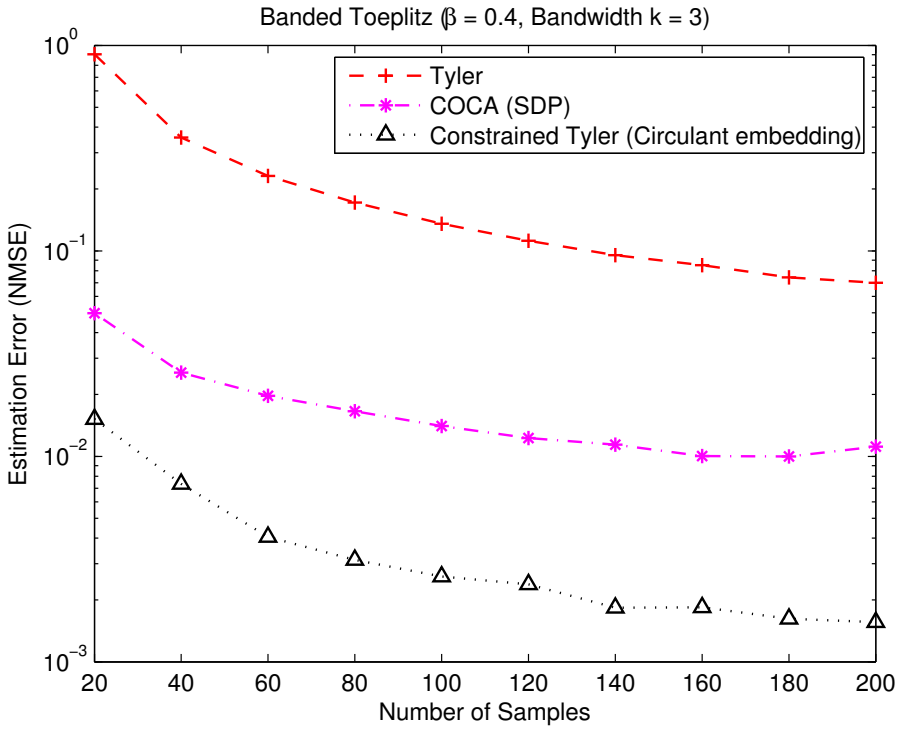


Figure 6.3: The estimation error (NMSE) of different estimators under the banded Toeplitz structure.

6.6.2 Banded Toeplitz Structure

Next we investigate the case that $\mathbf{R}_0$ is a k-banded Toeplitz matrix $B_k(\mathbf{R}_0)$, where $B_k(\mathbf{R}_0)$ defines a matrix with the ij-th entry equals to that of $\mathbf{R}_0$ if $|i-j| \leq k$, and equals zero otherwise. $\mathbf{R}_0 = \mathbf{R}(0.4)$ and the bandwidth k is chosen to be 3. The NMSE is plotted in Fig. 6.3, where the constrained Tyler's estimator achieves the smallest estimation error. Fig. 6.4 plots the average time consumed by COCA and the constrained Tyler's estimator. As the number of semidefinite constraints that COCA has is proportional to N, the time consumption is increasing in N, while the time cost by the algorithm for the constrained Tyler's estimator remains roughly the same as N grows. When N is small, the algorithm for COCA runs faster than ours since the scale of the SDP that COCA solves is small. In the regime that N is large, the computational cost of COCA increases, as reflected both in the time and the memory required to run the algorithm.

In the third simulation, we consider $\mathbf{R}_0$ being a non-banded Toeplitz matrix with the property that $(\mathbf{R}_0)_{ij}$ decays rapidly as $|i-j|$ increases. We investigate the cases of $\mathbf{R}_0 = \mathbf{R}(0.4)$ (fast decay) and $\mathbf{R}_0 = \mathbf{R}(0.8)$ (slow decay) and impose a banded Toeplitz structure on the Tyler's estimator with a varying bandwidth k to regularize the estimator. Fig. 6.5 shows

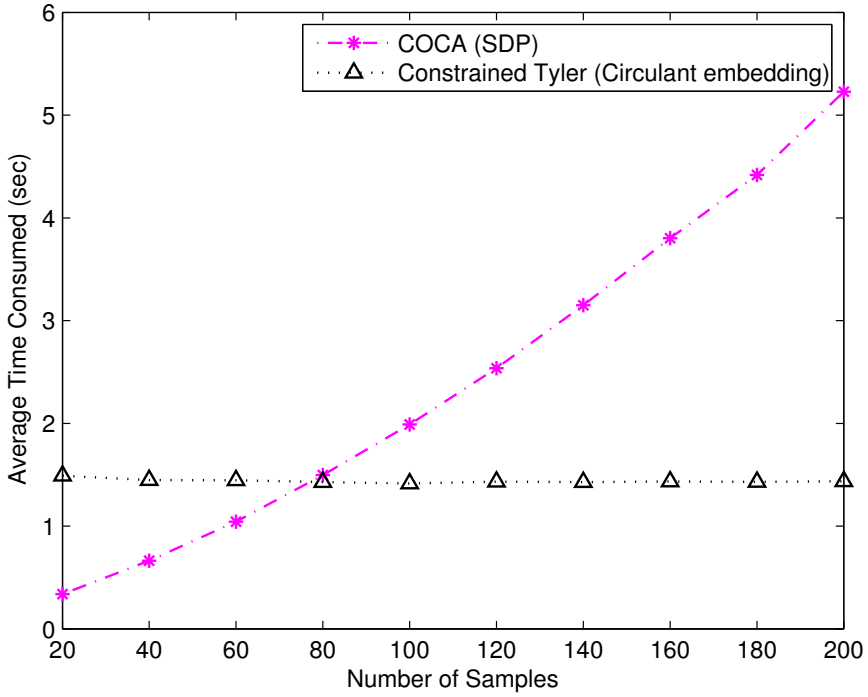


Figure 6.4: Average time (in seconds) consumed by COCA and constrained Tyler's estimator.

that the smallest error is obtained when $k = 3$ in the $\beta = 0.4$ case, and when $k = 13$ in the $\beta = 0.8$ case. In either case, with the right choice of bandwidth k, the regularized estimator outperforms the unbanded one when the number of samples is relatively small compared to the dimension of the covariance matrix to be estimated.

