이학박사학위논문

Positive-definite correction of covariance matrix estimators via linear shrinkage

선형 축소를 통한 공분산 행렬 추정량의 양정치 보정

2015년 8월

서울대학교 대학원

통계학과

최 영 근

Positive-definite correction
# of covariance matrix estimators
via linear shrinkage
선형 축소를 통한
공분산 행렬 추정량의 양정치 보정

지도교수 임 요 한

이 논문을 이학박사 학위논문으로 제출함

2015년 4월

서울대학교 대학원

통계학과

최 영 근

최영근의 이학박사 학위논문을 인준함

2015년 6월

위 원 장 Myunghee Cho Paik (인)

부위원장 원 중 호 (인)

위 원 김 용 대 (인)

위 원 임 요 한 (인)

위 원 전 용 호 (인)

Positive-definite correction of covariance matrix estimators via linear shrinkage

by

Young-Geun Choi

A Thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
August, 2015

Abstract

Young-Geun Choi
The Department of Statistics
The Graduate School
Seoul National University

In this paper, we study the positive definiteness (PDness) problem in covariance matrix estimation. For high dimensional data, the most common sample covariance matrix performs poorly in estimating the true matrix. Recently, as an alternative to the sample covariance matrix, many regularized estimators are proposed under structural assumptions on the true including sparsity. They are shown to be asymptotically consistent and rate-optimal in estimating the true covariance matrix and its structure. However, many of them do not take into account the PDness of the estimator and produce a non-PD estimate. Otherwise, additional regularizations (or constraints) are required on eigenvalues which make both the asymptotic analysis and computation much harder. To achieve the PDness, we propose a simple one-step procedure to update the regularized covariance matrix estimator which is not necessarily PD in finite sample. We revisit the idea of linear shrinkage (Stein, 1956; Ledoit and Wolf, 2004) and propose to take a convex combination between the first stage covariance matrix estimator (the regularized covariance matrix without PDness) and a given form of diagonal matrix. The proposed one-step correction, which we denote as LSPD (linear shrinkage for positive definiteness) estimator, is shown to preserve the asymptotic properties of the first stage estimator if the shrinkage parameters are carefully selected. In addition, it has a closed form expression and its computation is optimization-free, unlike existing sparse PD estimators (Rothman, 2012; Xue et al., 2012). The LSPD estimator is numerically compared with other sparse PD estimators to understand its finite sample properties as well as its computational gain. Finally, it is applied to two multivariate procedures relying on the covariance matrix estimator - the linear minimax classification problem posed by Lanckriet et al. (2002) and the well-known mean-variance portfolio optimization problem - and is shown to substantially improve the performance of both procedures.

keywords : *Covariance matrix; positive definitess; high-dimensional estimation; linear shrinkage.*

Student Number : 2010-23063

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Covariance matrix and its consistent estimation are involved in many multivariate statistical procedures, where the sample covariance matrix is popularly used. To date, high dimensional data are prevalent for which the sample covariance matrix is known to be inconsistent (Marcenko and Pastur, 1967). To resolve the difficulty from high dimensionality, regularized procedures or estimators have been proposed under various structural assumptions on the true matrix. For instance, if the true covariance matrix is assumed to be sparse or banded, one thresholds the elements of the sample covariance matrix to satisfy the assuamptions (Bickel and Levina, 2008a,b; Cai et al., 2010; Cai and Low, 2011; Cai and Yuan, 2012; Cai and Zhou, 2012; Rothman et al., 2009) or penalizes the likelihood function using the l_1-norm of the covariance matrix (Bien and Tibshirani, 2011; Lam and Fan, 2009). The asymptotics of the regularized estimators are well understood and, particularly, they are shown to be consistent in estimating the true covariance matrix and its support (positions of its non-zero elements).

The main interest of this paper is positive definiteness (PDness) of covariance matrix estimator. The PDness of the covariance matrix is an important property since many multivariate statistical procedures are well defined only when the covariance matrix (or its estimate) is PD. In addition, it is an interesting feature by itself since it implies the positive variance of any linear combination of variables. However, the regularized covariance matrix estimators recently studied are often not PD. This is because they more focus on the given structural assumptions and do not impose PDness on the covariance matrix. For example, the banding or thresholding method regularizes the sample covariance matrix in element-wise and provides an explicit form of the estimator that satisfies the given structural assumptions (Bickel and Levina, 2008b; Rothman et al., 2009). Nonetheless, the eigen-structures of the estimator are completely unknown and, without doubt, the resulting covariance matrix estimate is often not PD.

A few efforts are made to find an estimator which attains both the sparsity and the PDness (Bien and Tibshirani, 2011; Lam and Fan, 2009; Liu et al., 2014; Rothman, 2012; Xue et al., 2012). In particular, the works of Rothman, Xue et al., and Liu et al. provide convex extensions of thresholding estimators. They view the soft-thresholding of the sample covariance (correlation) matrix as a convex minimization problem and add a convex penalty or constraint to the problem so that the solution is PD. We remark that all these methods are tailored to a specific regularized covariance matrix estimator - soft-thresholding estimator - and also require us to solve a large-scale optimization problem.

The estimator we propose in this paper is a generic framework of modifying a given regularized covariance matrix estimator to make it PD. The regularized estimators in the literature are often proven to be consistent and rate optimal for various matrix norms in estimating true covariance matrix $\mathbf{\Sigma}$. Thus, we aim to modify a given (or any) regularized estimator (denoted by $\widehat{\mathbf{\Sigma}}$) minorly to retain the same asymptotic properties as well as be PD. To be specific, we consider the distance minimization problem

$$\underset{\widehat{\mathbf{\Sigma}}^*}{\text{minimize}} \left\{ \left\| \widehat{\mathbf{\Sigma}}^* - \widehat{\mathbf{\Sigma}} \right\| \, : \, \gamma_{\min}(\widehat{\mathbf{\Sigma}}^*) \geq \epsilon, \; \operatorname{supp}(\widehat{\mathbf{\Sigma}}^*) = \operatorname{supp}(\widehat{\mathbf{\Sigma}}), \; \widehat{\mathbf{\Sigma}}^* = (\widehat{\mathbf{\Sigma}}^*)^\top \right\} \tag{1.1}$$

where $\epsilon > 0$ is a pre-determined small constant and $\gamma_{\min}(\widehat{\mathbf{\Sigma}}^*)$ denotes the smallest eigenvalue of $\widehat{\mathbf{\Sigma}}^*$. In solving (1.1), to make the correction simple, we further restrict the class of $\widehat{\mathbf{\Sigma}}^*$ to a family of linear shrinkage of $\widehat{\mathbf{\Sigma}}$ to the identity matrix that is

$$\widehat{\mathbf{\Sigma}}^* \equiv \mathbf{\Phi}_{\mu,\alpha}\big(\widehat{\mathbf{\Sigma}}\big) := \alpha\widehat{\mathbf{\Sigma}} + (1-\alpha)\mu\mathbf{I}, \quad \alpha \in [0,1], \; \mu \in \mathbb{R}, \tag{1.2}$$

Here, shrinking $\widehat{\Sigma}$ linearly to the identity enables us to directly analyze the eigen-struture of $\widehat{\Sigma}^*$ while preserving the support of $\widehat{\Sigma}$.

The proposed LSPD estimator has several interesting properties. First and most of all, the minimization problem (1) is analytically solvable for many popular norms and the LSPD estimator $\widehat{\mathbf{\Sigma}}^*$ can be explicitly expressible with the smallest and largest eigenvalues of the regularized estimator $\widehat{\mathbf{\Sigma}}$. Second, we analytically show that the LSPD estimator $\widehat{\mathbf{\Sigma}}$ achieves the same asymptotic properties with those of the original regularized estimator $\widehat{\mathbf{\Sigma}}$ with a suitable choice of μ and α in (2). A careful choice of μ and α can set the risk of $\widehat{\mathbf{\Sigma}}^*$ asymptotically comparable to those of $\widehat{\mathbf{\Sigma}}$. Third, the LSPD can be adapted to any possibly non-PD estimator, whereas each of existing procedures is customized to a specific regularized estimator.

The rest of the paper is organized as follows. In Section 2, we illustrate some simulated examples of estimators that have non-PD outcomes. Recent state-of-the-art sparse covariance estimators which guarantees PDness are also briefly reviewed. Our main results are presented in Section 6. The LSPD correction method is developed based on distance minimization scheme as (1.1) using (1.2). It leads to a generic framework of updating any (possibly) non-PD covariance matrix estimator to be PD, without solving optimization problem. We not only derive statistical convergence rate of LSPD-modified estimator, but discuss some implementation issues for the update procedure also. In Section 4, we numerically show that LSPD-updated soft thresholding estimator has comparable risks with recent optimization-based PD sparse covariance matrix estimators, whereas ours are computationally much simpler and faster. In Section 5, we illustrate the usefulness of LSPD-correction by plugging-in the LSPD-modified regularized covariance estimators into statistical procedures, along with real data analyses. The selected examples in this paper are linear minimax classification problem (Lanckriet et al., 2002) and Markowitz portfolio optimization with no shortsales (Jagannathan and Ma, 2003). Since the LSPD correction is applicable to any covariance matrix estimators including precision matrix (the inverse of covariance matrix) estimators, we briefly discuss this extendability in Section 6. Finally, Section 7 is for concluding remarks.

Notations. We assume all covariance matrices are of size $p \times p$. $\mathbf{S}$ denotes a sample covariance matrix from p-dimensional data. $\mathbf{\Sigma}$ is true covariance matrix, and $\widehat{\mathbf{\Sigma}}$ is its (any) estimator. For a symmetric matrix $\mathbf{A}$, $\gamma_i(\mathbf{A})$, $\gamma_{\max}(\mathbf{A})$, and $\gamma_{\min}(\mathbf{A})$ respectively indicate the i-th largest, maximum, and minimum eigenvalue of $\mathbf{A}$. In particular, $\gamma_i(\mathbf{\Sigma})$ is abbreviated to γ_i and $\gamma_i(\widehat{\mathbf{\Sigma}})$ to $\widehat{\gamma}_i$. The scaled Frobenius norm of $\mathbf{A}$ is defined by $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(\mathbf{A}^\top \mathbf{A})/p}$. The spectral norm of $\mathbf{A}$ is $\|\mathbf{A}\|_2 := \sqrt{\gamma_{\max}(\mathbf{A}^\top \mathbf{A})} \equiv \max_i |\gamma_i(\mathbf{A})|$.

Chapter 2

Literature review

In this chapter, we review recent regularization methods for the estimation of covariance matrix as well as precision matrix. Then we discuss the PDness problem, which is the main theme of this thesis, of the introduced estimators with numerical illustrations.

2.1. Regularized covariance matrix estimators

2.1.1. Banding estimators

The banded covariance matrix arises when the variables of data are ordered and serially dependent as in data for time series, climatology, or spectroscopy. Bickel and Levina (2008b) propose a class of banding estimators, which is

$$\widehat{\Sigma}_h^{\text{Band}} = (s_{ij} \cdot w_{|i-j|}), \quad w_m = I(m \leq h),\ m = 0, 1, \ldots, p-1,$$

for a fixed bandwidth h. Heuristically the estimator $\widehat{\Sigma}_h^{\text{Band}}$ discards the sample covariance s_{ij} if corresponding indices i and j are "distant". Later, Cai et al. (2010) considers a tapering estimator which is a 'smoothed' version of the banding estimator, in which the weight is differently defined as the weight w_m $(m = 0, 1, \ldots, p-1)$ as

$$w_m = \begin{cases} 1, & \text{when } m \leq h/2 \\ 2 - 2m/h, & \text{when } h/2 < m \leq h \\ 0, & \text{otherwise.} \end{cases}$$

The asymptotic properties of banding estimators is inspected on the following parameter space of 'bandable' covariance matrices,

$$\mathcal{F}_\alpha(M_0, M) = \Bigg\{\Sigma : \max_{1\leq i\leq p} \sum_{j=1}^{p} \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k, \\ \lambda_{\max}(\Sigma) \leq M_0\Bigg\},$$

which was proposed in Bickel and Levina (2008a). In $\mathcal{F}_\alpha(M_0, M)$, σ_{ij} decays to zero as (i, j) gets distant from the main diagonal and α determines how fast the decaying rate is. In this space, the convergence rate of banding estimators are given with some distributional assumptions and regularity conditions, by

$$\left\|\widehat{\Sigma} - \Sigma\right\|_2 = O_P\left(\min\left[\left(\frac{\log p}{n} + n^{-\frac{\alpha}{2\alpha+1}}\right), \sqrt{\frac{p}{n}}\right]\right).$$

Later, Cai et al. (2010) proves that this rate is optimal in minimax sense.

2.1.2. Thresholding estimators

Universal (generalized) threholding estimator 'Thresholding' technique means intuitively that one artificially discards an element with small magnititude to zero. Bickel and Levina (2008a) first proposes to hard-threshold a sample covariance matrix, and Rothman et al. (2009) generalizes the class of thresholding functions. Later Cai and Zhou (2012) proves the optimality of thresholding method in minimax sense if true covariance matrix has sparse structure.

A *thresholding operator* is defined as a function $s_\lambda : \mathbb{R} \to \mathbb{R}$ satisfying the three properties below : for fixed $\lambda \geq 0$,

(i) $|s_\lambda(z)| \leq |z|$;

(ii) $s_\lambda(z) = 0$ for $|z| \leq \lambda$;

(iii) $s_\lambda(z) = 0$ for $|z| \leq \lambda$.

$s_\lambda(\cdot)$ includes the following penalty functions widely used in the penalized regression literature:

(i) ('Hard thresholding') $s_\lambda^H(z) := z \times I(|z| > \lambda)$;

(ii) ('Soft thresholding') $s_\lambda^S(z) := \text{sign}(z)(|z| - \lambda)_+$;

(iii) ('SCAD') $s_\lambda^{SC}(z) := \lambda \left\{ I(|z| \leq \lambda) + \frac{(a\lambda - |z|)_+}{(a-1)\lambda} I(|z| > \lambda) \right\}$ for fixed $a > 2$.

Rothman et al. (2009) proposes a sample covariance matrix with generalized thresholding :

$$\widehat{\boldsymbol{\Sigma}}_\lambda^{\text{GT}} := s_\lambda(\mathbf{S}) \tag{2.1}$$

where a notational abuse $s_\lambda(\mathbf{S})$ means an elementwise operation, $s_\lambda(\mathbf{S}) := [s_\lambda(s_{ij})]_{1\le i,j\le p}$. This procedure conicides Bickel and Levina (2008a) when $s_\lambda() \equiv s_\lambda^H(\cdot)$. Consistency and convergence rate are proved to be uniform under $\mathcal{U}_\tau(q, c_0(p), M)$, a family of 'approximately sparse' covariance matrices which is defined by

$$\mathcal{U}_\tau(q, c_0(p), M) := \left\{ \mathbf{\Sigma} \,:\, \sigma_{ii} \le M,\, \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p) \right\}. \tag{2.2}$$

for $0 \le q < 1$. Typical examples are diagonal matrices and AR(1) matrices. Note that if $q = 0$, the second condition means the exact sparsity constraint. The convergence rate is proved to be uniform on $\mathcal{U}_\tau(q, c_0(p), M)$ with

$$\left\| \widehat{\mathbf{\Sigma}}_\lambda^{\mathrm{GT}} - \mathbf{\Sigma} \right\|_2 = O_P\left(c_0(p) \left(\frac{\log p}{n} \right)^{(1-q)/2} \right)$$

provided that $\lambda = M'\sqrt{(\log p/n)} = o(1)$ for a sufficiently large M' along with some distributional assumptions including sub-Gaussian tail. The rate can be interpreted as the square root of $\log p/n$ multipled by true sparsity. Cai and Zhou (2012) constructs similar classes and proves that this rate cannot be improved.

One short remark is that in some of the variants of generalized thresholding methods, one discards only the off-diagonal elements of sample covariance matrix. Even in this case, asymptotic properties including consistency and convergence rate are preserved.

Adaptive thresholding estimator One invokes that the thresholding rules in the previous paragraph are *universal* in a sense that a single level of cutoff value is applied to all the entries of sample covariance matrix. Basically universal thresholding rules consider each element having homoscedastic noises. Cai et al. (2011) points out and illustrates, however, that sample covariance matrix may have elements with heteroscedastic noises from its validity nature. Instead, the authors propose an *adaptive* thresholding rule generalizing universal thresholding. Let $\mathbf{X} \equiv [x_{ki}]_{1\leq k\leq n, 1\leq i\leq p}$ be a data matrix with n observations and p variables. Then the adaptive thresholding estimators is described by

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Lambda}}^{\text{AT}} := s_{\boldsymbol{\Lambda}}(\mathbf{S}) := [s_{\lambda_{ij}}(s_{ij})]_{1\leq i,j\leq p}. \tag{2.3}$$

The threshold matrix $\boldsymbol{\Lambda} := [\lambda_{ij}]$ is suggested to be

$$\lambda_{ij} = \delta\sqrt{\frac{\eta_{ij}\log p}{n}},$$

where η_{ij} represents the variability of each s_{ij} and proposed to be estimated by

$$\widehat{\eta}_{ij} := \frac{1}{n}\sum_{k=1}^{n}\left[(x_{ki}-\bar{x}_{\cdot i})(x_{kj}-\bar{x}_{\cdot j})-s_{ij}\right]^2, \quad \bar{x}_{\cdot i} := \frac{1}{n}\sum_{k=1}^{n}x_{ki}.$$

$\delta \in [0,4]$ is selected by cross-validation, or letting $\delta = 2$ is also allowed in theory.

The convergence rate of adaptive thresholding estimator is inspected on broder class than that of universal thresholding estimator. Cai et al. (2014)

considers

$$\mathcal{H}(c_{n,p}) = \left\{ \Sigma : \max_{1\le i\le p} \sum_{j=1}^{p} \min\left((\sigma_{ii}\sigma_{jj})^{1/2}, \frac{|\sigma_{ij}|}{\sqrt{(\log p)/n}} \right) \le c_{n,p} \right\}.$$

The authors explain the meaning of $\mathcal{H}(c_{n,p})$ as follows.

> The comparison between the noise level $\sqrt{(\sigma_{ii}\sigma_{jj}\log p)/n}$ and the signal level $|\sigma_{ij}|$ captures the essence of the sparsity of the model.

$\mathcal{H}(c_{n,p})$ possess many classes of sparse covariance matrices proposed: in particular, the parameter spaces considered for universal thresholding estimator by Bickel and Levina (2008b); Rothman et al. (2009); Cai and Zhou (2012). It is also notable that $\mathcal{H}(c_{n,p})$ allows arbitrarily large σ_{ii}, while the parameter spaces for $\widehat{\mathbf{\Sigma}}^{\mathrm{GT}}$ does not. The convergence rate of $\widehat{\mathbf{\Sigma}}^{\mathrm{AT}}$ on $\mathcal{H}(c_{n,p})$ is expressed by

$$\mathbb{P}\left(\left\| \widehat{\mathbf{\Sigma}}^{\mathrm{AT}} - \mathbf{\Sigma} \right\|_2 \le C \cdot c_{n,p} \sqrt{\frac{\log p}{n}} \right) \ge 1 - O\left(p^{2-\delta} \sqrt{\frac{1}{\log p}} \right).$$

It is known that $\widehat{\mathbf{\Sigma}}^{\mathrm{AT}}$ is rate-optimal estimator in minimax sense in $\mathcal{H}(c_{n,p})$ (Cai et al., 2014) and $\widehat{\mathbf{\Sigma}}^{\mathrm{GT}}$ is sub-optimal. In short, since $\widehat{\mathbf{\Sigma}}^{\mathrm{AT}}$ is able to handling heterogeneous variability, its convergence rate can be applied to wider class than that of $\widehat{\mathbf{\Sigma}}^{\mathrm{GT}}$.

POET estimator In some practical applications, it is more reasonable to believe that data has common factors which is hidden. Static factor model is one easy way to describe the hidden structure, representing an observation by a linear combination of hidden (latent) variables. Suppose that each element of data matrix is generated by

$$x_{ij} = \mathbf{b}_j^\top \mathbf{f}_i + u_{ij}, \quad i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, p,$$

or, equivalently,

$$\mathbf{x}_i = \mathbf{B}^\top \mathbf{f}_i + \mathbf{u}_i, \quad i = 1, 2, \ldots, n.$$

Here, $\mathbf{x}_i := (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is i-th observation, $\mathbf{f}_i$ is a $K \times 1$ vector standing for common factors, $\mathbf{B} := [\mathbf{b}_1; \mathbf{b}_2; \cdots; \mathbf{b}_p]^\top$ is a $p \times K$ matrix of factor loading parameters, and $\mathbf{u}_i := (u_{i1}, u_{i2}, \ldots, u_{ip})^\top$ is a noise vector. We assume that $\mathbf{b}_i$'s and $\mathbf{u}_i$'s are respectively iid and that $\{\mathbf{b}_i\}$ and $\{\mathbf{u}_i\}$ are uncorrleated. K is also assumed to be fixed. Then the true covariance matrix has the following structure :

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Sigma}_{\text{factor}} + \mathbf{\Sigma}_{\text{noise}} \\ &:= \mathbf{B}\text{Var}(\mathbf{f}_i)\mathbf{B}^\top + \text{Var}(\mathbf{u}_i). \end{aligned} \tag{2.4}$$

Here, $\mathbf{\Sigma}_{\text{noise}}$ is called *idiosyncratic component* and its elements means the conditional (co-)variances given the common factors. It is desriable to assume that $\mathbf{\Sigma}_{\text{noise}}$ is sparse in a belief that the noises would be conditionally weakly correlated. Fan et al. (2013) asserts that if the top K eigenvalues of $\mathbf{\Sigma}$ are large relatively to the others, then the two terms, $\mathbf{\Sigma}_{\text{factor}}$ and $\mathbf{\Sigma}_{\text{noise}}$ in (2.4), could be distinguished by the spectral decomposition and estimated separately. Based on this intuition, Fan et al. (2013) proposes to retain the first K principal component of sample covariance matrix and threshold the residual (Principal Orthogonal complEment Thresholding, POET). Consider the spectral decompositon of sample covariance matrix

$$\mathbf{S} \equiv \mathbf{Q}\mathbf{L}\mathbf{Q}^\top \equiv \sum_{j=1}^{p} l_j \mathbf{q}_j \mathbf{q}_j^\top, \quad \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}, \quad \mathbf{L} \equiv \operatorname{diag}(l_1, l_2, \ldots, l_p)$$

with natural ordering $l_1 \geq l_2 \geq \ldots \geq l_p$. Then POET estimator is defined as

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{K,\mathbf{\Lambda}}^{\text{POET}} &:= \widehat{\mathbf{\Sigma}}_{\text{factor}}(K) + \widehat{\mathbf{\Sigma}}_{\text{noise}}(\mathbf{\Lambda}, K) \\ &= \sum_{i=1}^{K} l_i \mathbf{q}_i \mathbf{q}_i^\top + s_{\mathbf{\Lambda}} \left(\sum_{i=K+1}^{p} l_i \mathbf{q}_i \mathbf{q}_i^\top \right). \end{aligned} \tag{2.5}$$

where $s_{\mathbf{\Lambda}}(\cdot)$ is the adaptive threhsolding rule defined in (2.3). The former term estimates the factor covariance part while the latter term does the idiosyncratic part of $\mathbf{\Sigma}$. The number of true factor is possibly unknown; the detail of consistent estimation of K is referred to the paper. Convergence rate is spearately analyzed for $\widehat{\mathbf{\Sigma}}_{\text{noise}}$ and $\widehat{\mathbf{\Sigma}}_{\text{factor}}$; the key assumption is that the K top eigenvalues of $\mathbf{\Sigma}_{\text{factor}}$ diverge with rate $O(p)$ as $p \to \infty$. Then under some additional assumptions on distributions, regularities and thresholds,

$$\begin{aligned} \left\| \widehat{\mathbf{\Sigma}}_{\text{noise}}(\mathbf{\Lambda}, \widehat{K}) - \mathbf{\Sigma}_{\text{noise}} \right\|_2 &= O_p \left\{ m_p \left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}} \right)^{1-q} \right\}, \\ \frac{1}{\sqrt{p}} \left\| \widehat{\mathbf{\Sigma}}^{-\frac{1}{2}} \widehat{\mathbf{\Sigma}}_{\widehat{K},\mathbf{\Lambda}}^{\text{POET}} \widehat{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{I} \right\|_F &= O_p \left\{ \frac{\sqrt{p}\log p}{n} + m_p \left(\sqrt{\frac{1}{p}} + \sqrt{\frac{\log p}{n}} \right)^{1-q} \right\} \end{aligned}$$

where m_p indicates the sparsity of true $\mathbf{\Sigma}_{\text{noise}}$ for $0 \leq q < 1$. The first result of $\widehat{\mathbf{\Sigma}}_{\text{noise}}$ coincides the convergence rate of thresholding estimators if without $p^{-1/2}$. This additional rate $p^{-1/2}$ is explained to be the price to estimate the unknown K. On the other hand, the overall convergence rate of POET-estimator is given by the relative error, not by the normed difference from the true. Fan et al. (2013) explains the reason, that one can hardly obtain a satisfactory accuracy from any estimator when the first K eigenvalues of $\mathbf{\Sigma}$ are diverging, in which case the high dimensional factor model has essential limitations no matter what the estimatior is.

2.2. Regularized precision matrix estimators

In the literature for large-scale covariance matrix estimation, it is common to estimate $\mathbf{\Omega} := \mathbf{\Sigma}$ directly rather than to invert the estimator of $\mathbf{\Sigma}$. The reasons are clear: via direct formulation, not only one can acheive a desired structure (e.g. sparsity), but also can calculate estimator fast comparing to the combination of covariance matrix estimation and matrix inversion. In this section, the most popular schemes for sparse precision matrix estimation are introduced.

2.2.1. Penalized likelihood estimators

Penalized likelihood method indicates an estimation procedure that minimizing the log-likelihood of multivariate normal distribution with l_1 penalty in terms of $\mathbf{\Omega}$. Suppose that the data matrix $\mathbf{X}$ was generated from $\mathcal{N}\left(\boldsymbol{\mu}, \mathbf{\Omega}^{-1}\right)$. With plugging-in the sample mean vector (the MLE of $\boldsymbol{\mu}$), the log-likelihood of $\mathbf{\Sigma}$ comes with (up to constant)

$$\mathcal{L}\left(\mathbf{\Omega}\right) := \log\det\mathbf{\Omega} - \operatorname{tr}\left(\mathbf{S}\mathbf{\Omega}\right).$$

Yuan and Lin (2007) first proposed to minimize negative log-likelihood with respect to $\mathbf{\Omega}$, with elementwise l_1-norm constraint :

$$\mathcal{Q}_{\text{PL}}\left(\mathbf{\Omega}\right) := -\log\det\mathbf{\Omega} + \operatorname{tr}\left(\mathbf{S}\mathbf{\Omega}\right) + \lambda|\mathbf{\Omega}|_1, \tag{2.6}$$

of which the minimizer, say $\mathbf{\Omega}^{\text{PL}}$, is the estimator for true $\mathbf{\Omega}$. The term ‘PL’ abbreviates ‘penalized likelihood’. There are a huge number of algorithms proposed in the literature to solve the PL problem and its variants, for example, Banerjee et al. (2008); Danaher et al. (2014); D’Aspremont et al. (2008); Friedman et al. (2008); Hsieh (2014); Mazumder and Hastie (2012a,?); Ravikumar et al. (2011); Yuan and Lin (2007); Yuan (2008); Witten et al. (2011) to name a few. One short remark is that some PL methods use $|\mathbf{\Omega}|_{1,\text{off}}$ instead of $|\mathbf{\Omega}|_1$ to regularize the off-diagonal terms only, like the sample covariance matrix thresholding literature. The most popular algorithm solving PL problem is *graphical lasso* (Friedman et al., 2008), which is to be discussed below.

Asymptotic properties of penalized likelihood estimation has been inspected in Rothman et al. (2008); Lam and Fan (2009); Ravikumar et al. (2011) under various regularity conditions. They assume in common: data is generated from Gaussian or sub-Gaussian distribution; t the number of nonzero elements of true precision matrix is bounded by s; the specrum of true precision matrix is bounded below from zero and above from infinity. Then the convergence rate of penalized likelihood estimator is expressed by

$$\left\|\mathbf{\Omega}^{\text{PL}} - \mathbf{\Omega}\right\|^2 = O_P\left((1+s)\frac{\log p}{n}\right)$$

in both spectral norm and Frobenius norm.

Graphical lasso algorithm for penalized likelihood estimator This algorithm is now available as the `R` package `glasso`. It involves block coordinate descent algorithm as follows. Since (2.6) is convex in $\mathbf{\Theta}$, it suffices to solve the normal equation coming from the subgradient :

$$-\mathbf{U}+\mathbf{S}+\lambda\mathbf{\Gamma}=\mathbf{O}, \tag{2.7}$$

uhere $\mathbf{U} := \mathbf{\Theta}^{-1}$ and $\mathbf{\Gamma} \in \text{sign}(\mathbf{\Theta})$, uhich means $\gamma_{ij} = \text{sign}(\omega_{ij})$ if $\omega_{ij} \neq 0$; otheruise $\gamma_{ij} \in [-1, 1]$. A keypoint of the algorithm is to solve the equation in column-by-column fashion iteratively, with respect to $\mathbf{U}$. Without loss of generality, we demonstrate the update of the last column (row). Let the matrices be divided by blocks with sizes $p-1$ and 1 :

$$\mathbf{\Theta}=\begin{bmatrix}\mathbf{\Theta}_{11} & \boldsymbol{\omega}_{12}\\ \boldsymbol{\omega}_{12}^\top & \omega_{22}\end{bmatrix},\quad \mathbf{U}=\begin{bmatrix}\mathbf{U}_{11} & \mathbf{u}_{12}\\ \mathbf{u}_{12}^\top & u_{22}\end{bmatrix},\quad \mathbf{S}=\begin{bmatrix}\mathbf{S}_{11} & \mathbf{s}_{12}\\ \mathbf{s}_{12}^\top & s_{22}\end{bmatrix},\quad \mathbf{\Gamma}=\begin{bmatrix}\mathbf{\Gamma}_{11} & \boldsymbol{\gamma}_{12}\\ \boldsymbol{\gamma}_{12}^\top & \gamma_{22}\end{bmatrix}. \tag{2.8}$$

Now some observations are presented. From $\mathbf{U}\mathbf{\Theta} = \mathbf{I}$,

$$\mathbf{U}_{11}\boldsymbol{\omega}_{12}+\mathbf{u}_{12}\omega_{22} = \mathbf{0}\ ; \tag{2.9}$$

$$\mathbf{u}_{12}^\top\boldsymbol{\omega}_{12}+u_{22}\omega_{22} = 1. \tag{2.10}$$

On the other hand, one can observe the following, from (2.7),

$$u_{ii} = s_{ii}+\lambda, \quad i = 1, 2, \ldots, p\ ; \tag{2.11}$$

$$-\mathbf{u}_{12}+\mathbf{s}_{12}+\lambda\boldsymbol{\gamma}_{12} = 0. \tag{2.12}$$

The four equations above lies under the core of the algorithm. First, (2.11) shall be maintained during all iterations. Second, it is true from (2.9) that $\mathbf{u}_{12} = -\mathbf{U}_{11}\boldsymbol{\omega}_{12}/\omega_{22}$ and $\text{sign}(\boldsymbol{\omega}_{12}) = -\text{sign}(\mathbf{U}_{11}^{-1}\mathbf{u}_{12})$ since $\omega_{22} > 0$. Now letting $\boldsymbol{\beta} := \boldsymbol{\omega}_{12}/\omega_{22} \equiv -\mathbf{U}_{11}^{-1}\mathbf{u}_{12}$, one can rewirte (2.12) as

$$\mathbf{U}_{11}\boldsymbol{\beta} + \mathbf{s}_{12} + \lambda\boldsymbol{\nu} = \mathbf{0}, \tag{2.13}$$

where $\boldsymbol{\nu} \in \text{sign}(\boldsymbol{\beta})$. Now assume that $\mathbf{U}_{11} \succ 0$ is given fixed. Then solving (2.13) is equivalent to the following lasso-type optimization program :

$$\begin{aligned} &\text{minimize } \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{U}_{11}\boldsymbol{\beta} + \mathbf{s}_{12}^\top\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \\ \Longleftrightarrow\ &\text{minimize } \frac{1}{2}\|\mathbf{U}_{11}^{1/2}\boldsymbol{\beta} - \mathbf{b}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &\text{w.r.t.} \quad \boldsymbol{\beta} \in \mathbb{R}^{p-1} \end{aligned} \tag{2.14}$$

where $\mathbf{b} := -\mathbf{U}_{11}^{-1/2}\mathbf{s}_{12}$. Note that one can solve (2.14) efficiently by elementwise coordinate descent algorithm (Friedman et al., 2007), resulting in a sparse solution $\widehat{\boldsymbol{\beta}}$. Then $\widehat{\mathbf{u}}_{12}$ is calculated as $\widehat{\mathbf{u}}_{12} = -\mathbf{U}_{11}\widehat{\boldsymbol{\beta}}$ from (2.9). Once the last column is updated, The algorithm updates every column sequentially and repeatedly until convergence to obtain the solution $\widehat{\mathbf{U}}$. Finally, recovery $\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{U}}^{-1}$ from $\widehat{\mathbf{U}}$. This inverse computation is relatively cheap if all $\boldsymbol{\beta}$'s have been stored. To see why, solve (2.9) and (2.10) with respect to $\boldsymbol{\omega}_{12}$ and ω_{22} to write

$$\begin{aligned} \boldsymbol{\omega}_{12} &= -\mathbf{U}_{11}^{-1}\mathbf{u}_{12}\omega_{22} \ ; \\ \omega_{22} &= 1/\left(u_{22} - \mathbf{u}_{12}^\top\mathbf{U}_{11}^{-1}\mathbf{u}_{12}\right). \end{aligned}$$

From $\widehat{\boldsymbol{\beta}} = -\mathbf{U}_{11}^{-1}\widehat{\mathbf{u}}_{12}$, $\widehat{\omega}_{22} = 1/(u_{22} - \widehat{\mathbf{u}}_{12}^{\top}\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\omega}}_{12} = \widehat{\boldsymbol{\beta}}\widehat{\omega}_{22}$. One remarks that sparsity of $\widehat{\omega}_{12}$ is from $\widehat{\boldsymbol{\beta}}$.

2.2.2. Penalized regression-based estimators

Regression-based methods share the same motivation that each element of precision matrix is equivalent to the regression coefficient when a variable is regressed by the other variables. To express this quantitively, say $\mathbf{X} = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$. Then it is known that

$$X_j | X_{-j} \sim \mathcal{N}\left(-\omega_{jj}^{-1}\boldsymbol{\Omega}_{j,-j}^{\top}X_{-j},\, -\omega_{jj}^{-1}\right).$$

where the subscription '$-j$' means that 'all but j-th coordinate'. For example, $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$, and $\boldsymbol{\Omega}_{j,-j}$ denotes the j-th row of $\boldsymbol{\Omega}$ with j-th column removed.

The least square method comes of use for estimating the coefficients (the elements of precision matrix). The celebrating work was done by Meinshausen and Bühlmann (2006), which proposes to estimate $\boldsymbol{\Omega}$ by the solving the following lasso problem

$$\operatorname*{argmin}_{\beta} \|\mathbf{X}_j - \mathbf{X}_{-j}\beta\|_2^2 + \lambda\|\beta\|_1.$$

Note that this estimation does not guarantee the symmetry of estimates, which is an essential property of precision matrix. Nevertheless, a penalized regression approach became a important inspiration for precision matrix estimation since Meinshausen and Bühlmann (2006).

Pseudo-likelihood estimators In some of the penalized regression methods, the derived objective functions turns out to be rewritten in the following penalized pseudo-likelihood form

$$Q_{\mathrm{PPL}}(\mathbf{\Omega}) := -\log\det G(\mathbf{\Omega}) + \operatorname{tr}(\mathbf{S}H(\mathbf{\Omega})) + \lambda|\mathbf{\Omega}|_1. \tag{2.15}$$

The term 'pseudo-likelihood' is motivated by the fact that (2.15) can be seen as a generalization of Gaussian likelihood. Indeed, if $G(\mathbf{\Omega}) = H(\mathbf{\Omega}) = \mathbf{\Omega}$, (2.15) coincides the Gaussian likelihood. The choice of $G(\mathbf{\Omega})$ and $H(\mathbf{\Omega})$ varys through the regression-based methods, for example (up to reparametrization):

- $G(\mathbf{\Omega}) = \mathbf{\Omega}_D$ and $H(\mathbf{\Omega}) = \mathbf{\Omega}\mathbf{\Omega}_D^{-2}\mathbf{\Omega}$ ('SPACE', Peng et al., 2009) ;
- $G(\mathbf{\Omega}) = \mathbf{\Omega}_D$ and $H(\mathbf{\Omega}) = \mathbf{\Omega}\mathbf{\Omega}_D^{-1}\mathbf{\Omega}$ ('Symmetric lasso', Friedman et al., 2010) ;
- $G(\mathbf{\Omega}) = \mathbf{\Omega}_D^2$ and $H(\mathbf{\Omega}) = \mathbf{\Omega}$ ('CONCORD', Khare et al., 2015) ;

where $\mathbf{\Omega}_D := \operatorname{diag}(\mathbf{\Omega})$. Asymptotic properties are established for some of regression-based methods, SPACE and CONCORD for example.

CONCORD estimator CONCORD (CONvex partial CORrelation selection methoD) (Khare et al., 2015) is the most recent penalized pseudo-likelihood method for sparse precision matrix estimation. It is known to guarantee the three important properties simultaneously whereas other regression-based methods do not: symmetry, computational convergence, consistency.

The CONCORD estimation procedures are as follows. The objective function can be written in a coordinate-wise version by

$$Q_{\text{CON}}(\mathbf{\Omega}) := -\sum_{i=1}^{p} \log \omega_{ii} + \frac{1}{2n} \sum_{i=1}^{p} \left\| \omega_{ii} \mathbf{X}^i + \sum_{j \neq i} \omega_{ij} \mathbf{X}^j \right\|_2^2 + \lambda \sum_{1 \leq i < j \leq p} |\omega_{ij}|.$$

where $\mathbf{X}^i$ denote the i-th column of the $n \times p$ data matrix $\mathbf{X}$. Khare et al. (2015) proposes a coordinatewise minimization algorithm for minimizing $Q_{\text{CON}}(\mathbf{\Omega})$. This algorithm minimizes $Q_{\text{CON}}(\mathbf{\Omega})$ as a function of one coordinate at a time while fixing the others, and alternates this minization sequentially. Given the sample covariance matrix $\mathbf{S} = [s_{ij}]$ and having current iterate $\mathbf{\Omega} = [\omega_{ij}]$, the updated coordinate can be written explicitly by

$$\begin{aligned} \omega_{ii}^{\text{new}} &= \frac{-\sum_{j \neq i} w_{ij} s_{ij} + \sqrt{\left(\sum_{j \neq i} w_{ij} s_{ij}\right)^2 + 4 s_{ii}}}{2 s_{ii}}, \qquad \text{for } 1 \leq i \leq p, \\ \omega_{ij}^{\text{new}} &= \frac{s_\lambda \left\{ -\left(\sum_{j' \neq j} w_{ij'} s_{jj'} + \sum_{i' \neq i} w_{ji'} s_{ii'} \right) + 4 s_{ii} \right\}}{s_{ii} + s_{jj}} \quad \text{for } 1 \leq i < j \leq p, \end{aligned}$$

where s_λ denotes the soft thresholding operator $s_\lambda(t) := \text{sign}(t)(|t| - \lambda)_+$. The authors proved the computational convergence of this algorithm to a global minimum even if $n < p$.