6.6.3 Direction of Arrival Estimation

In this subsection, we examine the robustness of the proposed estimator in the context of the direction of arrival estimation problem.

A number of $m = 5$ random signals are assumed arriving from directions $-10°$, $10°$, $15°$, $35°$, $40°$ with equal power $p = 1$ and the noise power is set to be $\sigma^2 = 0.1$. The received signal is assumed to be elliptically distributed. The number of sensors is $K = 15$.

We first estimate $\mathbf{R}$ and then apply the MUSIC algorithm to estimate the arriving angles. The performance of SCM, Tyler's estimator, COCA and the constrained Tyler's estimator are compared. For the latter two estimators, which require a specification of the structure set $\mathcal{S} = \left\{\mathbf{R}|\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H\right\}$ parameterized by $\mathbf{P} \succeq \mathbf{0}$, we construct the matrix $\mathbf{A}\left(\boldsymbol{\theta}_e\right)$ according to (6.4.5) and (6.4.6) with $\boldsymbol{\theta}_e = \left[-90^\circ, -85^\circ, \ldots, 80^\circ, 85^\circ\right]$. Fig. 6.6 shows the estimated arrival direction using different estimators with the number of snapshots $N = 20$, and only

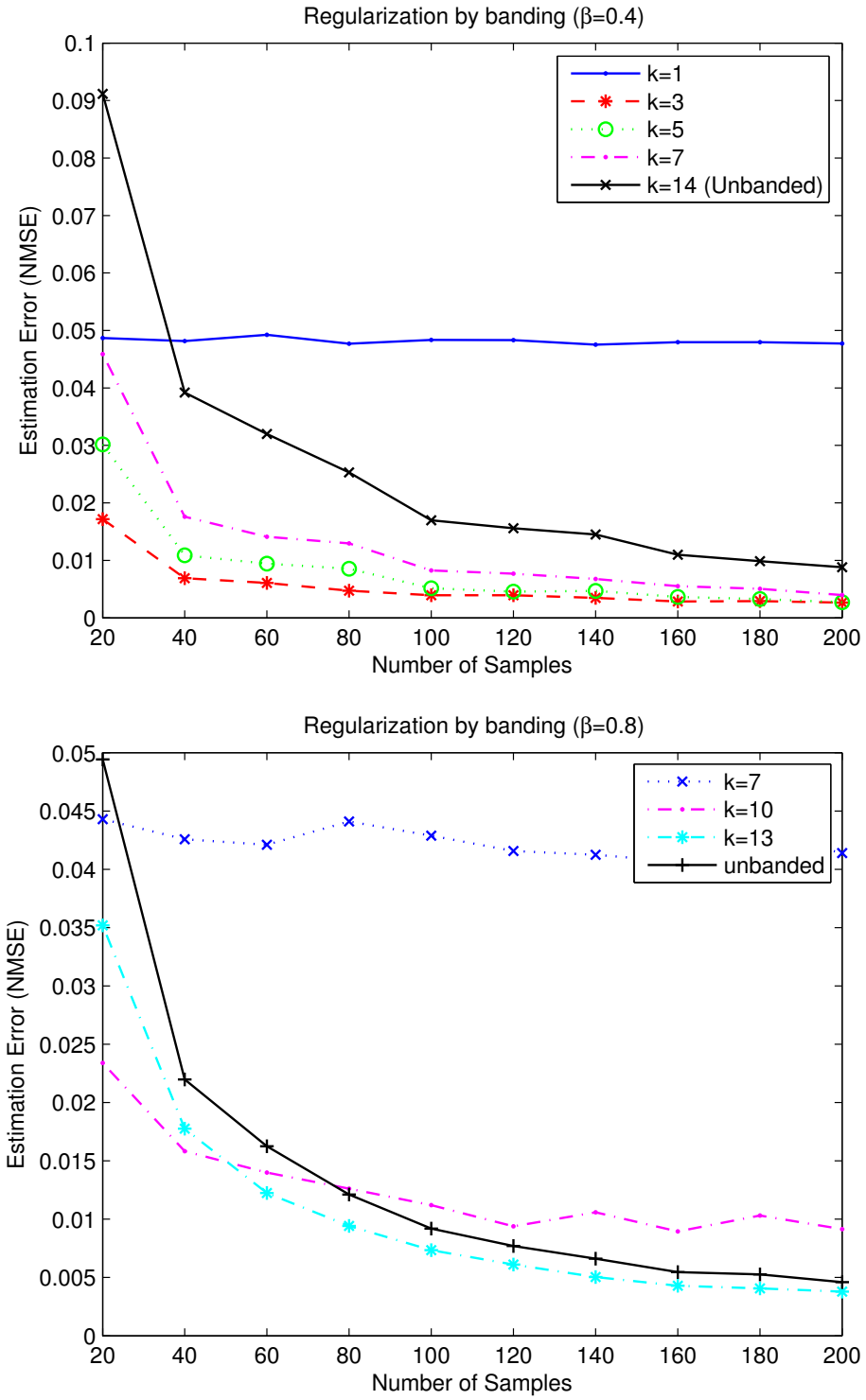


Figure 6.5: NMSE of the regularized Tyler's estimator by imposing the banded Toeplitz structure of different bandwidth k when $\mathbf{R}_0 = \mathbf{R}\,(0.4)$ and $\mathbf{R}_0 = \mathbf{R}\,(0.8)$.