The consistency of CONCORD is established under sub-Gaussinity, bounded eigenvalues and ‘incoherence condition’. Since description of the consistency for CONCORD is somewhat complicated, the detail is referred to the original paper.

2.2.3. Other methods

CLIME estimator CLIME (Constrainted l_1-minimization for Inverse Matrix Estimation) in Cai et al. (2011) directly approximates $\|\mathbf{\Omega\Sigma} - \mathbf{I}\|_\infty$ to the zero with sparsity constraint. The succeeding paper Cai et al. (2012) reflect the heteroscedastic variablility among the elements of estimated precision matrix and discuss the data-adaptive choice of tuning parameter. they propose to solve

$$\operatorname{argmin}\left\{\|\mathbf{\Omega}\|_1 : \|\mathbf{S\Omega} - \mathbf{I}\|_\infty \leq \tau\right\},$$

where $\tau = C\|\mathbf{\Omega}\|_1\sqrt{\log p/n}$. This formulation is equivalent to independent p of linear programming problem which is implemented well in standard packages. As for asymptotic properties, the parameter space in consideration is

$$\mathcal{HP}(c_{n,p}, M) = \left\{\begin{array}{c} \mathbf{\Omega} : \max_{1\leq i\leq p}\sum_{j\neq i}\min\left(1, \frac{|\omega_{ij}|}{\sqrt{(\log p)/n}}\right) \leq c_{n,p}, \\ M^{-1} \leq \gamma_{\min}(\mathbf{\Omega}),\, \max_i \omega_{ii} \leq M,\, \mathbf{\Omega} \succ 0 \end{array}\right\}.$$

Note that even if $\mathcal{HP}(c_{n,p}, M)$ might appears similar to $\mathcal{H}(c_{n,p})$ in adaptive thresholding estimator for covariance matrix estimation, $\mathcal{HP}(c_{n,p}, M)$ has stronger assumptions of bounded marginal variance and bounded eigenvalues. The result for convergence rate is as follows. If $c_{n,p}(\log p/n)^{1/2} = o(1)$, then for all $\mathbf{\Omega} \in \mathcal{HP}(c_{n,p}, M)$,

$$\left\|\widehat{\mathbf{\Omega}}^{\text{CLIME}} - \mathbf{\Omega}\right\|_2^2 = O_P\left(c_{n,p}^2\frac{\log p}{n}\right).$$

This rate has been proved to be minimax optimal (Cai et al., 2014).

2.3. Discussion: positive definiteness problem

Through previous sections, we see that the regularized estimation of covariance matrix (precision matrix, resp.) estimation methods involves thresholding, penalization, or constraint on the elements of covariance matrix (precision matrix, resp.). These regularizations are key techniques to establish the sparsity and consistency of the provided estimatiors when the true parameter matrices are sparse.

However, the sacrifice is the PDness of the estimators: aforementioned regularization techniques are not designed to reflect the spectral structure of matrices. All the thresholding, penalization and constraint are acting on matrices in elementwise fashion. Since the spectral structure of a matrix is complicated to analyze, none is known for how the element regularization techniques affects eigenvalues. Thus it is not surprising that they could give an estimate which is not PD from a finite sample.

As an illustration, we numerically inspected the smallest eigenvalues of regularized covariance matrix estimators. A dataset of $n = 100$ and $p = 400$ was generated from the multivariate t-distribution of degrees of freedom 5, with true covariance matrix $\mathbf{M}_1$ defined in Chapter 4. Then we calculated the universal thresholding estimator ($\widehat{\mathbf{\Sigma}}_\lambda^{\text{Thr}}$) with hard/soft/SCAD thresholding, and the banding estimator ($\widehat{\mathbf{\Sigma}}_\lambda^{\text{Band}}$) with banding/tapering techniques. Figure 2.1 plots the minimum eigenvalues of these estimators while tuning parameters are variously chosen. As shown in the graph, $\widehat{\mathbf{\Sigma}}_\lambda^{\text{Thr}}$ is not PD for moderately small thresholding level and even $\widehat{\mathbf{\Sigma}}_\lambda^{\text{band}}$ is not PD for most of the choice of bandwidth. One may question that we could exclude the non-PD estimates during the tuning parameter selection. However, a theory-supported five-fold cross validation introduced in Bickel and Levina (2008a) selected non-PD estimates. In short, covariance regularization techniques resulted in non-PDness of its estimates.

We further remark that even the penalized likelihood estimators can be non-PD in finite sample. Although the term log det $\mathbf{\Omega}$ in its objective function automatically guratantees the PDness of the global optimum, an algorithm to solve the PL problem could fail to reach the optimum and offer a non-PD estimate. A typical example is graphical lasso algorithm (Mazumder and Hastie, 2012).

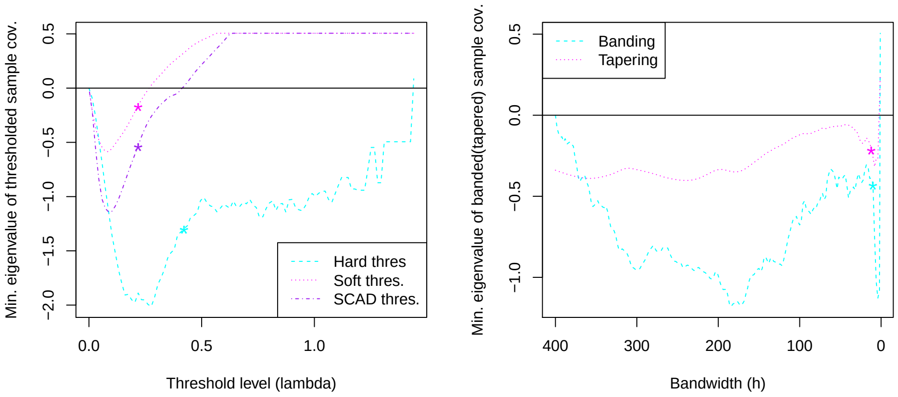

Figure 2.1: Miminum eigenvalues of regularized covariance matrix estimators ($n = 100$, $p = 400$). Left for thresholding estimators with threshold (λ) and right for banding estimators with varying bandwidth (h). The star ($*$)-marked point is the optimal threshold (bandwidth) selected by 5-fold cross validation. Note that all the CV-selected estimates are non-PD.

2.3.1. Related works

Although PDness is crucial for practical applications, a limited number of works are related to the PDness problem of sparse covariance matrix estimators (Rothman, 2012; Xue et al., 2012; Liu et al., 2014). The listed papers all start with the finding that the soft thresholding estimator can be obtained by minimizing the convex function:

$$\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}(\lambda) = (\text{sign}(s_{ij}) \cdot (|s_{ij}| - \lambda)_+) = \underset{\boldsymbol{\Sigma}}{\text{argmin}} \ \|\boldsymbol{\Sigma} - \mathbf{S}\|_{\text{F}}^2 + \lambda \sum_{i \leq j} |\sigma_{ij}|. \quad (2.16)$$

The listed papers add a constraint or penalty function to the objective to make the solution be PD. For example, Rothman (2012) considers a log-determinant penalty:

$$\widehat{\boldsymbol{\Sigma}}^{\text{logdet}}(\lambda) := \underset{\boldsymbol{\Sigma}}{\text{argmin}} \ \|\boldsymbol{\Sigma} - \mathbf{S}\|_{\text{F}}^2 + \tau \log\det(\boldsymbol{\Sigma}) + \lambda \sum_{i<j} |\sigma_{ij}|. \quad (2.17)$$

where $\tau > 0$ is fixed a small value. One short note is that (2.17) does not penalize the diagonal elements. The additional log-determinant behaves a convex barrier that naturally ensures **the** PDness of the solution and preserves the convexity the objective function. He solve the normal equations with respect to Σ, where a column of the current estimate Σ is updated jointly by solving a p-variate lasso regression. We note that, even the lasso regression can be fastly solved, solving it for every column alternatively leads to $O(p^3)$ flops for the whole $\boldsymbol{\Sigma}$ to be updated.

On the other hand, Xue et al. (2012) proposes to solve

$$\widehat{\mathbf{\Sigma}}^{\text{EigCon}}(\lambda) := \operatorname*{argmin}_{\mathbf{\Sigma}\,:\,\mathbf{\Sigma}\succeq\epsilon\mathbf{I}} \|\mathbf{\Sigma}-\mathbf{S}\|_{\text{F}}^2 + \lambda\sum_{i<j}|\sigma_{ij}|. \qquad (2.18)$$

where $\mathbf{\Sigma} \succeq \epsilon\mathbf{I}$ means that $\mathbf{\Sigma} - \epsilon\mathbf{I}$ is positive semidefinite. The modified objective function preserves convexity so that the global optimum exists. The paper optimizes (2.18) by an algorithm of alternating direction method of multipliers (ADMM) (Boyd et al., 2010), in which each iteration contains eigenvalue decomposition and thresholding of p by p matrix. Liu et al. (2014) use the same form of objective function as (2.16) and the ADMM algorithm with the sample correlation matrix instead of $\mathbf{S}$. As for computation cost, the ADMM algorithms used iterative solve eigen-decomposition problem with a complexity of $O(p^3)$.

To obtain deeper understanding, we tried a small simulation of the aforementioned PD-sparse covariance matrix estimators. Table 2.1 displays empirical risks and computation times of the proposed PD-sparse covariance estimators (by Rothman (2012) and Xue et al. (2012)) and soft thresholding estimator, over 100 replications from the same simulation setting with that of Figure 2.1. The PD-sparse covariance estimators show comparable performance to the soft-thresholding estimator, which is shown to be rate-optimal, in view of empirical risks. However, the computation time of PD sparse estimators is substantially increased than that of the soft thresholding. Combining with five-fold cross validation for tuning parameter selection with 100 candidates of λ, for example, it takes more than 2,500 sec to get either eigenvalue constraint or log-determinant barrier estimate, while the soft thresholding method only consumes 6 sec. It is no doubt that PDness is an essential property of covariance matrix to be plugged-in other statistical procedures, though, we question that it is worth heavy computation to guarantee only PDness additionally.

To summarize and conclude the literature review, the non-PDness problem is prevalent for many regularized covariance/precision matrix estimators and a systemic remedy to overcome it is yet to be established.

	Matrix l_1	Matrix l_2	Frobenius	#(PD)	Time (s)
Soft thres.	19.34 (1.12)	7.47 (0.30)	25.94 (0.40)	0/100	0.01
Eig. constraint	18.92 (1.05)	7.45 (0.30)	25.92 (0.39)	100/100	4.93
log-det barrier	18.90 (1.05)	7.45 (0.30)	25.99 (0.39)	100/100	2.33

Table 2.1: Comparison of average computing time (s.e.) for one fixed tuning parameter under 100 replications, measured on R version 3.1.2, Intel Core i7-2600 3.4GHz CPU and 16GB RAM. The data is generated under $n = 100$, $p = 400$, MV-t distribution, and $\mathbf{\Sigma} = \mathbf{M}_1$. Computational convergence criteria is set the relative error of 10^{-7}. ϵ and τ are respectvely 10^{-2} for eigenvalue constraint method and log-determinant barrer method.

Chapter 3

The linear shrinkage for positive definiteness (LSPD)

Let $\widehat{\mathbf{\Sigma}}$ be a possibly non-PD covariance matrix estimator. Recall that we propose a class of linear shrinkage as a correction of $\widehat{\mathbf{\Sigma}}$,

$$\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) := \alpha\widehat{\mathbf{\Sigma}} + (1-\alpha)\mu\mathbf{I},$$

where $\alpha \in [0,1]$ and $\mu \in \mathbb{R}$. The problem in consideration is the distance minimization between $\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}})$ and $\widehat{\mathbf{\Sigma}}$, while keeping the minimum eigenvalue of $\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}})$ positive. We first state this constrainted minimization problem quantitvely (Section 3.1.) and derive the minimum for α while fixing μ as a constant (Section 3.2.). Then we find the condition of μ which minimizes the distance for spectral norm and Frobenius norm (Section 3.3.). For chosen μ and σ, the statistical convergence rate of $\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}})$ is established (Section

3.4.). We also discuss some other implementation issues such as tuning parameter selection and computation (Section 3.5. and 3.6., respectively).

One notes that we need no shrinkage when $\widehat{\boldsymbol{\Sigma}}$ is already PD. The only case of interest is when $\widehat{\boldsymbol{\Sigma}}$ is not PD. From numerical reason, we set a small cut-point $\epsilon > 0$ to determine whether $\widehat{\boldsymbol{\Sigma}}$ is PD or not. In other words, we will let $\widehat{\boldsymbol{\Sigma}}$ as it is when $\widehat{\gamma}_{\min} \geq \epsilon$, and do some correction when $\widehat{\gamma}_{\min} < \epsilon$. We assume that ϵ is a fixed constant and specify its condition to guarantee the modified covariance estimator be statistically efficient.

3.1. Distance minimization

Let $\widehat{\gamma}_{\min} < \epsilon$. As in the introduction, we solve the distance minimization problem from $\widehat{\boldsymbol{\Sigma}}$ in which the correction is restricted to the family of linear shrinkage:

$$\begin{aligned} \underset{\mu,\alpha\in\mathbb{R}}{\text{minimize}} \quad & \left\| \boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}) - \widehat{\boldsymbol{\Sigma}} \right\| \\ \text{subject to} \quad & \alpha\widehat{\gamma}_{\min} + (1-\alpha)\mu \geq \epsilon\,; \\ & \alpha \in [0,1). \end{aligned} \tag{3.1}$$

In (3.1), the first constraint enforces the minimum eigenvalue of $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ to be at least ϵ. The second specifies the range of α (the magnitude of the shrinkage of eigenvalues to μ) where $\alpha = 0$ is for a complete shrinkage to μ and $\alpha = 1$ and $\alpha = 1$ is for no-shrinkage.

The range of μ, which is (ϵ, ∞), comes together with the first constraint and $\widehat{\gamma}_{\min} < \epsilon$. They imply that for any $\alpha \in [0, 1)$,

$$\mu \geq \frac{1}{1-\alpha}\epsilon - \frac{\alpha}{1-\alpha}\widehat{\gamma}_{\min} > \frac{1}{1-\alpha}\epsilon - \frac{\alpha}{1-\alpha}\epsilon = \epsilon.$$

This tells the shrinkage target μ should be larger than numerical bound ϵ.

3.2. The choice of α

We for a moment assume $\mu \in (\epsilon, \infty)$ is a fixed constant and solve (3.1). Since the correction is only applied to the case $\widehat{\gamma}_{\min} < \epsilon$, we have $\widehat{\gamma}_{\min} < \epsilon < \mu$ and

$$1 - \alpha \geq \frac{\epsilon - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min}} \tag{3.2}$$

from the constraints of (3.1). Note that $\|\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\| = (1-\alpha)\|\mu\mathbf{I} - \hat{\mathbf{\Sigma}}\|$ in the objective function (3.1). Since μ is fixed and $\widehat{\mathbf{\Sigma}}$ is given, only $(1-\alpha)$ can vary in $(1-\alpha)\|\mu\mathbf{I} - \hat{\mathbf{\Sigma}}\|$, and this can be minimized when $(1-\alpha)$ touches the lower bound of the inequality in (3.2). Thus we have the following lemma.

Lemma 3.1. Let $\widehat{\mathbf{\Sigma}}$ be given and assume $\epsilon > \widehat{\gamma}_{\min}$. Then for any $\mu \in (\epsilon, \infty)$, the problem (3.1), with respect to α, is minimized at

$$\alpha^* = \frac{\mu - \epsilon}{\mu - \widehat{\gamma}_{\min}}. \tag{3.3}$$

3.3. The choice of μ

Substituting (3.3) to the problem (3.1) yields a reduced problem which depends only on μ:

$$\underset{\mu\,:\,\mu>\epsilon}{\text{minimize}} \quad \left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\| \quad = \quad \frac{\epsilon - \widehat{\gamma}_{\text{min}}}{\mu - \widehat{\gamma}_{\text{min}}} \left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|. \tag{3.4}$$

The solution of this problem depends on how the matrix norm $\|\cdot\|$ is defined. We consider two most popular matrix norms in below, spectral norm $\|\cdot\|_2$ and Frobenius norm $\|\cdot\|_{\text{F}}$.

Lemma 3.2 (Spectral norm)**.** If $\widehat{\mathbf{\Sigma}}$ is given and $\epsilon > \widehat{\gamma}_{\text{min}}$, then

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_2 \geq \epsilon - \widehat{\gamma}_{\text{min}} \quad \text{for all } \mu > \epsilon.$$

The equality condition is specified by

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_2 = \epsilon - \widehat{\gamma}_{\text{min}} \iff \mu \geq \mu_{\text{S}} := \frac{\widehat{\gamma}_{\text{max}} + \widehat{\gamma}_{\text{min}}}{2}.$$

Proof. By the definition of spectral norm,

$$\begin{aligned}\frac{\epsilon - \widehat{\gamma}_{\text{min}}}{\mu - \widehat{\gamma}_{\text{min}}} \left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|_2 &= \frac{\epsilon - \widehat{\gamma}_{\text{min}}}{\mu - \widehat{\gamma}_{\text{min}}} \max_i |\mu - \widehat{\gamma}_i| \\ &= \frac{\epsilon - \widehat{\gamma}_{\text{min}}}{\mu - \widehat{\gamma}_{\text{min}}} \max\left\{|\mu - \widehat{\gamma}_{\text{max}}|, |\mu - \widehat{\gamma}_{\text{min}}|\right\}.\end{aligned}$$

Consider $\psi_2(t) := \max\left\{|t-a|, |t-b|\right\}/(t-a)$ with $a < b$, $t \in (a, \infty)$. One can easily check that $\psi_2(t) \equiv 1$ when $\mu \geq (a+b)/2$, and $\psi_2(t) > 1$ when $a < t < (a+b)/2$. Now substitute $t \leftarrow \mu$, $a \leftarrow \widehat{\gamma}_{\text{min}}$, and $b \leftarrow \widehat{\gamma}_{\text{max}}$. □

Lemma 7 implies that whenever we take μ such that $\mu \geq \mu_{\mathrm{S}}$, the distance (in spectral norm) between the PD-modified and the original estimator is exactly equals to the difference between their smallest eigenvalues. We remark that $\mu \in (\epsilon, \mu_{\mathrm{S}})$ is not a recommended choice. Indeed, if $\mu \in (\epsilon, \mu_{\mathrm{S}})$, it is easy to see that $\|\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_2$, as a function of μ, increases to infinity as $\mu \downarrow \epsilon$.

Lemma 3.3 (Frobenius norm)**.** If $\widehat{\mathbf{\Sigma}}$ is given and $\epsilon > \widehat{\gamma}_{\min}$, then

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \geq (\epsilon - \widehat{\gamma}_{\min}) \cdot \frac{1}{\sqrt{1 + \{(\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min})^2 / \mathrm{V}(\widehat{\gamma})\}}} \quad \text{for all } \mu > \epsilon,$$

where $\mathrm{M}(\widehat{\gamma}) := \sum_i \widehat{\gamma}_i / p$ and $\mathrm{V}(\widehat{\gamma}) := (\sum_i \widehat{\gamma}_i^2 / p) - \mathrm{M}(\widehat{\gamma})^2$. The equality holds if and only if $\mu = \mu_{\mathrm{F}} := \mathrm{M}(\widehat{\gamma}) + \{\mathrm{V}(\widehat{\gamma}) / (\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min})\}$. In particular,

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} < \epsilon - \widehat{\gamma}_{\min} \quad \text{for all } \mu \geq \mathrm{M}(\widehat{\gamma}) + \{\mathrm{V}(\widehat{\gamma}) / (\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min})\}.$$

Proof. Observe that

$$\begin{aligned} \frac{\epsilon - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min}} \left\|\mu \mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} &= \frac{\epsilon - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min}} \cdot \frac{1}{\sqrt{p}} \sqrt{\mathrm{tr}[(\mu \mathbf{I})^\top (\mu \mathbf{I})] - 2\mathrm{tr}[(\mu \mathbf{I})^\top \widehat{\mathbf{\Sigma}}] + \mathrm{tr}[\widehat{\mathbf{\Sigma}}^\top \widehat{\mathbf{\Sigma}}]} \\ &= \frac{\epsilon - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min}} \sqrt{\mu^2 - \frac{2}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}})\mu + \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^\top \widehat{\mathbf{\Sigma}})}. \end{aligned}$$

To make the problem simpler, set $a := \widehat{\gamma}_{\min}$, $b := \mathrm{tr}(\widehat{\mathbf{\Sigma}})/p$, and $c := \mathrm{tr}(\widehat{\mathbf{\Sigma}}^\top \widehat{\mathbf{\Sigma}})$. It suffices to inspect the behavior of the funcion $\psi_{\mathrm{F}}(t) := (t^2 - 2bt + c)/(t - a)^2$ on $t \in (a, \infty)$, given constant $a < b < c$. One can easily observe that (1) ψ_{F} has unique minimum at $t^* = b + (c - b^2)/(b - a)$ with $\psi_{\mathrm{F}}(t^*) = \{1 + (b - a)^2/(c - b^2)\}^{-1}$, (2) ψ_{F} is strictly increasing on (t^*, ∞), and (3)

$\lim_{t\to\infty}\psi_{\mathrm{F}}(t) = 1$.

To finish the proof, recall that $b = \mathrm{tr}(\widehat{\mathbf{\Sigma}})/p = \sum_i \widehat{\gamma}_i/p \equiv \mathrm{M}(\widehat{\gamma})$ and $c - b^2$ $= \mathrm{tr}(\widehat{\mathbf{\Sigma}}^\top\widehat{\mathbf{\Sigma}})/p - [\mathrm{tr}(\widehat{\mathbf{\Sigma}})/p]^2 = (\sum_i \widehat{\gamma}_i^2/p) - (\sum_i \widehat{\gamma}_i/p)^2 \equiv \mathrm{V}(\widehat{\gamma})$. □

In the proof, we see that the Frobenius-norm distance function $\|\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_{\mathrm{F}}$ (as a function of μ) has the unique minimum at $\mu = \mu_{\mathrm{F}}$. The function is strictly increasing but bounded by $\epsilon - \widehat{\gamma}_{\min}$ on $[\mu_{\mathrm{F}}, \infty)$ and also increases to infinity as $\mu \downarrow \epsilon$. This suggests us to choose $\mu \geq \mu_{\mathrm{F}}$ to make the Frobenius-norm distance smaller than $\epsilon - \widehat{\gamma}_{\min}$.

Finally, as a summary, we have the following theorem.

Theorem 3.4 (Distance between the original and LSPD). *Assume $\widehat{\mathbf{\Sigma}}$ and $\epsilon > 0$ be given. Set*

$$\alpha^* = \begin{cases} 1 & \textit{if } \widehat{\gamma}_{\min} \geq \epsilon \\ 1 - \frac{\epsilon - \widehat{\gamma}_{\min}}{\mu^* - \widehat{\gamma}_{\min}} & \textit{if } \widehat{\gamma}_{\min} < \epsilon \end{cases}; \qquad \mu \in [\mu_{\mathrm{SF}}, \infty),$$

where $\mu_{\mathrm{SF}} := \max(\mu_{\mathrm{S}}, \mu_{\mathrm{F}})$ with μ_{S} and μ_{F} defined in Lemma 7 and 3.3, respectively. Then

1. *$\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_2$ exactly equals to $(\epsilon - \widehat{\gamma}_{\min})_+$ for any μ ;*

2. *$\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_{\mathrm{F}}$ is increasing in μ and bounded by*

$$\frac{(\epsilon - \widehat{\gamma}_{\min})_+}{\sqrt{1 + \{(\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min})^2/\mathrm{V}(\widehat{\gamma})\}}} \leq \left\|\mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \leq (\epsilon - \widehat{\gamma}_{\min})_+.$$

Remark. Hereafter, we call $\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}})$ $(\mu \in [\mu_{\mathrm{SF}}, \infty))$ by *LSPD-estimator induced by* $\widehat{\mathbf{\Sigma}}$, or *LSPD-modifier of* $\widehat{\mathbf{\Sigma}}$. And if there is no confusion, we abbreviate $\mathbf{\Phi}_{\mu,\alpha^*}\left(\hat{\mathbf{\Sigma}}\right)$ as $\mathbf{\Phi}_{\mu}\left(\hat{\mathbf{\Sigma}}\right)$ or just $\mathbf{\Phi}\left(\hat{\mathbf{\Sigma}}\right)$ for notational simplicity.

In the theorem, for any μ in $[\mu_{\mathrm{SF}}, \infty)$, the distance between the original estimator $\widehat{\mathbf{\Sigma}}$ and its LSPD-updated one is smaller than $(\epsilon - \widehat{\gamma}_{\min})_+$. Indeed, the Monte Carlo experiments in Section 4 show that the empirical risk of $\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}})$ is robust to the choice of μ. However, $\mu = \mu_{\mathrm{SF}}$ would be a more preferable choice if one wants to minimize the Frobenius-norm distance from $\mathbf{\Sigma}$ while keeping the spectral-norm distance as minimal. In addition, the few smallest or largest eigenvalues of $\widehat{\mathbf{\Sigma}}$ are likely to be illusionary outcomes from the high-dimensionality of the data, and it would be reasonable to shrink them to the center by choosing μ not too high.

We remark that a special case of the proposed LSPD estimator is discussed by a group of researchers (Section 5.2 in Cai et al. (2014)). In the paper, the authors propose to modify the original estimator $\widehat{\mathbf{\Sigma}}$ as $\widehat{\mathbf{\Sigma}}+(\epsilon-\widehat{\gamma}_{\min})\mathbf{I}$. This coincides our LSPD estimator with $\mu = \infty$ (understand this as taking $\mu \to \infty$); in this case, $\mathbf{\Phi}_{\infty,\alpha^*}(\widehat{\mathbf{\Sigma}}) = \widehat{\mathbf{\Sigma}} + (\epsilon - \widehat{\gamma}_{\min})\mathbf{I}$.

3.4. Statistical properties of LSPD-estimator

The convergence rate of LSPD-estimator to the true Σ is based on the triangle inequality

$$\begin{aligned} \left\|\mathbf{\Phi}\left(\hat{\mathbf{\Sigma}}\right) - \mathbf{\Sigma}\right\| &\leq \left\|\mathbf{\Phi}\left(\hat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\| + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \\ &\leq (\epsilon - \widehat{\gamma}_{\min}) + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|. \end{aligned}$$

As in Xue et al. (2012), we first set ϵ be a constant smaller than $\gamma_{\min}$, the smallest eigenvalue of the true $\mathbf{\Sigma}$, which is assumed to be positive but can approaches 0 as the dimension p increases. For this choice of ϵ, we will show that

$$(\epsilon - \widehat{\gamma}_{\min}) \leq \gamma_{\min} - \widehat{\gamma}_{\min} \leq \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|, \tag{3.5}$$

for both spectral and Frobenius norms. In (3.5), the first inequality is simply from $\epsilon \leq \gamma_{\min}$. Now we show the second inequality for general symmetric matrices $\mathbf{A}$ and $\mathbf{B}$. For spectral norm, the perturbation inequality tells us that

$$\max_i \{\gamma_i(\mathbf{A}) - \gamma_i(\mathbf{B})\} \leq \|\mathbf{A} - \mathbf{B}\|_2 .$$

In case of Frobenius norm, note that

$$\|\mathbf{A}\|_{\mathrm{F}} = [\mathrm{tr}(\mathbf{A}^\top \mathbf{A})/p]^{1/2} = [\sum_i \gamma_i(\mathbf{A})/p]^{1/2}$$

and

$$\gamma_{\min}(\mathbf{A}) - \gamma_{\min}(\mathbf{B}) \leq \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} - \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{B}\mathbf{x} \leq \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top (\mathbf{A} - \mathbf{B})\mathbf{x} = \gamma_{\min}(\mathbf{A} - \mathbf{B}).$$

Finally, using the inequalities (3.5), we have the following theorem on the convergence rate of the LSPD estimator.

Theorem 3.5 (Convergence rate of LSPD-estimator)**.** *Let $\widehat{\mathbf{\Sigma}}$ is any estimator of unknown true covariance matrix $\mathbf{\Sigma}$. If ϵ is equal to or less than the smallest eigenvalue of $\mathbf{\Sigma}$, then for α^* defined in Theorem 3.4 and $\mu \in [\mu_{\mathrm{SF}}, \infty)$,*

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Sigma}}\right) - \mathbf{\Sigma} \right\| \leq 2 \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|,$$

for both spectral and Frobenius norm.

Theorem 3.5 shows that the convergence rate of the LSPD estimator is at least equivalent to that of the original estimator. In particular, LSPD-modifier inherits the minimax rate optimality (with respect to spectral norm or Frobenius norm) of its original estimator. Examples include the banding or tapering estimator (Cai et al., 2010), the adaptive block thresholding estimator (Cai and Yuan, 2012), the universal thresholding estimator (Cai and Zhou, 2012), and the adaptive thresholding estimator (Cai and Liu, 2011)

3.5. If tuning parameter selection is involved for $\widehat{\mathbf{\Sigma}}$

The selection of tuning parameter (for example, the thresholding value in Bickel and Levina (2008a); Cai and Liu (2011) and the size of the bandwidth in the banding estimator (Bickel and Levina, 2008b)) is a very central step in most of recent regularized covariance estimators. Let $\widehat{\mathbf{\Sigma}}(\lambda)$ be a regularized estimator indexed by tuning parameter λ. Usually the derivation of convergence rate of $\widehat{\mathbf{\Sigma}}(\lambda)$ assumes $\lambda \equiv \lambda_n \to 0$ in certain order as $n \to \infty$, without giving any guidance to choose λ. Practically the selection of λ often follows the procedure: (a) iteratively split the data into training and testing sets, (b) compute the estimate from the training data set (say $\widehat{\mathbf{\Sigma}}_{\mathrm{Train}}(\lambda)$) and evaluate the risk by $\|\widehat{\mathbf{\Sigma}}_{\mathrm{Train}}(\lambda) - \mathbf{S}_{\mathrm{Test}}\|_{\mathrm{F}}^2$, where $\mathbf{S}_{\mathrm{Test}}$ denotes the sample covariance matrix from the testing-set, and (c) do (a) and (b) for each λ and select λ^* to minimize the evaluated risk. Theorem 3 of Bickel and Levina (2008a) proves that $\widehat{\mathbf{\Sigma}}(\lambda^*)$ with λ^* chosen by step (a)-(c) acheives the same convergence rate with $\widehat{\mathbf{\Sigma}}(\lambda_n)$.

This framework for λ-tuning is also applied to the PD-regularized estimatiors of Rothman (2012); Xue et al. (2012); Liu et al. (2014), which are introduced in Section **??**. They first establish convergence rates by assuming $\lambda_n \to 0$, and follow step (a)-(c) for data-driven choice of λ to utilize Theorem 3 of Bickel and Levina (2008a). However, the PD-regularized estimators are based on large-scale optimization, then it leads to a trenendous computation to compute of $\widehat{\mathbf{\Sigma}}^{n_1}(\lambda)$ repetitively for each λ and each training dataset. To reduce computational burden, one could consider *one-step update* estimate $\widehat{\mathbf{\Sigma}}^{\text{logdet}}(\lambda^*)$ or $\widehat{\mathbf{\Sigma}}^{\text{EigCon}}(\lambda^*)$, where λ^* was chosen from the soft-thresholding estimator. However, the asymptotic behavior of $\widehat{\mathbf{\Sigma}}^{\text{logdet}}(\lambda^*)$ (or $\widehat{\mathbf{\Sigma}}^{\text{EigCon}}(\lambda^*)$) is unknown.

Unlike those PD-regularized estimators, our LSPD-estimator supports *one-step update* framework: it suffices to calculate $\mathbf{\Phi}\big(\widehat{\mathbf{\Sigma}}(\lambda^*)\big)$ *once* after obtaining $\widehat{\mathbf{\Sigma}}(\lambda^*)$ where λ^* is selected by (a)-(c) with respect to $\widehat{\mathbf{\Sigma}}$. Our Theorem 2 tells that $\mathbf{\Phi}\big(\widehat{\mathbf{\Sigma}}(\lambda^*)\big)$ has a same convergence rate no slower than that of $\widehat{\mathbf{\Sigma}}(\lambda^*)$. Thus if $\widehat{\mathbf{\Sigma}}(\lambda^*)$ is asymptotically equivalent to $\widehat{\mathbf{\Sigma}}(\lambda_n)$, so is $\mathbf{\Phi}\big(\widehat{\mathbf{\Sigma}}(\lambda^*)\big)$ to $\widehat{\mathbf{\Sigma}}(\lambda_n)$. It makes the LSPD-correction is simply and fastly applied to any regularized covariance estimation with λ-tuning. For example, in the case of $n = 100$ and $p = 400$, the soft thresholding estimator with five-fold lcross-validation costs the computation of about 6 sec, and its one step LSPD-update takes 0.06 sec. In the simulation study later, all LSPD corrections will be based on one-step update, and it will be seen that the empirical errors after LSPD correction are still comparable to original regularized estimators or other proposals.

3.6. Computation

The proposed LSPD estimator by itself has great computational advantages over the existing estimators (Rothman, 2012; Xue et al., 2012; Liu et al., 2014). Most of all, it is optimization-free and does not need do eigenvalue decomposition, which costs $O(p^3)$ computation, iteratively. Intuitively, the LSPD estimator needs it only one time. However, in truth, it is much faster than the intuition. Recall that the LSPD estimator depends on the four functionals of $\widehat{\mathbf{\Sigma}}$, which are are $\widehat{\gamma}_{\max}$, $\widehat{\gamma}_{\min}$, $\mathrm{M}(\widehat{\gamma})$, and $\mathrm{V}(\widehat{\gamma})$. For $\widehat{\gamma}_{\max}$ and $\widehat{\gamma}_{\min}$, the largest and smallest eigenvalue of a symmetric matrix can be calculated independently from the whole spectrum by Krylov subspace methods (see Chapter 7 of Demmel, 1997 or Chapter 10 of Golub and Van Loan, 2012 for example), which is faster than the usual eigenvalue decomposition for large-scale sparse matrices. Some of these algorithms are implemented into the package; the MATLAB built-in function `eigs()` (based on Lehoucq and Sorensen, 1996; Sorensen, 1990) and the user-defined function `eigifp()` (Golub and Ye, 2002) available at the author's homepage.[1] For the other two functionals $\mathrm{M}(\widehat{\gamma})$ and $\mathrm{V}(\widehat{\gamma})$, they are written as

$$\begin{aligned}
\mathrm{M}(\widehat{\gamma}) &= \sum_i \widehat{\gamma}_i / p = \mathrm{tr}(\widehat{\mathbf{\Sigma}})/p \\
\mathrm{V}(\widehat{\gamma}) &= (\sum_i \widehat{\gamma}_i^2 / p) - \mathrm{M}(\widehat{\gamma})^2 = \mathrm{tr}(\widehat{\mathbf{\Sigma}}^2)/p - \{\mathrm{tr}(\widehat{\mathbf{\Sigma}})/p\}^2,
\end{aligned}$$

and can also be evaluated without the computation of the whole spectrum.

[1] During the simulation, we use `eigifp()` instead of `eigs()`, since `eigs()` failed to converge in some smallest eigenvalue computation for large matrices.

3.7. Inaccurate calculation of smallest eigenvalue

The extreme eigenvalues can be calculated inaccurately for large-scale matrix in practical. In LSPD estimation, an inaccurate approximcation of the smallest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$ could lead to two problems: (1) $\boldsymbol{\Phi}(\widehat{\boldsymbol{\Sigma}})$ could fail to be PD, and (2) $\boldsymbol{\Phi}(\widehat{\boldsymbol{\Sigma}})$ could fail to preserve the convergence rate of $\widehat{\boldsymbol{\Sigma}}$. Let $\widehat{\gamma}_{\min}$ be the accurately calculated smallest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$, and let $\widehat{\gamma}^1_{\min}$ be an inaccurate approximation of $\widehat{\gamma}_{\min}$. Since all the numerical algorithms to calculate smallest eigenvalue approximates the true with upward bias, we assume $\widehat{\gamma}^1_{\min} = \widehat{\gamma}_{\min} + \delta$ where $\delta > 0$. In this section, we specify the range of δ to keep PDness and convergence rate of $\boldsymbol{\Phi}(\widehat{\boldsymbol{\Sigma}})$ despite of inaccurate calculation of $\widehat{\gamma}_{\min}$. We assume that $\epsilon > \widehat{\gamma}^1_{\min}$.

3.7.1. PDness of $\boldsymbol{\Phi}(\widehat{\boldsymbol{\Sigma}})$

Recall that we chose α as $\alpha^* = \frac{\mu-\epsilon}{\mu-\widehat{\gamma}_{\min}}$. Then during the inaccurate approximation, α^* is incorrectly specified as $\alpha^{*1} := \frac{\mu-\epsilon}{\mu-\widehat{\gamma}^1_{\min}}$. The smallest eigenvalue of $\boldsymbol{\Phi}_{\mu,\alpha^{*1}}(\widehat{\boldsymbol{\Sigma}})$ satisfies

$$\begin{aligned}
\gamma_{\min}\left(\boldsymbol{\Phi}_{\mu,\alpha^{*1}}(\widehat{\boldsymbol{\Sigma}})\right) &= \frac{1}{\mu - \widehat{\gamma}^1_{\min}}\left((\mu-\epsilon)\widehat{\gamma}_{\min} + (\epsilon - \widehat{\gamma}^1_{\min})\mu\right) \\
&= \frac{\mu - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min} - \delta}\epsilon - \frac{\delta}{\mu - \widehat{\gamma}_{\min} - \delta}\mu \\
&> \epsilon - \frac{\delta}{\mu - \widehat{\gamma}_{\min} - \delta}\mu.
\end{aligned}$$

We note that if δ were zero, the second line coincides ϵ. We will guarantee the final term of the third line to be nonnegative. To do this, we should ensure that

$$\epsilon \geq \frac{\delta\mu}{\mu - \widehat{\gamma}_{\min} - \delta} \iff \delta \leq \epsilon \cdot \frac{\mu - \widehat{\gamma}_{\min}}{\mu + \epsilon}.$$

Since $\epsilon > \widehat{\gamma}^1_{\min} > \widehat{\gamma}_{\min}$, from the increasing property $\frac{\mu - \widehat{\gamma}_{\min}}{\mu+\epsilon}$ (in $\mu \in (\epsilon, \infty)$),

$$\delta \leq \epsilon \cdot \frac{\mu - \epsilon}{\mu + \epsilon} \implies \delta \leq \epsilon \cdot \frac{\mu - \widehat{\gamma}_{\min}}{\mu + \epsilon}.$$

Thus we can conclude that if the inaccuracy measure δ is sufficiently small than $\epsilon \cdot \frac{\mu-\epsilon}{\mu+\epsilon}$, $\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}})$ is not guaranteed to be ϵ.