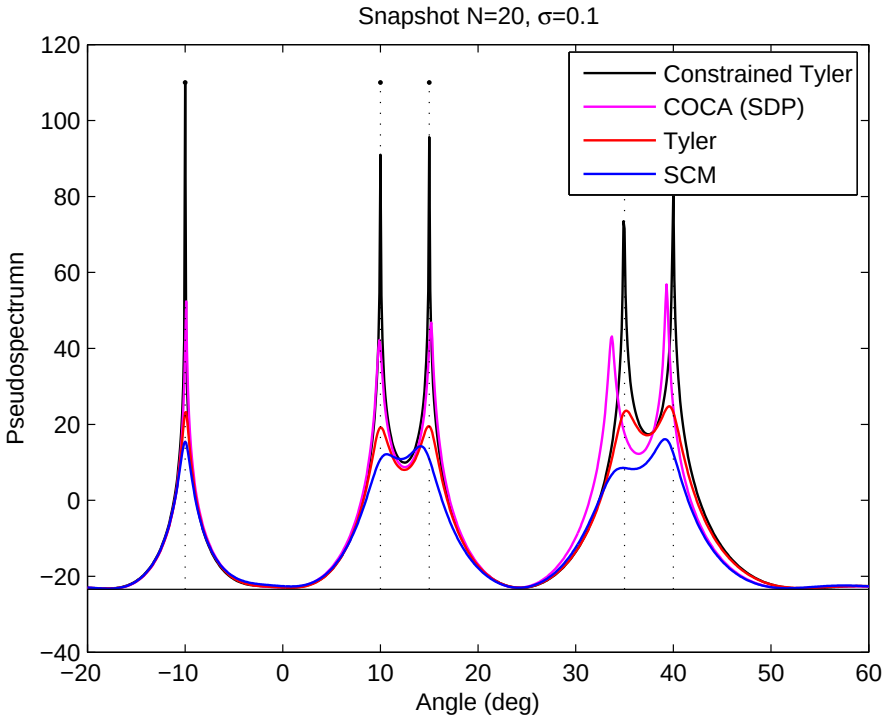


Figure 6.6: Arrival angle estimated by MUSIC with different covariance estimators.

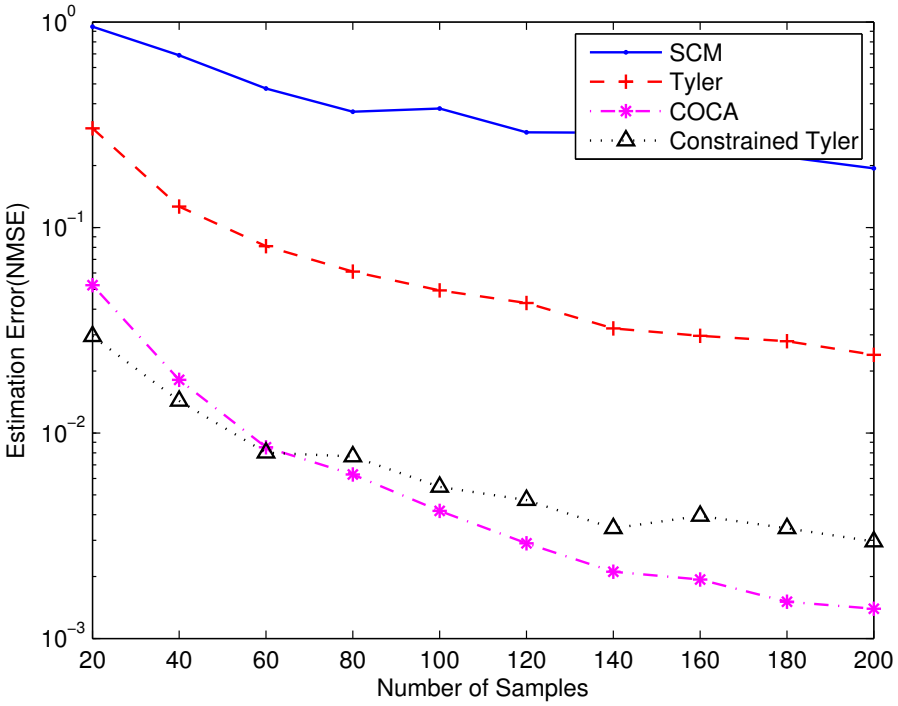
the constrained Tyler's estimator correctly recovers all of the arriving angles.

Fig. 6.7 shows the performance of different estimators in terms of NMSE and the estimation error of noise subspace evaluated by

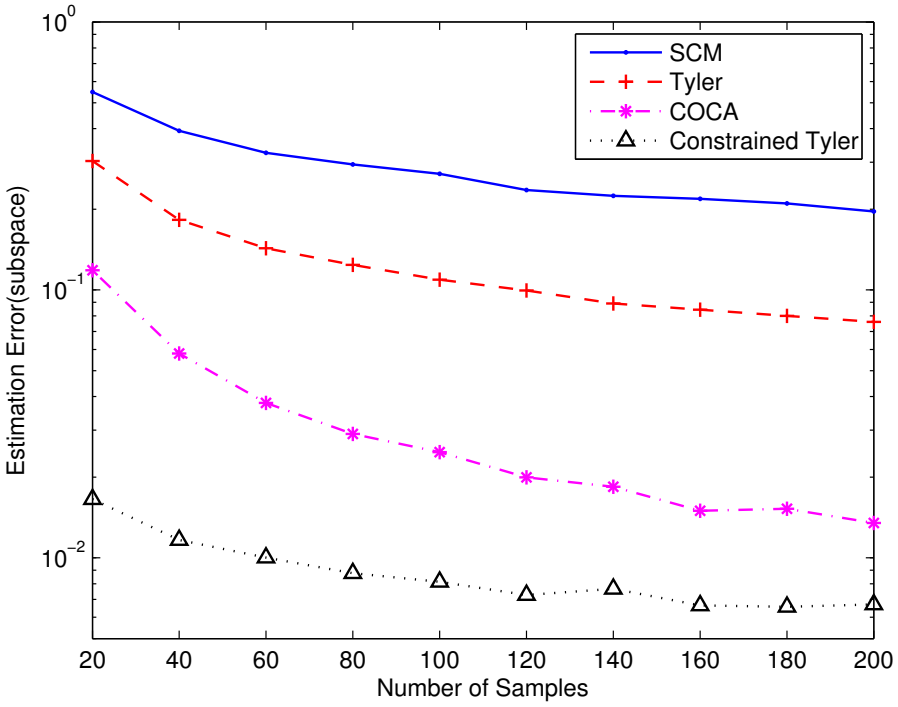
$$\left\| \hat{\mathbf{E}}_c \hat{\mathbf{E}}_c^H - \mathbf{E}_c \mathbf{E}_c^H \right\|_F, \tag{6.6.3}$$