3.7.2. Convergence rate

Since $\|\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\| = (1-\alpha)\|\mu\mathbf{I} - \hat{\mathbf{\Sigma}}\|$,

$$\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\| = \frac{\epsilon - \widehat{\gamma}^1_{\min}}{\mu - \widehat{\gamma}^1_{\min}}\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\|.$$

We claim the argument into spectral norm and Frobenius norm separately. For spectral norm, the result is summarized to the following lemma. We note that μ_{S} is also inaccurately calculated as $\mu \leq \mu^1_{\mathrm{S}} := (\widehat{\gamma}_{\max} + \widehat{\gamma}^1_{\min})/2$. Lemma A.1. implies that if we set the inaccuracy δ sufficiently small relative to $\widehat{\gamma}_{\max} - \widehat{\gamma}_{\min}$, the convergence rate of $\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}})$ can be preserved even if we calculate μ_{S} by $\mu \geq \mu^1_{\mathrm{S}}$.

Lemma 3.6 (Spectral norm)**.** Write $\delta = \eta(\widehat{\gamma}_{\max} - \widehat{\gamma}_{\min})$. If $\widehat{\gamma}^1_{\min} < \epsilon$,

$$\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_2 \leq (\epsilon - \widehat{\gamma}_{\min}) \cdot \frac{1+\eta}{1-\eta},$$

for all $\mu \geq \mu_{\mathrm{S}}^{1} := (\widehat{\gamma}_{\max} + \widehat{\gamma}_{\min}^{1})/2$.

Proof.

$$\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_2 = \frac{\epsilon - \widehat{\gamma}_{\min}^{1}}{\mu - \widehat{\gamma}_{\min}^{1}} \max\left\{|\mu - \widehat{\gamma}_{\max}|, |\mu - \widehat{\gamma}_{\min}|\right\}.$$

Let $\mu \geq \mu_{\mathrm{S}}^{1}$. Since $\mu_{\mathrm{S}}^{1} > \mu_{\mathrm{S}}$,

$$\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\| = \epsilon - \widehat{\gamma}_{\min}^{1} \cdot \frac{\mu - \widehat{\gamma}_{\min}}{\mu - \widehat{\gamma}_{\min}^{1}}.$$

The left factor of RHS is bounded by $\epsilon - \widehat{\gamma}_{\min}$ by the assumption $\widehat{\gamma}_{\min}^{1} < \epsilon$. It is also easy to show that the second factor is bounded by $\frac{1+\eta}{1-\eta}$ provided that $\mu \geq \mu_{\mathrm{S}}^{1}$. □

As for Frobenius norm, we also let $\mu_{\mathrm{F}}^{1} := \mathrm{M}(\widehat{\gamma}) + \{\mathrm{V}(\widehat{\gamma})/(\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min}^{1})\}$ to indicate the inaccurate approximate of μ_{F}. $\mathrm{M}(\widehat{\gamma}$ and $\mathrm{V}(\widehat{\gamma})$ is free of inaccurate approximation, since they can be computed from trace operators as in Section 3.6.. It is also easy to show $\mu_{\mathrm{F}}^{1} > \mu_{\mathrm{F}}$.

Lemma 3.7 (Frobenius norm)**.** If $\widehat{\gamma}_{\min}^{1} < \epsilon$ and $\widehat{\gamma}_{\min}^{1} < \mathrm{M}(\widehat{\gamma})$, then

$$\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_{\mathrm{F}} \leq \epsilon - \widehat{\gamma}_{\min},$$

for all $\mu \geq \mu_{\mathrm{F}}^{1} := \mathrm{M}(\widehat{\gamma}) + \{\mathrm{V}(\widehat{\gamma})/(\mathrm{M}(\widehat{\gamma}) - \widehat{\gamma}_{\min}^{1})\}$.

Proof. First, $\|\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_{\mathrm{F}}$ is rewritten as

$$\frac{\epsilon - \widehat{\gamma}_{\min}^{1}}{\mu - \widehat{\gamma}_{\min}^{1}} \left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} = (\epsilon - \widehat{\gamma}_{\min}^{1}) \cdot \frac{1}{\mu - \widehat{\gamma}_{\min}^{1}} \sqrt{\mu^2 - \frac{2}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}})\mu + \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^{\top}\widehat{\mathbf{\Sigma}})}.$$

The first factor of RHS is also bounded $\epsilon - \widehat{\gamma}_{\min}$ by the assumption. It suffices to show that the function

$$\psi_{\mathrm{F}}(t) = \frac{1}{\mu - \widehat{\gamma}_{\min}} \sqrt{\mu^2 - \frac{2}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}})\mu + \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^{\top}\widehat{\mathbf{\Sigma}})}$$

is lower than 1 for $\mu \geq \mu_{\mathrm{F}}^1$. It goes parallel with the proof of Lemma 3.3, since we assume $\widehat{\gamma}_{\min}^1 < \mathrm{M}(\widehat{\gamma})$. □

From the two lemmas, we can conclude that if (1) δ sufficiently small comparing to $\widehat{\gamma}_{\max} - \widehat{\gamma}_{\min}$, (2) $\widehat{\gamma}_{\min}^1 \leq \min\{\epsilon, \mathrm{M}(\widehat{\gamma})\}$ and (3) $\mu \geq \mu_{\mathrm{SF}}^1 := \max\{\mu_{\mathrm{S}}^1, \mu_{\mathrm{F}}^1\}$, we can ensure that $\mathbf{\Phi}_{\mu,\alpha^{*1}}(\widehat{\mathbf{\Sigma}})$ has at least the same convergence rate with $\widehat{\mathbf{\Sigma}}$.

Chapter 4

Simulation study

In this section, we numerically compare the finite sample performances of the LSPD estimator and other existing estimators. In the comparison, we use the soft-thresholding estimator (2.16) as an initial non-PD estimator. It is shown that the LSPD estimator is comparable to the existing optimization-based PD estimators (Section 4.2.), while its computation is much faster than others (Section 4.3.).

4.1. Data generation

For the study, we generated independently and identically distributed (IID) random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from (i) the p-dimensional multivariate normal distribution and (ii) the multivariate t distribution with degrees of freedom 5, where the sample size $n = 100$ and the dimension $p = 100, 200, 400$ are chosen. We consider two choices of the true covariance matrix $\boldsymbol{\Sigma}$:

1. ('Linearly tapered Toeplitz') $(\mathbf{M}_1)_{ij} := (1 - \frac{|i-j|}{10})_+$

2. ('Overlapped block-diagonal') The row (column) index $\{1, 2, \ldots, p\}$ is partitioned into $K := p/20$ subsets which are non-overlapping and of equal size. Denoting by i_k the maximum index in I_k, define $(\mathbf{M}_2)$ by

$$(\mathbf{M}_2)_{ij} := \begin{cases} 1, & \text{if } i = j; \\ 0.4, & \text{if } i \neq j \text{ and } (i,j) \in (I_k \cup \{i_k + 1\}) \times (I_k \cup \{i_k + 1\}); \\ 0, & \text{otherwise.} \end{cases}$$

3. ('Randomly sparse') $\mathbf{M}_3^0$ is a sparse matrix with sparsity $15/(p/100)\%$ in which diagonal elements are assigned to 1 and nonzero off-diagonal elements to 0.5. Then we set $\mathbf{M}_3 := \mathbf{M}_3^0 + \delta\mathbf{I}$ so that the smallest eigenvalue of $\mathbf{M}_3$ can be 10^{-2}.

Finally, we generate 100 datasets for each of all combinations of the underlying distribution, the true covariance matrix, the sample size n, and the dimension p.

4.2. Empirical risk

We compute sample covariance matrix $\mathbf{S} := \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, where $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and the soft-thresholding estimator $\widehat{\mathbf{\Sigma}}^{\text{Soft}}(\lambda)$ defined in (2.16), where the universal threshold λ is applied to all off-diagonal (not the main diagonal) elements. To find the optimal threshold λ^*, we do five-fold cross validation introduced using the validation error introduced in Section 3.5..

As the candidates of λ, we consider grids $\lambda \in \big\{ \big(k/100\big) \cdot \|\mathbf{S} - \mathrm{diag}(\mathbf{S})\|_{\max} : k = 0, 1, \ldots, 100 \big\}$, where $\|\mathbf{A}\|_{\max}$ denotes the maximum absolute value of the elements of a matrix $\mathbf{A}$ and $\mathrm{diag}(\mathbf{S})$ is a diagonal matrix whose diagonal elements coincide those of $\mathbf{S}$.

First, we investigate the spectrum of the soft thresholding estimator to understand how often and in what magnitude it becomes non-PD. Table 4.2 reports the summary of the negative eigenvalues of the soft-thresholding estimator from the simulated data set. In both cases of $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, we find that the soft-thresholding estimator is easily non-PD. In addition, the smallest eigenvalue often becomes more smaller than 0 as p increases.

In the comparison, we consider the following four PD covariance matrix estimators:

1. (LSPD(μ_{SF})) LSPD $\mathbf{\Phi}_{\mu,\alpha^*}\Big(\widehat{\mathbf{\Sigma}}^{\mathrm{Soft}}(\lambda^*)\Big)$ with $\mu = \mu_{\mathrm{SF}}$

2. (LSPD(∞)) LSPD $\mathbf{\Phi}_{\mu,\alpha^*}\Big(\widehat{\mathbf{\Sigma}}^{\mathrm{Soft}}(\lambda^*)\Big)$ with $\mu = \infty$

3. (Eig. con.) Xue et al. (2012)'s eigenvalue-constraint $\widehat{\mathbf{\Sigma}}^{\mathrm{EigCon}}(\lambda^*)$ in (2.18)

4. (log-det) Rothman (2012)'s log-determinant barrier $\widehat{\mathbf{\Sigma}}^{\mathrm{logdet}}(\lambda^*)$ in (2.17)

In applying the methods, we put ϵ, which should be predetermined in the two LSPD estimators - $\mathbf{\Phi}_{\mu,\alpha^*}\Big(\widehat{\mathbf{\Sigma}}^{\mathrm{Soft}}(\lambda^*)\Big)$ with $\mu = \mu_{\mathrm{SF}}$ and ∞ - and $\widehat{\mathbf{\Sigma}}^{\mathrm{EigCon}}(\lambda^*)$ as $\epsilon = 10^{-2}$, and the log-determinant barrier τ in $\widehat{\mathbf{\Sigma}}^{\mathrm{logdet}}(\lambda^*)$ is also set as $\tau = 10^{-2}$.

Table 4.3 shows the empirical risks based on three popular matrix norms (the matrix l_1 norm, spectral norm, and Frobenius norm) of the four estimators considered. In the table, the results for $p = 100$, 200 of $\mathbf{M}_2$ (the

		$\mathbf{M}_1$ (Tapered)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	-0.035 (0.004)	0.017 (0.001)	16/100
	t	-0.066 (0.006)	0.020 (0.001)	9/100
200	$\mathcal{N}$	-0.061 (0.003)	0.017 (0.001)	2/100
	t	-0.114 (0.020)	0.018 (0.001)	4/100
400	$\mathcal{N}$	-0.086 (0.003)	0.016 (0.001)	0/100
	t	-0.270 (0.026)	0.017 (0.001)	0/100
		$\mathbf{M}_2$ (Overlap. block diag.)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	0.257 (0.003)	0.000 (0.000)	100/100
	t	0.292 (0.012)	0.000 (0.000)	100/100
200	$\mathcal{N}$	0.138 (0.003)	0.000 (0.000)	100/100
	t	0.133 (0.018)	0.001 (0.000)	87/100
400	$\mathcal{N}$	-0.039 (0.003)	0.006 (0.000)	7/100
	t	-0.150 (0.026)	0.014 (0.001)	7/100
		$\mathbf{M}_3$ (Randomly sparse)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	0.944 (0.012)	0.000 (0.000)	100/100
	t	0.619 (0.020)	0.000 (0.000)	98/100
200	$\mathcal{N}$	1.214 (0.010)	0.000 (0.000)	100/100
	t	0.506 (0.039)	0.003 (0.001)	82/100
400	$\mathcal{N}$	1.163 (0.008)	0.000 (0.000)	100/100
	t	0.226 (0.039)	0.004 (0.001)	57/100

Table 4.1: The non-PDness of the soft thresholding estimator. Each column means minimum eigenvalue (Min. Eig), the proportion of negative eigenvalues (#(Neg. eig.)/p), and the number of cases the estimates are PD (#(PD)) over 100 replications.

overlapped block-diagonal matrix) and the normally distributed case of $p =$ 100, 200, 400 of $\mathbf{M}_3$ are the same across the estimators, since the corresponding soft-thresholding estimates themselves are mostly PD. As for the t-distrubuted case for $\mathbf{M}_3$, the algorithm of log-determinant estimator failed to convergent.

Now we compare the empirical risks between estimators. In the linearly tapered model ($\mathbf{M}_1$), LSPD(μ_{SF}) has lower risks than the soft thresholding

		$\mathbf{M}_1$ (Tapered)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	-0.377 (0.004)	0.185 (0.001)	0/100
	t	-0.401 (0.012)	0.200 (0.004)	0/100
200	$\mathcal{N}$	-0.425 (0.004)	0.200 (0.001)	0/100
	t	-0.446 (0.028)	0.185 (0.006)	4/100
400	$\mathcal{N}$	-0.459 (0.005)	0.217 (0.001)	0/100
	t	-0.526 (0.027)	0.154 (0.007)	2/100
		$\mathbf{M}_2$ (Overlap. block diag.)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	0.027 (0.005)	0.004 (0.001)	69/100
	t	0.196 (0.019)	0.001 (0.000)	90/100
200	$\mathcal{N}$	-0.123 (0.005)	0.019 (0.001)	0/100
	t	0.047 (0.024)	0.005 (0.001)	43/100
400	$\mathcal{N}$	-0.333 (0.006)	0.069 (0.001)	0/100
	t	-0.263 (0.031)	0.024 (0.002)	9/100
		$\mathbf{M}_3$ (Randomly sparse)		
p		Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	1.871 (0.014)	0.000 (0.000)	100/100
	t	0.970 (0.038)	0.000 (0.000)	96/100
200	$\mathcal{N}$	1.914 (0.015)	0.000 (0.000)	100/100
	t	0.638 (0.045)	0.003 (0.001)	81/100
400	$\mathcal{N}$	1.626 (0.014)	0.000 (0.000)	100/100
	t	0.301 (0.043)	0.004 (0.001)	64/100

Table 4.2: The non-PDness of the SCAD thresholding estimator. Each column means minimum eigenvalue (Min. Eig), the proportion of negative eigenvalues (#(Neg. eig.)/p), and the number of cases the estimates are PD (#(PD)) over 100 replications.

estimates in marix l_1 norm and spectral norm, and higher or lower risks in Frobenius norm. The risks of LSPD(∞) are higher than soft thresholding (except spectral norm-measured multivariate normal data), with risk inflation within 4%. The two optimization-based methods reports the equal or lower risks than soft thresholding.

In the overlapped block diagonal model ($\mathbf{M}_2$), the risks are almost the same across the five estimators, for all cases of $p = 100$, 200. This is because the corresponding soft-thresholding estimates themselves are mostly PD. For $p = 400$, the trend of risk is similar to those in the model 2.

As for the randomly sparse model ($\mathbf{M}_3$), the results of estimators for multivariate normal distribution are the same across in all norms since the soft thresholding estimator was already PD. The results of the multivariate t-generated data is interesting, since the risks are reduced as μ gets large in LSPD estimation. We conjecture that is because the soft thresholding estimates for $\mathbf{M}_3$ has bias downward from the true and the diagonal shifts in LSPD correction reduced bias.

The overall result shows that each estimator have risk difference are within 4% from others, and, furthermore, the standard errors of risks shows that each estimator has the confidence interval in which other risks are possessed. Therefore, we can conclude that the PD-correction methods have comparable error with the soft-thresholding estimator in finite sample.

It is interesting that LSPD methods have comparable performance with other two optimization-based estimators, despite of its methodological simplicity. We further investgated how the linear shrinkage affects the risk of original estimator. In Table A.2 through A.4 attached in the Appendix, we partitioned the matrix into the three part (the diagonals, the support of off-diagonal elements of the true matrix, the zero set of off-diagonal elements of the true matrix) and measured the risks separately for three norms. As one expects, the comparable risk of LSPDs can be interpreted as a result of trade-off. In the off-diagonal parts, the LSPD shrinks the original estimate. The tables tells that the risk of LSPD is more reduced in the zero set of off-diagonal elements of the true matrix than the original estimate; and is more biased in the support of off-diagonal elements. In the diagonal parts, the shrinkage toward μ from LSPD inflated the risk of the original in the case $\mathbf{M}_1$ and $\mathbf{M}_2$, and reduced the risk in the case $\mathbf{M}_3$.

4.3. Computation time

In previous subsection, we observe that, in finite sample, the empirical risks of the LSPD estimators are comparable to the soft-thresholding estimator as well as others; all estimators are minimax optimal in asymptotic if true covariance matrix is sparse. We now numerically show that the proposed LSPD estimator is computationally much faster and simpler than the optimization-based PD estimators.

To measure the computation times, we calculate of the four PD covariance matrix estimates and the soft-thresholding estimate for one data, generated from the sample size $n = 100$ and varying dimension $p = 400$, 1200, and 3600. The distribution is multivariate normal with the true covariance matrix $\mathbf{\Sigma} = \mathbf{M}_1$ and $\mathbf{M}_2$. Here, λ is fixed as a constant and the row 'NZ' indicates the corresponding proportion of non-zero elements of each estimator. The computation of this section is made using MATLAB, operated on the computer of Intel Core i7 CPU (3.4GHz) and 16GB RAM. As for the optimization-based estimators, we consider the optimization procedure converged if $\|\widehat{\mathbf{\Sigma}}^{(\mathrm{New})} - \widehat{\mathbf{\Sigma}}^{(\mathrm{Old})}\| / \|\widehat{\mathbf{\Sigma}}^{(\mathrm{Old})}\| < 10^{-7}$. The results are summarized in Table 4.4. It is not very surprising that both LSPD estimators apparently faster than both the eigenvalue constraint and log-determinant barrier estimator. To find the reason of speedup, we invoke that the eigenvalue constraint method need iterative eigenvalue decompositon, whereas LSPD does not demand even one eigenvalue decomposition. And log-determinant method needs iterative computation of $O(p^3)$ flops which consume time as much as eigenvalue decomposition. For the comparison of two LSPD estimators, LSPD(∞) is always faster than LSPD(μ_{SF}) since LSPD(μ_{SF}) needs more computation due to $\widehat{\gamma}_{\max}$, $\mathrm{M}(\widehat{\gamma})$, and $\mathrm{V}(\widehat{\gamma})$.

Table 4.3: The result of simulations for empirical risk comparison of the universal soft thresholding estimator and its PD-variants. Each column means the norm used to measure risk. The definitions for $\mathbf{M}_i$'s are introduced in Section 4.2..

	Multivariate normal			Multivariate t		
	Matrix l_1	Spectral	Frobenius	Matrix l_1	Spectral	Frobenius
			$\mathbf{M}_1$, $p = 100$			
Soft thres.	6.21 (0.11)	3.59 (0.05)	7.18 (0.07)	9.20 (0.35)	5.06 (0.12)	10.37 (0.17)
LSPD(μ_{SF})	6.20 (0.11)	3.59 (0.05)	7.25 (0.07)	9.12 (0.34)	5.04 (0.12)	10.45 (0.17)
LSPD(∞)	6.20 (0.11)	3.56 (0.05)	7.21 (0.07)	9.23 (0.35)	5.04 (0.13)	10.41 (0.17)
Eig. con.	6.21 (0.11)	3.59 (0.05)	7.18 (0.07)	9.19 (0.34)	5.06 (0.12)	10.37 (0.17)
log-det.	6.21 (0.11)	3.59 (0.05)	7.22 (0.06)	9.18 (0.34)	5.06 (0.12)	10.40 (0.17)
			$\mathbf{M}_1$, $p = 200$			
Soft thres.	7.08 (0.08)	4.24 (0.04)	11.35 (0.06)	13.40 (0.77)	6.25 (0.19)	17.18 (0.42)
LSPD(μ_{SF})	7.06 (0.08)	4.23 (0.04)	11.54 (0.05)	13.12 (0.72)	6.18 (0.18)	17.49 (0.43)
LSPD(∞)	7.05 (0.08)	4.16 (0.04)	11.40 (0.05)	13.51 (0.78)	6.24 (0.21)	17.40 (0.45)
Eig. con.	7.08 (0.08)	4.23 (0.04)	11.35 (0.06)	13.29 (0.74)	6.25 (0.19)	17.18 (0.42)
log-det.	7.07 (0.08)	4.23 (0.04)	11.41 (0.05)	13.28 (0.74)	6.25 (0.19)	17.23 (0.42)
			$\mathbf{M}_1$, $p = 400$			
Soft thres.	7.91 (0.08)	4.72 (0.03)	17.75 (0.06)	19.34 (1.12)	7.47 (0.30)	25.94 (0.40)
LSPD(μ_{SF})	7.86 (0.07)	4.71 (0.03)	18.14 (0.06)	18.58 (1.02)	7.28 (0.28)	26.92 (0.51)
LSPD(∞)	7.93 (0.08)	4.62 (0.03)	17.86 (0.06)	19.60 (1.15)	7.57 (0.34)	26.87 (0.57)
Eig. con.	7.90 (0.08)	4.72 (0.03)	17.74 (0.06)	18.92 (1.05)	7.45 (0.30)	25.92 (0.39)
log-det.	7.88 (0.08)	4.72 (0.03)	17.84 (0.06)	18.90 (1.05)	7.45 (0.30)	25.99 (0.39)
			$\mathbf{M}_2$, $p = 100$			
Soft thres.	3.18 (0.03)	1.68 (0.02)	5.18 (0.02)	4.95 (0.14)	2.30 (0.03)	7.56 (0.11)
LSPD(μ_{SF})	3.18 (0.03)	1.68 (0.02)	5.18 (0.02)	4.95 (0.14)	2.30 (0.03)	7.56 (0.11)
LSPD(∞)	3.18 (0.03)	1.68 (0.02)	5.18 (0.02)	4.95 (0.14)	2.30 (0.03)	7.56 (0.11)
Eig. con.	3.18 (0.03)	1.68 (0.02)	5.18 (0.02)	4.95 (0.14)	2.30 (0.03)	7.56 (0.11)
log-det.	3.18 (0.03)	1.68 (0.02)	5.18 (0.02)	4.95 (0.14)	2.30 (0.03)	7.56 (0.11)
			$\mathbf{M}_2$, $p = 200$			
Soft thres.	6.04 (0.05)	2.97 (0.03)	9.87 (0.03)	9.54 (0.26)	4.04 (0.06)	14.54 (0.23)
LSPD(μ_{SF})	6.04 (0.05)	2.97 (0.03)	9.87 (0.03)	9.51 (0.26)	4.03 (0.06)	14.54 (0.23)
LSPD(∞)	6.04 (0.05)	2.97 (0.03)	9.87 (0.03)	9.55 (0.27)	4.04 (0.06)	14.55 (0.23)
Eig. con.	6.04 (0.05)	2.97 (0.03)	9.87 (0.03)	9.52 (0.26)	4.04 (0.06)	14.54 (0.23)
log-det.	6.04 (0.05)	2.97 (0.03)	9.87 (0.03)	9.51 (0.26)	4.04 (0.06)	14.54 (0.23)
			$\mathbf{M}_2$, $p = 400$			
Soft thres.	11.77 (0.10)	5.64 (0.05)	19.29 (0.07)	19.67 (0.58)	7.44 (0.08)	27.91 (0.37)
LSPD(μ_{SF})	11.78 (0.10)	5.62 (0.05)	19.39 (0.07)	19.10 (0.52)	7.31 (0.07)	28.23 (0.38)
LSPD(∞)	11.73 (0.10)	5.59 (0.05)	19.32 (0.07)	19.83 (0.60)	7.39 (0.09)	28.32 (0.40)
Eig. con.	11.77 (0.10)	5.64 (0.05)	19.28 (0.07)	19.34 (0.54)	7.44 (0.08)	27.89 (0.37)
log-det.	11.75 (0.10)	5.63 (0.05)	19.23 (0.07)	19.31 (0.54)	7.43 (0.08)	27.86 (0.38)
			$\mathbf{M}_3$, $p = 100$			
Soft thres.	12.63 (0.07)	7.09 (0.03)	19.57 (0.03)	13.62 (0.07)	8.02 (0.03)	22.56 (0.08)
LSPD(μ_{SF})	12.63 (0.07)	7.09 (0.03)	19.57 (0.03)	13.62 (0.07)	8.02 (0.03)	22.56 (0.08)
LSPD(∞)	12.63 (0.07)	7.09 (0.03)	19.57 (0.03)	13.62 (0.07)	8.02 (0.04)	22.55 (0.08)
Eig. con.	12.63 (0.07)	7.09 (0.03)	19.57 (0.03)	13.62 (0.07)	8.02 (0.03)	22.56 (0.08)
log-det.	12.63 (0.07)	7.09 (0.03)	19.57 (0.03)		(Algorithm failed to converge)	
			$\mathbf{M}_3$, $p = 200$			
Soft thres.	13.85 (0.06)	8.25 (0.01)	31.25 (0.03)	14.79 (0.06)	8.93 (0.02)	35.89 (0.15)
LSPD(μ_{SF})	13.85 (0.06)	8.25 (0.01)	31.25 (0.03)	14.74 (0.06)	8.93 (0.02)	35.78 (0.14)
LSPD(∞)	13.85 (0.06)	8.25 (0.01)	31.25 (0.03)	14.76 (0.06)	8.90 (0.03)	35.64 (0.14)
Eig. con.	13.85 (0.06)	8.25 (0.01)	31.25 (0.03)	14.78 (0.06)	8.93 (0.02)	35.88 (0.15)
log-det.	13.85 (0.06)	8.25 (0.01)	31.25 (0.03)		(Algorithm failed to converge)	
			$\mathbf{M}_3$, $p = 400$			
Soft thres.	15.08 (0.05)	8.68 (0.01)	46.07 (0.03)	15.94 (0.06)	9.28 (0.02)	52.63 (0.18)
LSPD(μ_{SF})	15.08 (0.05)	8.68 (0.01)	46.07 (0.03)	15.87 (0.05)	9.27 (0.02)	52.36 (0.18)
LSPD(∞)	15.08 (0.05)	8.68 (0.01)	46.07 (0.03)	15.87 (0.05)	9.20 (0.02)	51.72 (0.21)
Eig. con.	15.08 (0.05)	8.68 (0.01)	46.07 (0.03)	15.94 (0.05)	9.28 (0.02)	52.62 (0.18)
log-det.	15.08 (0.05)	8.68 (0.01)	46.07 (0.03)		(Algorithm failed to converge)	

	$\mathbf{M}_1$: Tapered			$\mathbf{M}_2$: Overlap. block diag.		
	$p = 400$	$p = 1200$	$p = 3600$	$p = 400$	$p = 1200$	$p = 3600$
Soft thres.	0.00	0.02	0.23	0.00	0.03	0.24
(NZ)	13.6%	4.0%	1.4%	15.0%	5.7%	2.1%
LSPD(μ_{SF})	0.01	0.12	0.66	0.01	0.13	0.82
(NZ)	13.6%	4.0%	1.4%	15.0%	5.7%	2.1%
LSPD(∞)	0.01	0.09	0.50	0.01	0.09	0.58
(NZ)	13.6%	4.0%	1.4%	15.0%	5.7%	2.1%
log-det	2.33	99.14	3157.80	2.30	97.28	3156.08
(NZ)	13.1%	3.8%	1.3%	14.6%	5.5%	2.0%
Eig. con.	4.93	190.68	7757.47	2.42	106.13	4470.47
(NZ)	13.5%	3.9%	1.3%	14.7%	5.5%	2.0%

Table 4.4: Computational times of the four PD estimators and the soft-thresholding estimator for a given λ. In each cell, The first row means the compuation time measured in seconds and he second row indicates the proportion of non-zero elements of the estimators.

Chapter 5

Two applications

In this chapter, we apply the LSPD-modified covariance matrix estimators to two statistical procedures in literature - linear minimax classifiacation and Markowitz's portfolio allocation, which require the PD estimator of the covariance matrix. Here, both applications are illustrated with real data examples. The linear minimax classifier is illustrated with an example on speech recognition problem (Tsanas et al., 2014) and the Markowitz's portfolio allocation is illustrated with the Dow Jones' stock return example in (Won et al., 2013).

5.1. Liniar minimax classifier with application to speech recognition

5.1.1. Linear minimax probability machine

In a binary classification problem, Lanckriet et al. (2002) propose a new classifier, denoted by linear minimax probability machine (linear MPM), which does not make any distributional assumption except the mean vector and covariance matrix. The linear MPM finds a separating hyperplane that minimizes the maximum probability of misclassification over all distributions with given mean vectors and covariance matrices:

$$\begin{aligned} \max_{\alpha, \mathbf{a}\neq 0, b} \text{ s.t. } &\inf_{\mathbf{x}\sim(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)} P\left\{\mathbf{a}^\top \mathbf{x} \leq b\right\} \geq \alpha \\ &\inf_{\mathbf{y}\sim(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)} P\left\{\mathbf{a}^\top \mathbf{x} \geq b\right\} \geq \alpha \end{aligned} \tag{5.1}$$

where $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ (and $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, respectively) is a pair of the population mean vector and covariance matrix for group 0 (and 1, respectively). The classifier is known as a robust classifier since it is distribution free and also minimizes the worst-case misclassification probability. Lanckriet et al. (2002) shows that the problem (5.1) can be rewritten as

$$\max_{\mathbf{a}} \frac{\left|\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right|}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_0 \mathbf{a}} + \sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_1 \mathbf{a}}}. \tag{5.2}$$

if the denominator is replaced with $\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_0 \mathbf{a} + \mathbf{a}^\top \boldsymbol{\Sigma}_1 \mathbf{a}}$, its solution equals to the well-known Fisher's discriminant analysis and has an explicit form as $\mathbf{a} \propto (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. However, the linear MPM does not have an explicit form and is given as a solution to (5.2) which is a convex optimization problem. The PDness of the covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ is important because it makes the convex problem have the unique solution and an algorithm that globally converges to it.

In practice, the true $\boldsymbol{\mu}$'s and $\mathbf{\Sigma}$'s are unknown, we plug in their estimators to the MPM problem (5.2). Lanckriet et al. (2002) uses the sample mean vectors and covariance matrices when they are well defined. In case the sample covariance matrix $\mathbf{S}$ is singular, they suggest to use $\mathbf{S} + \delta\mathbf{I}$ with a given constant δ instead of $\mathbf{S}$.

5.1.2. Example: Speech recognition

We illustrate the performance of the linear MPM with various PD covariance matrix estimators using real data on speech recognition (Tsanas et al., 2014). The dataset is on 126 signals which are the pronounce of `/a/` by 14 Parkinson's disease patients; each patient makes 9 trials. Each signal is pre-processed into 309 features, and is labeled as 'acceptable' or 'unacceptable' by an expert. The dataset is the form of 126 × 309 matrix with binary labels. It is on-line available from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

To estimate the classification accuracy, we randomly split the dataset into 90% of training samples and 10% of testing samples, where the linear MPM is constructed using the training samples and the classification accuracy is measured using the testing samples. In building the linear MPM, we use various PD covariance matrices discussed in Section 3 for the true covariance matrices, whereas the true mean vectors are estimated with the sample means. The PD covariance estimators used are: (1) “Samp”, the sample covariance matrix with diagonal shift $\delta = 10^{-2}$, (2) “Diagsamp”, the diagonal matrix of the sample variances, (3) “LW”, the linear shrinkage estimator by Ledoit and Wolf (2004), (4) “Condreg”, the condition number regularized estimator by Won et al. (2013), (5) “Adap+eig.con.”, the eigenvalue constraint estimator by Xue et al. (2012) based on the adaptive thresholding estimator by Cai and Liu (2011), (6) “Adap+LSPD”, the proposed LSPD estimate ($\alpha = \alpha^*, \mu = \mu_{\mathrm{SF}}$) based on the adaptive thresholding estimator. Here, unlike the numerical study in Section 4, we use the adaptive thresholding estimator as an initial regularized covariance estimator instead of the (universal) soft-thresholding estimator. This is because the marginal variances are unknown and could be unequal over variables. In addition, we omit Rothman's log-determaninant estimator based on adaptive thresholding estimator since it did not converge.

Some technical details of the covariance matrix estimation are as follows. The five-fold cross-validation method is used to select the tuning parameter of the adaptive thresholding estimator. The pre-determined constant ϵ for Adap+eig.con. and Adap+LSPD is set as 10^{-2}. We make 100 random partitions (90% of training and 10% of testing) to compute classification accuracy which is evaluated from testing samples. The average and standard deviation of classification accuracy (the rate of correct classification) over 100 random partitions are reported in Table 5.1. In the result, all regularization methods show better classification accuracy than naive sample covariance matrix. In addition, the two linear shrinkage methods (LW, Adap+LSPD) have the highest accuracies among six considered. In particular, the LSPD update of adaptive thresholding estimator has 89.2% average accuracy with 9.2% standard deviation. This rate is almost equal to the rate 90% reported in the original paper Tsanas et al. (2014) that is based on random forest and support vector machine after a feaature selection algorithms named as LOGO Sun et al. (2010). We also find that it is interesting that Adap+LSPD showed better performance better than Adap+eig.con, since both perform similarly in the numerical study in Chapter 4.

Samp	Diagsamp	LW	Condreg	Adap+eig.con.	Adap+LSPD
73.8 (12.4)	76.4 (12.6)	86.9 (10.7)	75.3 (13.6)	82.0 (18.1)	89.2 (9.2)

Table 5.1: The average of the rates of correct classification (s.d.) for linear MPM with selected PD covariance matrice estimatiors based on 100 random partitions of data into 10% of test set and 10% of training set. The abbreviations of the estimators are introduced in the mainbody of the Section.

5.2. Markowitz portfolio optimization

5.2.1. Minimum-variance portfolio (MVP) allocation and shortsale

In finance, portfolio refers to a family of (risky) assets held by an institution or a private individual. If there are many assets to invest, usually a combination of assets is considered and it becomes an important issue to select an optimal allocation of portflio. The mean-variance portfolio optimization (Markowitz, 1952) is one of well-established strategies for portfolio allocation. The author proposes to choose a portfolio which minimizes risk when the expected level of return is given. Here, risk is understood as the standard deviation of return. To describe the problem quantitively, let $\mathbf{r} := (r_1, \ldots, r_p)^\top$ be a p-variate random vector in which each r_j represents the return of the j-th asset constituing to the portfolio ($j = 1, \ldots, p$). Denote by $\boldsymbol{\mu} := \mathbb{E}(\mathbf{r})$ and $\boldsymbol{\Sigma} := \mathbb{V}\text{ar}(\mathbf{r})$ the expect return and covariance matrix of assets which are unknown. Write $\mathbf{w}$ as a p by 1 vector of weight of the investor's wealth such that each w_j stands for the weight of the j-th asset. Then the mean-variance portfolio optimization problem is expressed as

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{1} = 1,\ \mathbf{w}^\top \boldsymbol{\mu} = \beta, \tag{5.3}$$

where $\mathbf{1}$ is a vector of ones and β is a given level of expected return. The solution of (5.3) indicates an optimal allocation of the investor's resource to each asset in the portfolio.

As in the linear MPM in Section 5.1, (5.3) depends on the unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the sample mean vector and covariance matrix are plugged in when they are well defined. Here, we focus on the estimation of covariance matrix, we inspect the following *minimum variance portfolio* (MVP) optimization problem (Chan, 1999) which excludes the expected-return-constraint in (5.3):

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{1} = 1. \tag{5.4}$$

Both the MVPs (5.3) and (5.4) allow their solutions to have a weight negative or greater than 1. These weights are interpreted respectively short-sale and leverage in stock market. Indeed, if there is a dominating factor in true covariance structure, the solution would result in extreme negative weights (Green and Hollifield, 1992). However, shortsale and leverage are sometimes impractical because of legal issues. Thus it is also natural to consider an optimization problem with no-shortsale constraint :

$$\underset{\mathbf{w}}{\text{minimize}}\ \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{1} = 1,\ \ \mathbf{w} \geq \mathbf{0}, \tag{5.5}$$

where $\mathbf{w} \geq \mathbf{0}$ is defined componentwise. The combination of two constraints in (5.5) ensures the solution restrected to $[0, 1]$. Undoubtedly, it is problematic that the solution may have a bias from the true optimal weight due to nonnegativity enforcement, since a bias can inflate the risk of corresponding portfolio. For this issue, Jagannathan and Ma (2003) empirically and analytically illustrate that (5.5) could have smaller the risk than (5.4), even if the no-shortsale constraint is wrong. We note that the paper handles only the case that sample covariance matrix is well defined and plugged-in for unknown $\boldsymbol{\Sigma}$. In principle, $\boldsymbol{\Sigma}$ can be replaced by a suitable estimator which is PD.

5.2.2. Example: Dow Jones stock return

The aims of this analysis are not only to reproduce the findings of Jagannathan and Ma (2003) that no-shortsale constraint does not harm the risk of (5.4) in a different data, but also to inspect whether the same conclusion may hold empirically for another choice of PD covariance matrix estimator. We compare the two portfolio optimization scheme - minimum variance portfolio optimization with shortsale allowed (5.4) and with no-shortsale constraint (5.5). The covariance matrix $\boldsymbol{\Sigma}$ is plugged-in by the seven of PD covariance matrix estimators: the five estimators (1) “Samp”, (2) “LW”, (3) “Condreg”, (4) “Adap+eig.con”, and (5) “Adap+LSPD” are already introduced in Section 5.1.2. and we further import (6) “POET+eig.con.”, the eigenvalue constraint estimator based on POET estimator proposed by Fan et al. (2013) (see Xue et al.’s discussion section on the cited paper) and (7) “POET+LSPD”, our LSPD modification on the POET estimator. The reason why we consider (6) and (7) is that usually the stock return data have a factor structure and the POET estimation is believed to reflect the structure.