with N varying from 20 to 200, where $\mathbf{E}_c$ denotes the noise subspace and $\hat{\mathbf{E}}_c$ denotes its estimate. Fig. 7 (a) reveals that the constrained Tyler's estimator achieves the smallest NMSE when N is small, while COCA performs better when N is large. However, Fig. 7 (b) indicates that the constrained Tyler's estimator can estimate the noise subspace more accurately for all values of N, which is beneficial for algorithms that are based on $\hat{\mathbf{E}}_c$ such as MUSIC.

The average time cost by COCA and the constrained Tyler's estimator is plotted in Fig. 6.8. It can be seen that the proposed method is much faster than COCA. In addition, unlike COCA, the consumed time of our algorithm is not sensitive to the number of samples N.



(a) NMSE



(b) Subspace error

Figure 6.7: The estimation error of different estimators under the DOA structure: (a) NMSE, (b) estimation error of the noise subspace given by different estimators evaluated by (6.6.3).

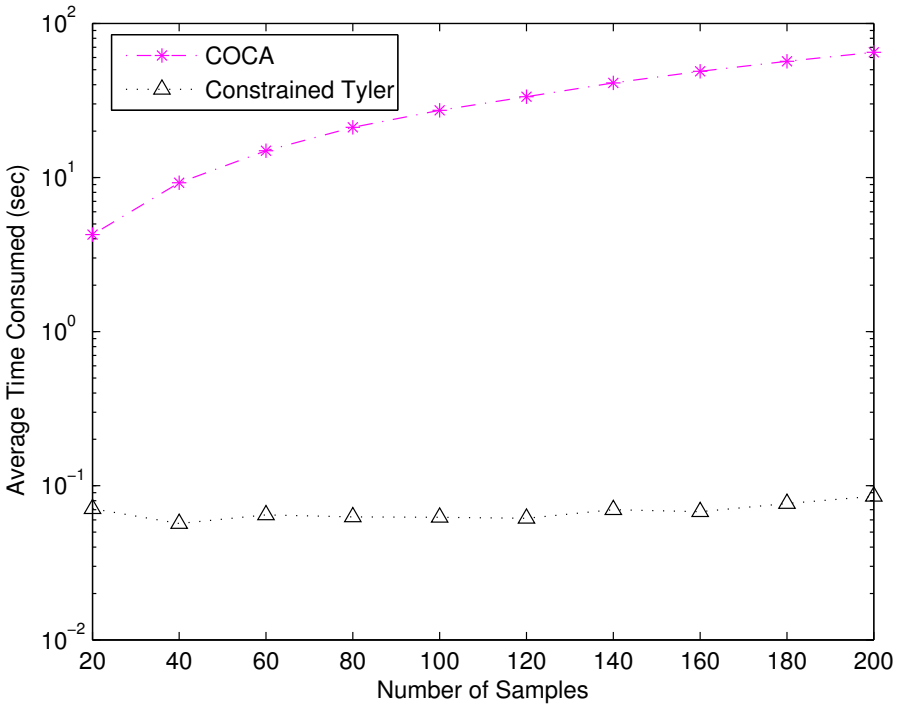


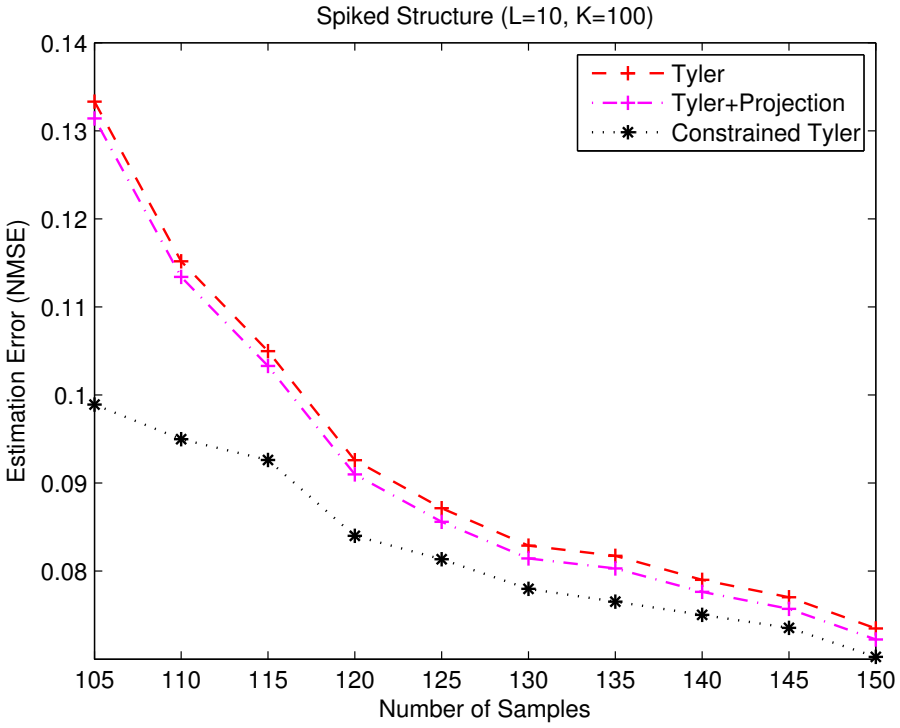
Figure 6.8: Average time (in seconds) consumed per data set by COCA and constrained Tyler's estimator.

6.6.4 Spiked Covariance Structure

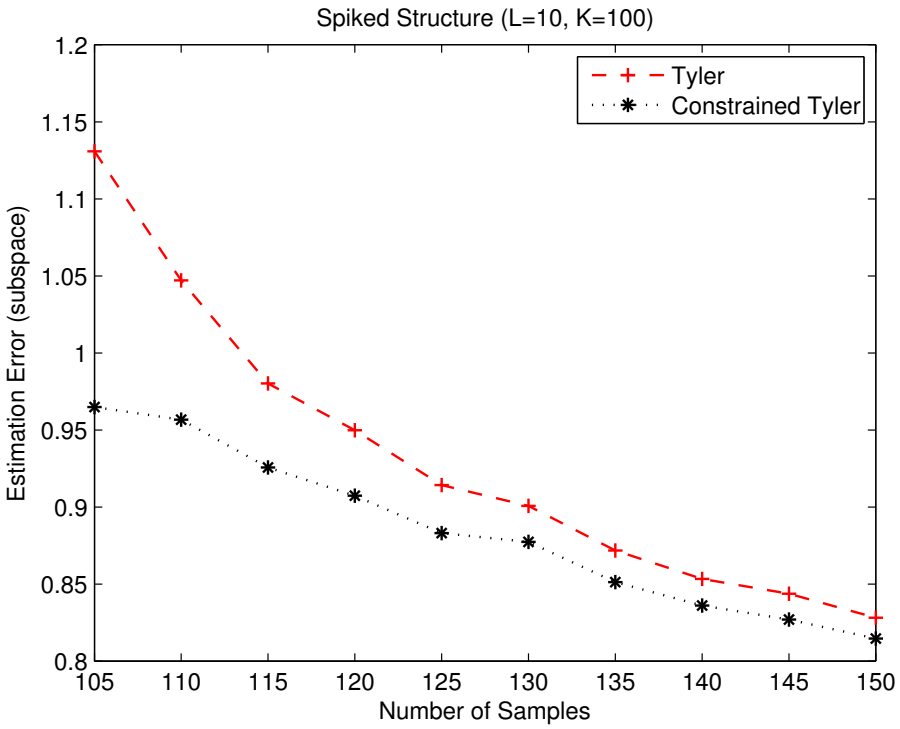
We construct the true covariance $\mathbf{R}_0$ by the following model:

$$\mathbf{R}_0 = \sum_{j=1}^{L} p_j \mathbf{a}_j \mathbf{a}_j^H + \sigma^2 \mathbf{I},$$

where the $\mathbf{a}_j$'s are randomly generated orthonormal basis and the p_j's are randomly generated corresponding eigenvectors uniformly distributed in $[0.01, 1]$. σ^2 is set to be 0.01. The number of spikes $L = 10$ is assumed to be known in prior. The matrix dimension is fixed to be $K = 100$, and the number of samples is varied from $N = 105$ to $N = 150$. As COCA applies only for convex structural set and cannot be used here, we replace it by the projected Tyler's estimator, which is a two step procedure that first obtains the Tyler's estimator and then performs projection according to (6.5.4). Fig. 6.9 shows that imposing the spiked structure helps in reducing the NMSE and subspace estimation error measured by (6.6.3).



(a) NMSE



(b) Subspace error

Figure 6.9: The estimation error of different estimators under the spiked covariance structure: (a) NMSE, (b) estimation error of the noise subspace given by different estimators evaluated by (6.6.3).

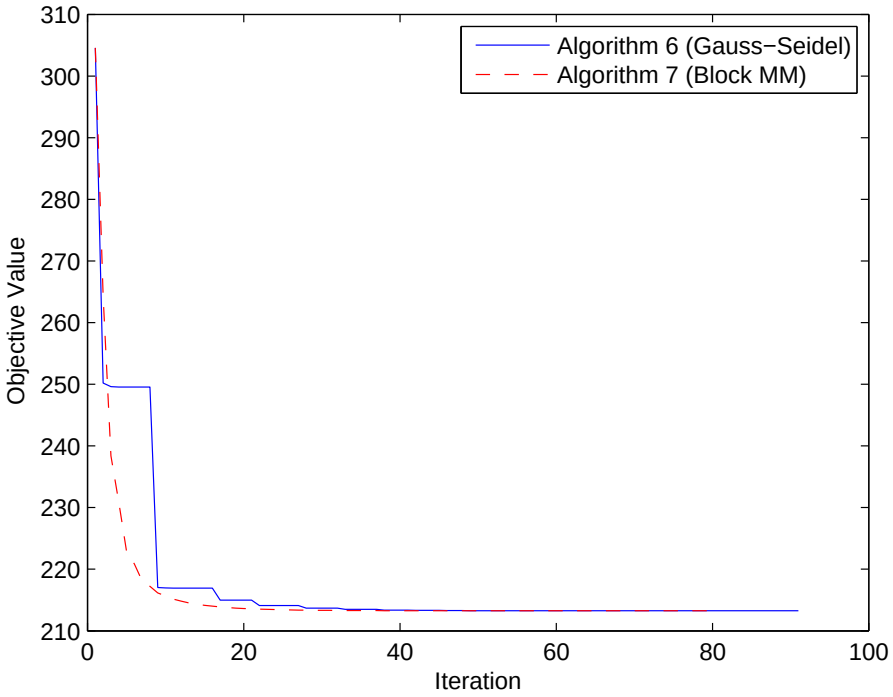


Figure 6.10: Convergence Comparison of Algorithm 6.6 and 6.7 under the Kronecker structure.