The dataset used is daily returns of 30 stocks which are constituted of the Dow-Jones industrial average (DJIA) at July 2008 and was previously analyzed in Won et al. (2013). It contains daily closing prices from December 1992 to June 2008, with adjusting splits and dividend distributions. The followings are how we constructed portfolios. For covariance matrix estimation, we use the stock returns of past 60 or 240 trading days (approximately 3 or 12 months respectively). Condreg, Adap and POET methods need tuning parameter selections and they are done by five-fold cross validation, treating the return of each day as independent sample. Once a portfolio is established by solving (5.4) and (5.5) with covariance matrix plugged-in by its estimators, we hold it thorugh 60 trading days. This process begins from February 18th, 1994 and continually is repeated until it ends in July 6th, 2008 with 60 of holding periods in total. We record all the returns for each trading day and summarize them in the form of realized return, realized risk, and Sharpe ratio. The *realized return* and *realized risk* of portfolio are respectively understood the average and standard deviation of daily returns from corresponding portfolio. The *Shape ratio* is a risk-adjusted index of performance, which is defined by {(realized return) - (risk-free rate)}/(realized risk). The risk-free rate is set as 5% per year.

	Realized return [%]		Realized risk [%]		Sharpe Ratio	
	Simple	No Short.	Simple	No Short.	Simple	No Short.
(Portfolio rebalancing based on 60 previous trading days)						
Sample	22.49	23.11	3.28	3.34	5.33	5.41
LedoitWolf	21.25	21.61	3.11	3.06	5.23	5.44
CondReg	24.70	24.70	4.17	4.16	4.73	4.73
Adap.+LSPD	22.74	23.26	3.59	3.62	4.95	5.04
Adap.+EigCon	22.74	23.38	3.35	3.40	5.30	5.40
POET+LSPD	20.99	21.85	3.13	3.07	5.11	5.49
POET+EigCon	20.81	21.28	3.18	3.06	4.96	5.32
(Portfolio rebalancing based on 240 previous trading days)						
Sample	22.42	23.07	3.33	3.37	5.24	5.36
LedoitWolf	21.72	23.08	2.89	2.94	5.78	6.14
CondReg	24.91	24.73	4.18	4.18	4.76	4.72
Adap.+LSPD	22.68	23.13	3.57	3.60	4.95	5.04
Adap.+EigCon	21.25	22.47	3.30	3.36	4.92	5.19
POET+LSPD	21.49	23.94	3.02	2.98	5.46	6.35
POET+EigCon	20.93	23.49	3.07	2.99	5.18	6.19

Table 5.2: Full table of empirical out-of-sample performances, from 30 constituents of DJIA with 60 days of holding, starting from 2/18/1994. All the rates are annualized.

We present some interpretations and conjectures based on the summarized results in Table 5.2. We first compare realized risks of simple MVPs and non-short MVPs. The 3rd and 4th columns of the table show that differnences of their realized risks are negligible. As for sample covariance matrix, it reproduces the findings of Jagannathan and Ma (2003). In addition, we find that the regularized covariance matrix estimators (LW, Condreg, Adap, POET) give also similar realized risks. For the realized returns (the 1st and 2nd columns), it is interesting that no-short MVPs give larger realized returns than simple MVPs except CondReg. This leads an improvement of

Sharpe ratios in the cases except CondReg. The reason of return improvement from no-shortsale constraint is still vague and we guess that further empirical data analyses are needed.

Next, we compare the results of 60-day- and 240-day-training for the construction of portfolios (the 1st-6th rows versus the 7th-12th rows). Unlike Samp and Adap, both LW and POET show higher realized returns and Sharpe ratio in 240-day-training than 60-day-training. This is true for both simple and no-short MVPs. We conjecture that the factored structure of the POET explain better the latent stracutures of stock in long-history data that in short-history one. On the other hand, the higher weight on the identity matrix in LW make the porfolio behaves as an equal-weight investment strategy, which is known to work well in long-term investment.

Condreg performs somewhat differently from other regualarized covariance matrix estimators. The realized return of Condreg dominates other methods. It coincides the results in Won et al. (2013), which show that the MVR with Condreg gives the highest the wealth growth. This shows an empirical evidence of the spirit of 'high risk, high return', since ConrReg also showes the highest realized risk.

Finally, we compare the two PD covariance matrix estimation methods - LSPD estimators (Adap+LSPD and POET+LSPD) and the eigenvalue constraint estimators (Adap+eig.con. and POET+eig.con.). For POET-based estimation, LSPD methods show higher Sharpe ratio than the eigenvalue constraint methods in all case. For adaptive thresholding estimation, the eigenvalue constraint methods have higher Sharpe ratio than LSPD methods expect the case simple MVR based on 60 trading days). However, the

POET+LSPD method with 240-trading day gives the highest Sharp ratio for both MVP with simple and no-short sale. In addition, the LSPD estimators are far simpler and faster than the eigenvalue constraint estimator and, thus they are more desirable for practicioners.

Chapter 6

Extension to other covariance matrix estimators: precision matrices

In this section, we apply the LSPD method to estimating the precision matrix estimator. Indeed, in the theory developed in section , the initial estimator $\hat{\mathbf{\Sigma}}$ is a generic symmetric matrix and Theorem 3.4 can be applied to the estimation of a PD precision matrix. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the unknown true precision matrix and $\widehat{\mathbf{\Omega}}$ be its initial estimator which is possibly non-PD. The parameters α^* and μ_{SF} in Theorem 3.4 defined for $\widehat{\mathbf{\Sigma}}$ are straightforwardly applied to $\widehat{\mathbf{\Omega}}$ along with its eigenvalues $\widehat{\gamma}_1 \geq \ldots \widehat{\gamma}_p$.

Corollary 6.1 (Convergence rate of LSPD-modified precision matrix estimator)**.** Let $\widehat{\mathbf{\Omega}}$ be any estimator of the unknown true precision matrix $\mathbf{\Omega}$. If

$\epsilon \leq \gamma_{\min}(\Omega)$, then

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Omega}}\right) - \mathbf{\Omega} \right\| \leq 2 \left\| \widehat{\mathbf{\Omega}} - \mathbf{\Omega} \right\|, \quad \mu \in [\mu_{\mathrm{SF}}(\widehat{\mathbf{\Omega}}), \infty),$$

in both spectral norm and Frobenius norm, where α^* and μ_{SF} are similarly defined as in Theorem 3.4 with the eigenvalues of $\widehat{\mathbf{\Omega}}$.

Corollary 6.1 implies that, as in the estimation of the PD covariance matrix, the LSPD-modified precision matrix estimator preserves both the support and the convergence rate in spectral and Frobenius norm of the initial estimator.

As in the covariance matrix, many estimators are proposed for sparse high dimensional precision matrix and they are frequently non-PD. Constrained l_1 regularization estimator such as CLIME (Cai et al., 2011) is one example. Penalized regression estimators - neighborhood selection (Meinshausen and Bühlmann, 2006), SPACE (Peng et al., 2009), symmetric lasso (Friedman et al., 2010), and CONCORD (Khare et al., 2015) - do not have a remedy for the PDness and are possiblly non-PD. Finally, penalized Gaussian likelihood methods ensures the PDness of the solution by its definition. However, often the optimization algorithms therein provide approximate solution which is not PD. For instance, Mazumder and Hastie (2012) points out that the popular `R` package `glasso` (version 1.7) for the graphical lasso algorithm by (Friedman et al., 2008; Witten et al., 2011) computes the precision matrix in an indirect way and could result in non-PD solution.

Depite the possibility of the non-PDness, the precesion matrix estimators listed above are likely to be PD and minimally suffered by its non-PDness.

Let $\widehat{\mathbf{\Omega}}(\lambda)$ stand for precison matrix estimators which we tried, equipped with a tuning paramter λ. Although the non-PDness of $\widehat{\mathbf{\Omega}}(\lambda)$ is often arises when λ is not large and $\widehat{\mathbf{\Omega}}(\lambda)$ is dense. However, we observe that $\widehat{\mathbf{\Omega}}(\lambda)$ with optimal selection of λ^* (selected by cross-validation) is PD in most of cases. The case $\mathbf{\Omega} = \mathbf{M}_1$ (tapered Toeplitz matrix) is the only case that we find non-PDness in some precision matrix estimators (CONCORD and symmetric lasso). Table 6 summarizes the spectral information of these estimators. Here, we find that even they yields non-PD estimates, the minimum eigenvalue is close to 0.

		CONCORD		Symmetric lasso	
p		Min. eig.	#(PD)	Min. eig.	#(PD)
100	$\mathcal{N}$	2.49e-03 (2.44e-05)	100/100	2.55e-03 (2.66e-05)	100/100
	t	2.12e-03 (4.07e-05)	100/100	2.23e-03 (4.89e-05)	100/100
200	$\mathcal{N}$	5.90e-04 (1.03e-05)	100/100	6.55e-04 (1.03e-05)	100/100
	t	3.05e-04 (2.41e-05)	92/100	4.11e-04 (2.33e-05)	94/100
400	$\mathcal{N}$	-3.44e-04 (4.09e-05)	14/100	-1.60e-04 (1.00e-07)	4/100
	t	-3.43e-04 (1.21e-04)	23/100	-2.92e-04 (2.80e-05)	9/100

Table 6.1: Spectral information of selected sparse precision matrix estimators, when true precision matrix is $\mathbf{M}_1$ (tapered Toeplitz).

Chapter 7

Concluding remarks

In this thesis, we propose a simple but novel one-step updating rule to make “any type of” covariance matrix estimator, which is possibly non-PD, be PD. We denote the updating rule as LSPD method and the resulting estimator as LSPD estimator. The existing PD covariance matrix estimators are variants of l_1-penalization estimators which are updates of a soft-thresholded sample covariance matrix. They are very adapted to the soft-thresholding estimator and not directly applicable to other types of sparse covariance matrix estimators. In addition, they are computationally expensive. Unlike the existing estimators, the LSPD estimato has many advantages in both theory and computation. First, the LSPD estimator not only perfectly preserves both sparse structure and convergence rate of the inital estimator, whihc is often known to be statistical optimal. Second, the LSPD update is very generic and can be applied to any non-PD covariance matrix estimators, not restrictied to a specific type of covariance matrix estimatores. In addition, as shown in Section 6. the method even can be applied to the estimation of PD precision matrix. Third, the LSPD method is optimization-free and computationally much simple compared to the existing estimators.

One could ask the behavior of μ_{S} and μ_{F} to characterize when they gets close in some asymptotic sense. Note that μ_{S} and μ_{F} depend only on the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. Thus it should be prior to analyze the spectral properties of $\widehat{\boldsymbol{\Sigma}}$. We first remark that if $\widehat{\boldsymbol{\Sigma}}$ is a general symmetric matrix, there would be no tendency on corresponding $\mu_{\mathrm{S}} - \mu_{\mathrm{F}}$. Figure 7.1 displays a set of spectrums of symmetric matrices whose largest and smallest eigenvalues are normalized to 1 and 0, respectively. From the normalization, μ_{S} for those matrices are all 0.5, and μ_{F} can vary on [0,1]. The figure tells us that μ_{F} could take indeed any value in [0,1]. Next, even if we regard $\widehat{\boldsymbol{\Sigma}}$ as an covariance matrix estimator, still $\widehat{\boldsymbol{\Sigma}}$ have many candidates as introduced in Section 1 and the spectral analysis of each estimator is on development. Indeed, there are only a limited amount of related works (Chapter 7 of Bai and Silverstein (2010) for example). We postpone this direction for further research.

We finally conclude the paper with the discussion on whether the proposed LSPD method is "optimal modification" in view of distance minimization given in Section 3. In Section 3, we formulate the correction of the PDness as a distance minimization problem that is:

$$\text{minimize} \left\{ \left\| \mathbf{T}^* - \mathbf{T} \right\| \, : \, \gamma_{\min}(\mathbf{T}^*) \geq \epsilon, \; \mathrm{supp}(\mathbf{T}^*) = \mathrm{supp}(\mathbf{T}) \right\}, \qquad (7.1)$$

where $\mathbf{T}$ is any symmetric matrix. The question at here can be rephrases as whether the linear shrinkage of $\mathbf{T}$ acheives lower bounds of $\|\mathbf{T}^* - \mathbf{T}\|$. We

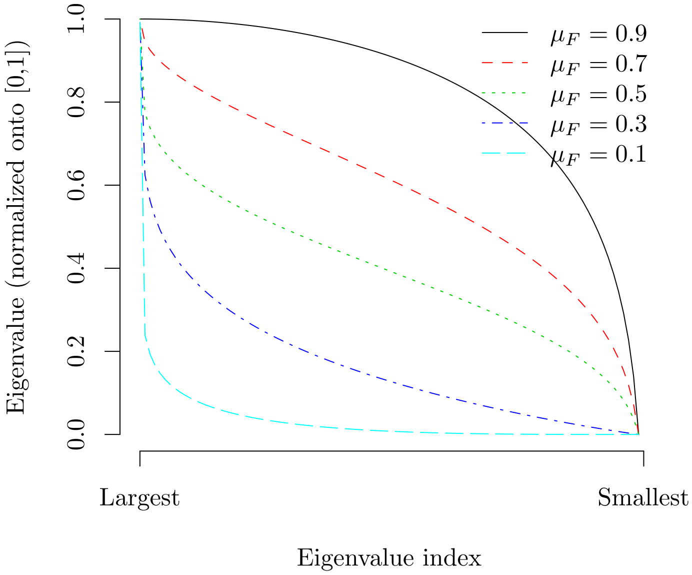

Figure 7.1: Note that $\mu_{\mathrm{S}} \equiv 0.5$ since the largest and smallest eigenvalues are normalized to 1 and 0, respectively.

answer this in spectral and Frobenius norm separately. First, for the spectral norm, it is easy to see that $\|\mathbf{T}^* - \mathbf{T}\|_2 \geq \gamma_{\min}(\mathbf{T}^*) - \gamma_{\min}(\mathbf{T}) \geq \epsilon - \gamma_{\min}(\mathbf{T})$ for any symmetric $\mathbf{T}^*$. Since Lemma implies that the LSPD acheives this bound exactly, we can conclude that LSPD estimator minimizes the distance in (7.1) and is optimal in spectral norm. On the other hand, for the Frobenius norm, it is unclear whether the LSPD estimator acheives the optimality. The Frobenius-norm distance is bounded by $\|\mathbf{T}^* - \mathbf{T}\|_{\mathrm{F}} \geq \frac{1}{p}\|\mathbf{T}^* - \mathbf{T}\|_2 \geq \frac{1}{p}(\epsilon - \gamma_{\min}(\mathbf{T}))$ for any symmetric $\mathbf{T}^*$ and $\mathbf{T}$. For the LSPD estimator, we can show that the LSPD estimator as $\|\mathbf{T}^{\mathrm{L}} - \mathbf{T}\|_{\mathrm{F}} \geq C(\epsilon - \gamma_{\min}(\mathbf{T}))$ for a constant C. This implies that the the LSPD estimator does not have $O(1/p)$-rate error and, at a first sighgt, is not able to acheive the lower bound. However, the sharpness of the lower bound $\frac{1}{p}(\epsilon - \gamma_{\min}(\mathbf{T}))$ is questionable since it does not rely on the same-support constraint $\mathrm{supp}(\mathbf{T}^*) = \mathrm{supp}(\mathbf{T}))$. We leave this as an open discussion to the readers.

Bibliography

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2nd edition.

Banerjee, O., Ghaoui, L. E., D'Aspremont, A., and El Ghaoui, L. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *The Journal of Machine Learning Research*, 9:485–516.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.

Cai, T. and Liu, W. (2011). Adaptive Thresholding for Sparse Covariance Matrix Estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T., Liu, W., and Luo, X. (2011). A Constrained l_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. and Low, M. (2011). A framework for estimation of convex functions.

Cai, T., Liu, W., and Zhou, H. (2012). Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation. *arXiv preprint arXiv:1212.2882*.

Cai, T., Ren, Z., and Zhou, H. H. (2014). Estimating Structured High-Dimensional Covariance and Precision Matrices : Optimal Rates and Adaptive Estimation. Technical report.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420.

Chan, L. (1999). On portfolio optimization: forecasting covariances and choosing the risk model. *Review of Financial Studies*, 12(5):937–974.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76(2):373–397.

D'Aspremont, A., Banerjee, O., and Ghaoui, L. E. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Mathematical Analysis and Applications*, 30(1):56–66.

Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. SIAM.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–41.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Stanford University, Stanford, CA.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*. Johns Hopkins University Press, fourth edition.

Golub, G. H. and Ye, Q. (2002). An Inverse Free Preconditioned Krylov Subspace Method for Symmetric Generalized Eigenvalue Problems.

Green, R. C. and Hollifield, B. (1992). When Will Mean-Variance Efficient Portfolios Be Well Diversified ? *The journal of finance*, 47(5):1785–1809.

Hsieh, C.-j. (2014). QUIC : Quadratic Approximation for Sparse Inverse Covariance Estimation. *Journal of Machine Learning Research*, 15:2911–2947.

Jagannathan, R. and Ma, T. (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *Journal of Finance*, 58(4):1651–1683.

Khare, K., Oh, S.-Y., and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, page to appear.

Lam, C. and Fan, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *Annals of statistics*, 37(6B):4254–4278.

Lanckriet, G. R., El Ghaoui, L., Bhattacharyya, C., and Jordan, M. I. (2002). A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Lehoucq, R. B. and Sorensen, D. C. (1996). Deflation Techniques for an Implicitly Restarted Arnoldi Iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821.

Liu, H., Wang, L., and Zhao, T. (2014). Sparse Covariance Matrix Estimation With Eigenvalue Constraints. *Journal of Computational and Graphical Statistics*, 23(2):439–459.

Luenberger, D. G. (2013). *Investment Science*. Oxford University Press, second edition.

Marcenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1):77–91.

Mazumder, R. and Hastie, T. (2012a). Exact covariance thresholding into connected components for large-scale Graphical Lasso. *Journal of Machine Learning Research*, 13:781–794.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6(August):2125–2149.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462.

Merton, R. C. (1980). On estimating the expected return on the market. *Journal of Financial Economics*, 8:323–361.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, 104(486):735–746.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing 1 -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(January 2010):935–980.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(January):494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Sorensen, D. C. (1990). Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13(1):357–385.

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multi- Variate Normal Distribution. In *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 197–206, Berkeley, CA. University of California Press.

Sun, Y., Todorovic, S., and Goodison, S. (2010). Local-learning-based feature selection for high-dimensional data analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1610–26.

Tsanas, A., Little, M. a., Fox, C., and Ramig, L. O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New Insights and Faster Computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2013). Condition Number Regularized Covariance Estimation. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(3):427–450.

Xue, L., Ma, S., and Zou, H. (2012). Positive-Definite 1 -Penalized Estimation of Large Covariance Matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

Yuan, M. (2008). Regularized Estimates in Gaussian Graphical Models. *Journal of Computational and Graphical Statistics*, 17(4):809–826.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35.

Appendix

A. Further simulation results

Table A.1: All result of simulations for empirical risk comparison of the universal SCAD thresholding estimator and its LSPD corrections. Each column means the norm used to measure risk. The abbreviaions for the estimators and covariance models $\mathbf{M}_i$'s are introduced in Section 4.2..