6.6.5 Kronecker Structure

The parameters are set to be $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{B}_0 = \mathbf{R}\,(0.8)$, $p = 10$, $q = 8$, in the simulations. We first plot the convergence curve of Algorithms 6.6 and 6.7 with the number of samples $N = 4$ in Fig. 6.10. The two algorithms converges in roughly the same number of iterations, and the objective value corresponds to Algorithm 6.7 (block MM) decreases more smoothly than Algorithm 6.6 (Gauss-Seidel), as the latter is a double loop algorithm while the former is a single loop algorithm.

Fig. 6.11 plots the NMSE of Tyler's estimator with a Kronecker constraint on $\mathbf{R}$ and that with both a Kronecker constraint on $\mathbf{R}$ and a Toeplitz constraint on $\mathbf{B}$. We can see that further imposing a Toeplitz structure on $\mathbf{B}$ helps in reducing the estimation error.

6.7 Conclusion

In this chapter, we have discussed the problem of robustly estimating the covariance matrix with a prior structure information. The problem has been formulated as minimizing the negative log-likelihood function of the angular central Gaussian distribution subject to the prior

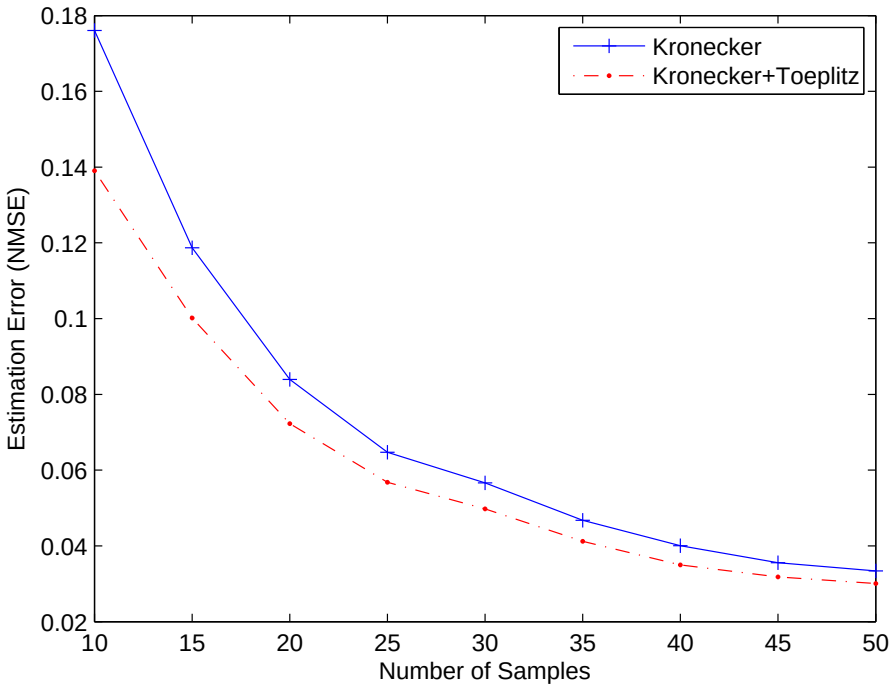


Figure 6.11: NMSE of Tyler's estimator with a Kronecker structural constraint versus that with both a Kronecker and a Toeplitz structural constraint.

structural constraint. For the general convex constraint, we have proposed a sequential convex programming algorithm based on the majorization minimization framework. The algorithm has been particularized with higher computational efficiency for several specific structures that are widely considered in the signal processing community. The spiked covariance model and the Kronecker structure, although belonging to the non-convex constraint, are also discussed and shown to be computationally tractable. The proposed estimator has been shown outperform the state-of-the-art methods in the numerical section.

Chapter 7

Conclusion

In general terms, this dissertation has considered the robust estimation of a covariance matrix.

The dissertation has been outlined in Chapter 1, where we have explained the necessity of robust estimators of a covariance matrix and the challenges of applying classical robust estimators to nowadays problems.

In Chapter 2, we have presented the majorization-minimization algorithm framework, which serves as the main tool to derive efficient algorithms throughout this dissertation.

In Chapter 4, the Tyler's scatter estimator has been introduced as one of the M-estimators for elliptically distributed samples. We have considered the problem of adapting the estimator to the high dimension regime. In particular, we have studied two kinds of shrinkage Tyler's estimators and provided the necessary and sufficient conditions for their existence. We have also established their uniqueness and shown that they are equivalent. Algorithms with provable convergence have been derived to compute the two estimators efficiently under the guidelines of the MM principle.

In Chapter 5, we have considered the joint estimation of mean and covariance matrix in the high dimension regime. Based on the cost function of the Cauchy MLE, which provides a good resistance against outliers, we have proposed a shrinkage estimator that shrinks the estimates towards a prior target mean and covariance matrix. Conditions for the existence and uniqueness of the estimator have been provided. Several algorithms have been derived and compared under the (block) MM framework.

In Chapter 6, we have investigated the problem of robustly estimating a structured covariance matrix. The problem has been formulated as minimizing the cost function of Tyler's estimator under a structural constraint. A sequential SDP algorithm has been derived based

on the MM algorithm for a general convex structural constraint. The algorithm then has been tailored to some special structures that are of wide interest in signal processing with closed-form update per iteration. Two non-convex structures have also been discussed at the end of the chapter.

Bibliography

[1] S. Visuri, H. Oja, and V. Koivunen, "Subspace-based direction-of-arrival estimation using nonparametric statistics," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 2060–2073, 2001.

[2] Y. Chen, A. Wiesel, and A. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4097–4107, 2011.

[3] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," *arXiv preprint arXiv:1311.6567*, 2013.

[4] V. Koivunen, "Nonlinear filtering of multivariate images under robust error criterion," *IEEE Trans. Image Processing*, vol. 5, no. 6, pp. 1054–1060, 1995.

[5] S. Visuri, H. Oja, and V. Koivunen, "Nonparametric statistics for subspace based frequency estimation," in *Proc. 10th European Signal Processing Conf. (EUSIPCO 2000)*, Tampere, Finland, 2000.

[6] F. Rubio, X. Mestre, and D. P. Palomar, "Performance analysis and optimal selection of large minimum variance portfolios under estimation risk," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 4, pp. 337–350, 2012.

[7] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

[8] U. A. Müller, M. M. Dacorogna, and O. V. Pictet, "Heavy tails in high-frequency financial data," *A Practical Guide to Heavy Tails: Statistical Techniques and Applications. Boston: Birkhaüser*, pp. 55–77, 1998.

[9] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 4, no. 1, pp. pp. 51–67, 1976.

[10] P. L. Davies, "Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices," *Ann. Statist.*, vol. 15, no. 3, pp. pp. 1269–1292, 1987.

[11] S. Van Aelst and P. Rousseeuw, "Minimum volume ellipsoid," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 1, no. 1, pp. 71–82, 2009.

[12] R. W. Butler, P. L. Davies, and M. Jhun, "Asymptotics for the minimum covariance determinant estimator," *Ann. Statist.*, vol. 21, no. 3, pp. 1385–1400, 09 1993.

[13] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*, ser. Wiley Series in Probability and Statistics. Wiley, 2006.

[14] P. Huber, *Robust Statistics*, ser. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2004.

[15] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 03 1987.

[16] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped cauchy distribution," *J. Appl. Statist.*, vol. 15, no. 2, pp. 247–254, 1988.

[17] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.

[18] G. Frahm, "Generalized elliptical distributions: theory and applications," Ph.D. dissertation, Universität zu Köln, 2004.

[19] Y. Abramovich and N. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *Proc. IEEE Int Conf. Acoust., Speech, Signal Process. (ICCASP), 2007.*, vol. 3, Honolulu, HI, April 2007, pp. III–1105–III–1108.

[20] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 29–38, 2012.

[21] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] F. J. Fabozzi, P. N. Kolm, D. Pachamanova, and S. M. Focardi, *Robust portfolio optimization and management*. John Wiley & Sons, 2007.

[23] D. R. Fuhrmann and M. I. Miller, "On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices," *IEEE Trans. Inf. Theory*, vol. 34, no. 4, pp. 722–729, 1988.

[24] A. Dembo, C. Mallows, and L. Shepp, "Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1206–1212, 1989.

[25] M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances," *Proc. IEEE*, vol. 75, no. 7, pp. 892–907, 1987.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[27] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, pp. 199–227, 2008.

[28] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 478–491, 2008.

[29] P. Shah and V. Chandrasekaran, "Group symmetry and covariance regularization," *Electronic Journal of Statistics*, vol. 6, pp. 1600–1640, 2012.

[30] P. Wirfalt and M. Jansson, "On Kronecker and linearly structured covariance matrix estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 6, pp. 1536–1547, March 2014.

[31] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970, vol. 30.

[32] T. T. Wu and K. Lange, "The MM alternative to EM," *Statistical Science*, vol. 25, no. 4, pp. 492–505, 2010.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[34] T. Blumensath, M. Yaghoobi, and M. E. Davies, "Iterative hard thresholding and l_0 regularisation," in *Proc. IEEE Int Conf. Acoust., Speech, Signal Process. (ICASSP), 2007*, vol. 3. IEEE, 2007, pp. III–877.

[35] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[36] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[37] G. Marjanovic, M. O. Ulfarsson, and A. O. Hero III, "Mist: l_0 sparse linear regression with momentum," *arXiv preprint arXiv:1409.7193*, 2014.

[38] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 899–925, 2013.

[39] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine Learning*, vol. 85, no. 1-2, pp. 3–39, 2011.

[40] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1627–1642, 2015.

[41] Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5143–5156, 2014.

[42] J. Song, P. Babu, and D. P. Palomar, "Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions," *IEEE Trans. Signal Process.*, vol. 63, no. 12, pp. 3096–3109, June 2015.

[43] Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *arXiv preprint arXiv:1506.05215*, 2015.

[44] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2178–2191, 2009.

[45] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Int Conf. Acoust., Speech, Signal Process. (ICASSP), 2011*. IEEE, 2011, pp. 1980–1983.

[46] Y. Cao, P. Eggermont, and S. Terebey, "Cross Burg entropy maximization and its application to ringing suppression in image reconstruction," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 286–292, Feb 1999.

[47] J. M. Bioucas-Dias, M. A. Figueiredo, and J. P. Oliveira, "Total variation-based image deconvolution: a majorization-minimization approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2006*, vol. 2. IEEE, 2006, pp. II–II.

[48] M. A. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization–minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, 2007.

[49] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[50] J. Song, P. Babu, and D. P. Palomar, "Optimization methods for designing sequences with low autocorrelation sidelobes," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3998–4009, Aug 2015.

[51] ——, "Sequence design to minimize the weighted integrated and peak sidelobe levels," *arXiv preprint arXiv:1506.04234*, 2015.

[52] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

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

[54] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, 1994.

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

[56] M. Jamshidian and R. I. Jennrich, "Conjugate gradient acceleration of the EM algorithm," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 221–228, 1993.

[57] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 425–437, 1995.

[58] ——, "A quasi-newton acceleration of the EM algorithm," *Statistica sinica*, vol. 5, no. 1, pp. 1–18, 1995.

[59] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[60] M. W. Jacobson and J. A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2411–2422, 2007.

[61] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward-backward algorithm," 2013.

[62] J. Mairal, "Incremental majorization-minimization optimization with application to large-scale machine learning," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 829–855, 2015.

[63] T. A. Louis, "Finding the observed information matrix when using the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 226–233, 1982.

[64] I. Meilijson, "A fast improvement to the em algorithm on its own terms," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 127–138, 1989.

[65] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. Conf. Info. Sci. Sys.*, 1993, pp. 611–616.

[66] R. M. Lewitt and G. Muehllehner, "Accelerated iterative reconstruction for positron emission tomography based on the em algorithm for maximum likelihood estimation," *IEEE Transactions on Medical Imaging*, vol. 5, no. 1, pp. 16–22, 1986.

[67] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth, "A comparison of new and old algorithms for a mixture estimation problem," *Machine Learning*, vol. 27, no. 1, pp. 97–119, 1997.

[68] E. Bauer, D. Koller, and Y. Singer, "Update rules for parameter estimation in bayesian networks," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1997, pp. 3–13.

[69] R. Salakhutdinov and S. Roweis, "Adaptive overrelaxed bound optimization methods," in *ICML*, 2003, pp. 664–671.

[70] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[71] N. Laird, N. Lange, and D. Stram, "Maximum likelihood computations with repeated measures: application of the em algorithm," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 97–105, 1987.

[72] M. Jamshidian and R. I. Jennrich, "Acceleration of the EM algorithm by using quasi-newton methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 3, pp. 569–587, 1997.

[73] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1130–1142, 2006.

[74] R. Varadhan and C. Roland, "Simple and globally convergent methods for accelerating the convergence of any em algorithm," *Scandinavian Journal of Statistics*, vol. 35, no. 2, pp. 335–353, 2008.

[75] H. Zhou, D. Alexander, and K. Lange, "A quasi-newton acceleration for high-dimensional optimization algorithms," *Statistics and computing*, vol. 21, no. 2, pp. 261–273, 2011.

[76] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 1995.

[77] E. Schlossmacher, "An iterative technique for absolute deviations curve fitting," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 857–859, 1973.

[78] P. Oğuz-Ekim, J. P. Gomes, J. Xavier, and P. Oliveira, "Robust localization of nodes and time-recursive tracking in sensor networks using noisy range measurements," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3930–3942, 2011.

[79] K. Lange and H. Zhou, "MM algorithms for geometric and signomial programming," *Mathematical programming*, vol. 143, no. 1-2, pp. 339–356, 2014.

[80] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography." *IEEE transactions on medical imaging*, vol. 14, no. 1, pp. 132–137, 1994.

[81] K. Lange and J. A. Fessler, "Globally convergent algorithms for maximum a posteriori transmission tomography," *IEEE Transactions on Image Processing*, vol. 4, no. 10, pp. 1430–1438, 1995.

[82] A. Hjørungnes, *Complex-valued matrix derivatives: with applications in signal processing and communications*. Cambridge University Press, 2011.

[83] M. Chiang, C. W. Tan, D. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *Wireless Communications, IEEE Transactions on*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[84] H. Zhou, L. Hu, J. Zhou, and K. Lange, "Mm algorithms for variance components models," *arXiv preprint arXiv:1509.07426*, 2015.

[85] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964.

[86] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *J. Amer. Statist. Assoc.*, vol. 84, no. 408, pp. 881–896, 1989.

[87] A. Lucas, "Robustness of the student t based M-estimator," *Comm. Statist. Theory Methods*, vol. 26, no. 5, pp. 1165–1182, 1997.

[88] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *Ann. Statist.*, vol. 19, no. 4, pp. 2102–2119, 12 1991.

[89] K. S. Tatsuoka and D. E. Tyler, "On the uniqueness of S-functionals and M-functionals under nonelliptical distributions," *Annals of Statistics*, pp. 1219–1243, 2000.

[90] J. T. Kent, D. E. Tyler, and Y. Vard, "A curious likelihood identity for the multivariate t-distribution," *Comm. Statist. Simulation Comput.*, vol. 23, no. 2, pp. 441–453, 1994.

[91] Y. I. Abramovich, "A controlled method for adaptive optimization of filters using the criterion of maximum signal-to-noise ratio," *Radio Eng. Elect. Phys*, vol. 25, no. 3, pp. 87–95, 1981.

[92] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[93] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365 – 411, 2004.

[94] ——, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of empirical finance*, vol. 10, no. 5, pp. 603–621, 2003.

[95] ——, "Honey, i shrunk the sample covariance matrix," *UPF Economics and Business Working Paper*, no. 691, 2003.

[96] M. Zhang, F. Rubio, D. Palomar, and X. Mestre, "Finite-sample linear filter optimization in wireless communications and financial systems," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 5014–5025, 2013.

[97] T. Lancewicki and M. Aladjem, "Multi-target shrinkage estimation for covariance matrices," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6380–6390, Dec 2014.

[98] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, 2002.

[99] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[100] R. Couillet and M. R. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *arXiv preprint arXiv:1401.4083*, 2014.

[101] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6059–6070, 2014.

[102] Y. Sun, P. Babu, and D. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5143–5156, Oct 2014.

[103] D. B. Williams and D. H. Johnson, "Robust estimation of structured covariance matrices," *IEEE Trans. Signal Process.*, vol. 41, no. 9, pp. 2891–2906, 1993.

[104] I. Soloveychik and A. Wiesel, "Group symmetry and non-Gaussian covariance estimation," *arXiv preprint arXiv:1306.4103*, 2013.

[105] ——, "Tyler's covariance matrix estimator in elliptical models with convex structure," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5251–5259, Oct 2014.

[106] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 29–38, 2012.

[107] ——, "Geodesic convexity and covariance estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6182–6189, 2012.

[108] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, Nov 2012.

[109] M. I. Miller and D. L. Snyder, "An alternating maximization of the entropy/likelihood function for image reconstruction and spectrum estimation," in *Proc. SPIE*. San Diego, CA: International Society for Optics and Photonics, 1986, pp. 163–166.

[110] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, pp. 295–327, 2001.

[111] D. Ruppert, *Statistics and data analysis for financial engineering*. Springer Science & Business Media, 2010.

[112] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[113] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Athena Scientific Belmont, 2003.

[114] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[115] T. Rapcsák, "Geodesic convexity in nonlinear optimization," *Journal of Optimization Theory and Applications*, vol. 69, no. 1, pp. 169–183, 1991.

[116] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[117] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.