	Multivariate normal			Multivariate t		
	Matrix l_1	Spectral	Frobenius	Matrix l_1	Spectral	Frobenius
$\mathbf{M}_1$, $p = 100$						
SCAD thres.	5.54 (0.11)	3.02 (0.06)	6.40 (0.07)	9.27 (0.40)	4.91 (0.17)	9.76 (0.24)
LSPD(μ_{SF})	5.33 (0.10)	2.94 (0.06)	7.05 (0.06)	8.82 (0.37)	4.75 (0.16)	10.23 (0.24)
LSPD(∞)	5.76 (0.12)	3.14 (0.07)	7.51 (0.08)	9.49 (0.42)	4.97 (0.19)	10.54 (0.27)
$\mathbf{M}_1$, $p = 200$						
SCAD thres.	6.42 (0.10)	3.45 (0.05)	9.82 (0.06)	13.87 (0.80)	6.28 (0.24)	16.31 (0.50)
LSPD(μ_{SF})	6.17 (0.08)	3.40 (0.04)	10.96 (0.05)	13.02 (0.72)	6.02 (0.22)	17.28 (0.51)
LSPD(∞)	6.71 (0.10)	3.51 (0.05)	11.59 (0.07)	14.28 (0.82)	6.42 (0.27)	17.75 (0.54)
$\mathbf{M}_1$, $p = 400$						
SCAD thres.	7.98 (0.12)	4.01 (0.05)	15.29 (0.06)	19.65 (1.03)	7.81 (0.33)	25.05 (0.57)
LSPD(μ_{SF})	7.40 (0.10)	3.92 (0.04)	17.09 (0.06)	18.28 (0.93)	7.45 (0.31)	26.84 (0.64)
LSPD(∞)	8.38 (0.13)	4.00 (0.06)	17.93 (0.08)	20.19 (1.05)	8.04 (0.37)	27.46 (0.71)
$\mathbf{M}_2$, $p = 100$						
SCAD thres.	3.19 (0.04)	1.64 (0.02)	5.13 (0.02)	4.71 (0.11)	2.33 (0.03)	7.78 (0.11)
LSPD(μ_{SF})	3.18 (0.04)	1.64 (0.02)	5.12 (0.02)	4.69 (0.11)	2.32 (0.03)	7.77 (0.11)
LSPD(∞)	3.19 (0.04)	1.63 (0.02)	5.13 (0.02)	4.72 (0.11)	2.32 (0.03)	7.77 (0.11)
$\mathbf{M}_2$, $p = 200$						
SCAD thres.	5.97 (0.06)	2.88 (0.03)	9.74 (0.04)	8.87 (0.19)	4.07 (0.04)	15.03 (0.23)
LSPD(μ_{SF})	5.92 (0.06)	2.84 (0.03)	9.84 (0.04)	8.76 (0.18)	4.04 (0.04)	15.05 (0.23)
LSPD(∞)	5.94 (0.06)	2.76 (0.03)	9.94 (0.04)	8.90 (0.19)	4.03 (0.05)	15.05 (0.23)
$\mathbf{M}_2$, $p = 400$						
SCAD thres.	11.59 (0.11)	5.49 (0.06)	19.08 (0.07)	18.55 (0.47)	7.57 (0.06)	28.82 (0.37)
LSPD(μ_{SF})	11.56 (0.10)	5.37 (0.05)	19.91 (0.07)	17.81 (0.39)	7.38 (0.06)	29.38 (0.37)
LSPD(∞)	11.64 (0.11)	5.17 (0.06)	20.31 (0.07)	18.73 (0.49)	7.39 (0.08)	29.65 (0.38)
$\mathbf{M}_3$, $p = 100$						
SCAD thres.	12.03 (0.06)	6.88 (0.03)	18.50 (0.02)	13.11 (0.08)	7.64 (0.04)	20.57 (0.06)
LSPD(μ_{SF})	12.03 (0.06)	6.88 (0.03)	18.50 (0.02)	13.11 (0.08)	7.64 (0.04)	20.57 (0.06)
LSPD(∞)	12.03 (0.06)	6.88 (0.03)	18.50 (0.02)	13.11 (0.08)	7.64 (0.04)	20.57 (0.06)
$\mathbf{M}_3$, $p = 200$						
SCAD thres.	12.97 (0.05)	7.64 (0.02)	27.99 (0.02)	14.48 (0.08)	8.40 (0.03)	32.39 (0.14)
LSPD(μ_{SF})	12.97 (0.05)	7.64 (0.02)	27.99 (0.02)	14.44 (0.07)	8.39 (0.03)	32.29 (0.12)
LSPD(∞)	12.97 (0.05)	7.64 (0.02)	27.99 (0.02)	14.47 (0.08)	8.37 (0.03)	32.25 (0.12)
$\mathbf{M}_3$, $p = 400$						
SCAD thres.	14.31 (0.05)	8.06 (0.01)	41.07 (0.03)	15.81 (0.07)	8.81 (0.02)	48.17 (0.18)
LSPD(μ_{SF})	14.31 (0.05)	8.06 (0.01)	41.07 (0.03)	15.74 (0.06)	8.80 (0.02)	47.91 (0.16)
LSPD(∞)	14.31 (0.05)	8.06 (0.01)	41.07 (0.03)	15.74 (0.06)	8.74 (0.02)	47.56 (0.16)

Table A.2: (Matrix l_1 norm) Empirical risks, measured on the restriction to three partition of the true covariance matrix. The column 'Diagonal' indicates the error measured only on diagonal elements, and 'Supp. (Non-Supp., respectively) off-diag' refers to the error measured on the support (the zero set) of the off-diagonal part of the true $\mathbf{\Sigma}$. The simulation settings and abbreviations follows Section 4.2..

	Multivariate normal			Multivariate t		
	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.
			$\mathbf{M}_1$, $p = 400$			
Soft thres.	0.44 (0.01)	6.00 (0.04)	5.05 (0.10)	0.79 (0.04)	7.77 (0.15)	14.63 (0.86)
LSPD(μ_{SF})	0.50 (0.01)	6.06 (0.04)	4.95 (0.10)	0.95 (0.06)	7.71 (0.11)	13.93 (0.78)
LSPD(∞)	0.53 (0.01)	6.00 (0.04)	5.05 (0.10)	1.04 (0.07)	7.77 (0.15)	14.63 (0.86)
Eig. con.	0.45 (0.01)	6.00 (0.04)	5.02 (0.10)	0.83 (0.05)	7.75 (0.15)	14.18 (0.79)
log-det	0.47 (0.01)	6.02 (0.04)	4.99 (0.10)	0.85 (0.05)	7.77 (0.15)	14.15 (0.79)
SCAD thres.	0.44 (0.01)	5.57 (0.06)	4.41 (0.09)	0.79 (0.04)	8.20 (0.23)	13.28 (0.72)
LSPD(μ_{SF})	0.75 (0.01)	5.80 (0.05)	4.07 (0.08)	1.13 (0.05)	8.01 (0.18)	12.26 (0.65)
LSPD(∞)	0.90 (0.01)	5.57 (0.06)	4.41 (0.09)	1.30 (0.06)	8.20 (0.23)	13.28 (0.72)
			$\mathbf{M}_2$, $p = 400$			
Soft thres.	0.46 (0.01)	10.30 (0.11)	6.01 (0.10)	0.83 (0.04)	14.09 (0.13)	13.90 (0.66)
LSPD(μ_{SF})	0.49 (0.01)	10.36 (0.11)	5.94 (0.10)	0.94 (0.04)	14.16 (0.12)	13.25 (0.61)
LSPD(∞)	0.51 (0.01)	10.30 (0.11)	6.01 (0.10)	1.02 (0.05)	14.09 (0.13)	13.90 (0.66)
Eig. con.	0.46 (0.01)	10.30 (0.11)	5.99 (0.10)	0.89 (0.04)	14.09 (0.13)	13.48 (0.62)
log-det	0.48 (0.01)	10.33 (0.11)	5.91 (0.10)	0.90 (0.04)	14.11 (0.13)	13.42 (0.62)
SCAD thres.	0.46 (0.01)	10.03 (0.13)	5.76 (0.10)	0.83 (0.04)	14.52 (0.12)	11.42 (0.51)
LSPD(μ_{SF})	0.69 (0.01)	10.35 (0.12)	5.41 (0.09)	1.00 (0.04)	14.62 (0.11)	10.64 (0.47)
LSPD(∞)	0.80 (0.01)	10.03 (0.13)	5.76 (0.10)	1.13 (0.04)	14.52 (0.12)	11.42 (0.51)
			$\mathbf{M}_3$, $p = 400$			
Soft thres.	2.27 (0.01)	12.39 (0.03)	3.14 (0.04)	3.44 (0.04)	12.54 (0.03)	3.88 (0.10)
LSPD(μ_{SF})	2.27 (0.01)	12.39 (0.03)	3.14 (0.04)	3.37 (0.03)	12.55 (0.03)	3.75 (0.09)
LSPD(∞)	2.27 (0.01)	12.39 (0.03)	3.14 (0.04)	3.37 (0.03)	12.54 (0.03)	3.88 (0.10)
Eig. con.	2.27 (0.02)	12.39 (0.03)	3.14 (0.04)	3.43 (0.04)	12.54 (0.03)	3.88 (0.10)
SCAD thres.	2.09 (0.02)	12.49 (0.03)	2.61 (0.04)	3.39 (0.04)	12.48 (0.03)	4.41 (0.11)
LSPD(μ_{SF})	2.09 (0.02)	12.49 (0.03)	2.61 (0.04)	3.32 (0.04)	12.49 (0.03)	4.29 (0.10)
LSPD(∞)	2.09 (0.02)	12.49 (0.03)	2.61 (0.04)	3.32 (0.04)	12.48 (0.03)	4.41 (0.11)

Table A.3: (Spectral norm) Empirical risks, measured on the restriction to three partition of the true covariance matrix. The column 'Diagonal' indicates the error measured only on diagonal elements, and 'Supp. (Non-Supp., respectively) off-diag' refers to the error measured on the support (the zero set) of the off-diagonal part of the true $\mathbf{\Sigma}$. The simulation settings and abbreviations follows Section 4.2..

	Multivariate normal			Multivariate t		
	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.
$\mathbf{M}_1$, $p = 400$						
Soft thres.	0.44 (0.01)	4.42 (0.03)	1.90 (0.03)	0.79 (0.04)	5.96 (0.11)	4.89 (0.26)
LSPD(μ_{SF})	0.50 (0.01)	4.50 (0.03)	1.86 (0.03)	0.95 (0.06)	6.02 (0.09)	4.67 (0.23)
LSPD(∞)	0.53 (0.01)	4.42 (0.03)	1.90 (0.03)	1.04 (0.07)	5.96 (0.11)	4.89 (0.26)
Eig. con.	0.45 (0.01)	4.42 (0.03)	1.89 (0.03)	0.83 (0.05)	5.97 (0.11)	4.84 (0.25)
log-det	0.47 (0.01)	4.45 (0.03)	1.88 (0.03)	0.85 (0.05)	5.99 (0.11)	4.83 (0.25)
SCAD thres.	0.44 (0.01)	3.66 (0.04)	1.69 (0.03)	0.79 (0.04)	5.99 (0.18)	4.47 (0.22)
LSPD(μ_{SF})	0.75 (0.01)	3.97 (0.04)	1.56 (0.03)	1.13 (0.05)	5.99 (0.14)	4.13 (0.20)
LSPD(∞)	0.90 (0.01)	3.66 (0.04)	1.69 (0.03)	1.30 (0.06)	5.99 (0.18)	4.47 (0.22)
$\mathbf{M}_2$, $p = 400$						
Soft thres.	0.46 (0.01)	5.40 (0.05)	1.82 (0.02)	0.83 (0.04)	7.14 (0.07)	4.11 (0.18)
LSPD(μ_{SF})	0.49 (0.01)	5.42 (0.05)	1.80 (0.02)	0.94 (0.04)	7.18 (0.07)	3.92 (0.17)
LSPD(∞)	0.51 (0.01)	5.40 (0.05)	1.82 (0.02)	1.02 (0.05)	7.14 (0.07)	4.11 (0.18)
Eig. con.	0.46 (0.01)	5.40 (0.05)	1.82 (0.02)	0.89 (0.04)	7.14 (0.07)	4.07 (0.18)
log-det	0.48 (0.01)	5.41 (0.05)	1.79 (0.02)	0.90 (0.04)	7.15 (0.07)	4.05 (0.18)
SCAD thres.	0.46 (0.01)	5.28 (0.06)	1.74 (0.02)	0.83 (0.04)	7.37 (0.07)	3.34 (0.14)
LSPD(μ_{SF})	0.69 (0.01)	5.43 (0.05)	1.63 (0.02)	1.00 (0.04)	7.43 (0.06)	3.12 (0.13)
LSPD(∞)	0.80 (0.01)	5.28 (0.06)	1.74 (0.02)	1.13 (0.04)	7.37 (0.07)	3.34 (0.14)
$\mathbf{M}_3$, $p = 400$						
Soft thres.	2.27 (0.01)	7.36 (0.00)	1.33 (0.01)	3.44 (0.04)	7.47 (0.00)	1.65 (0.04)
LSPD(μ_{SF})	2.27 (0.01)	7.36 (0.00)	1.33 (0.01)	3.37 (0.03)	7.48 (0.00)	1.59 (0.03)
LSPD(∞)	2.27 (0.01)	7.36 (0.00)	1.33 (0.01)	3.37 (0.03)	7.47 (0.00)	1.65 (0.04)
Eig. con.	2.27 (0.01)	7.36 (0.00)	1.33 (0.01)	3.43 (0.04)	7.47 (0.00)	1.65 (0.04)
SCAD thres.	2.09 (0.02)	7.44 (0.00)	1.17 (0.01)	3.39 (0.04)	7.43 (0.01)	1.84 (0.04)
LSPD(μ_{SF})	2.09 (0.02)	7.44 (0.00)	1.17 (0.01)	3.32 (0.04)	7.44 (0.01)	1.79 (0.04)
LSPD(∞)	2.09 (0.02)	7.44 (0.00)	1.17 (0.01)	3.32 (0.04)	7.43 (0.01)	1.84 (0.04)

Table A.4: (Frobenius norm) Empirical risks, measured on the restriction to three partition of the true covariance matrix. The column ‘Diagonal’ indicates the error measured only on diagonal elements, and ‘Supp. (Non-Supp., respectively) off-diag’ refers to the error measured on the support (the zero set) of the off-diagonal part of the true $\mathbf{\Sigma}$. The simulation settings and abbreviations follows Section 4.2..

	Multivariate normal			Multivariate t		
	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.	Diagonal	Supp. of off-diag.	Non-Supp. of off-diag.
			$\mathbf{M}_1$, $p = 400$			
Soft thres.	1.67 (0.14)	16.28 (0.08)	6.44 (0.07)	2.12 (0.34)	23.63 (0.34)	9.50 (0.26)
LSPD(μ_{SF})	1.77 (0.20)	16.69 (0.07)	6.32 (0.07)	2.52 (0.71)	24.20 (0.34)	9.10 (0.23)
LSPD(∞)	1.84 (0.23)	16.28 (0.08)	6.44 (0.07)	2.69 (0.78)	23.63 (0.34)	9.50 (0.26)
Eig. con.	1.67 (0.15)	16.29 (0.08)	6.39 (0.07)	2.13 (0.35)	23.65 (0.34)	9.39 (0.25)
log-det	1.70 (0.16)	16.41 (0.08)	6.31 (0.07)	2.12 (0.36)	23.75 (0.34)	9.32 (0.25)
SCAD thres.	1.67 (0.14)	13.91 (0.07)	5.66 (0.06)	2.12 (0.34)	22.90 (0.55)	8.82 (0.24)
LSPD(μ_{SF})	2.73 (0.30)	14.42 (0.07)	5.21 (0.06)	3.05 (0.71)	23.46 (0.53)	8.16 (0.22)
LSPD(∞)	3.12 (0.33)	13.91 (0.07)	5.66 (0.06)	3.39 (0.76)	22.90 (0.55)	8.82 (0.24)
			$\mathbf{M}_2$, $p = 400$			
Soft thres.	1.68 (0.11)	17.40 (0.08)	7.78 (0.05)	2.12 (0.33)	25.85 (0.40)	8.94 (0.27)
LSPD(μ_{SF})	1.71 (0.14)	17.54 (0.08)	7.70 (0.05)	2.33 (0.48)	26.11 (0.40)	8.56 (0.25)
LSPD(∞)	1.74 (0.16)	17.40 (0.08)	7.78 (0.05)	2.48 (0.55)	25.85 (0.40)	8.94 (0.27)
Eig. con.	1.69 (0.11)	17.40 (0.08)	7.77 (0.05)	2.13 (0.33)	25.85 (0.40)	8.86 (0.27)
log-det	1.69 (0.12)	17.41 (0.08)	7.62 (0.05)	2.13 (0.34)	25.86 (0.40)	8.75 (0.26)
SCAD thres.	1.68 (0.11)	17.31 (0.09)	7.48 (0.05)	2.12 (0.33)	27.35 (0.39)	7.37 (0.22)
LSPD(μ_{SF})	2.42 (0.29)	17.65 (0.09)	7.03 (0.05)	2.51 (0.50)	27.66 (0.38)	6.88 (0.20)
LSPD(∞)	2.72 (0.34)	17.31 (0.09)	7.48 (0.05)	2.79 (0.57)	27.35 (0.39)	7.37 (0.22)
			$\mathbf{M}_3$, $p = 400$			
Soft thres.	25.87 (1.68)	36.78 (0.01)	9.99 (0.06)	35.46 (4.48)	37.18 (0.01)	11.34 (0.20)
LSPD(μ_{SF})	25.87 (1.68)	36.78 (0.01)	9.99 (0.06)	35.18 (4.51)	37.20 (0.01)	10.98 (0.17)
LSPD(∞)	25.87 (1.68)	36.78 (0.01)	9.99 (0.06)	34.12 (4.96)	37.18 (0.01)	11.34 (0.20)
Eig. con.	25.87 (1.68)	36.78 (0.01)	9.99 (0.06)	35.45 (4.48)	37.18 (0.01)	11.33 (0.20)
SCAD thres.	15.53 (1.45)	37.04 (0.02)	8.52 (0.06)	28.10 (4.15)	37.04 (0.02)	12.56 (0.22)
LSPD(μ_{SF})	15.53 (1.45)	37.04 (0.02)	8.52 (0.06)	27.76 (4.06)	37.05 (0.02)	12.26 (0.19)
LSPD(∞)	15.53 (1.45)	37.04 (0.02)	8.52 (0.06)	27.02 (4.09)	37.04 (0.02)	12.56 (0.22)

국문초록

선형 축소를 통한
공분산 행렬 추정량의 양정치 보정
Positive-definite correction
of covariance matrix estimators
via linear shrinkage

본고에서는 공분산 행렬 추정에서 발생하는 양정치성 문제를 논한다. 흔히 쓰이는 표본 공분산 행렬은 자료가 고차원인 경우 모집단의 공분산 행렬 (참 공분산 행렬)을 잘 추정하지 못하는 점이 알려져 있다. 최근에는 표본 공분산 행렬 대신, 참 공분산 행렬에 구조적 가정(이를테면 희소성)을 부여하고 정규화된 추정량들이 제안되어 왔다. 이 정규화된 추정량들은 가정된 구조를 추정할 때 점근적으로 일치하거나 비율-최적성을 가짐이 알려져 있다. 그러나 제안된 추정량 중 다수가 추정량의 양정치성을 설명하고 있지 않으며 이는 비-양정치 추정치 문제를 야기한다. 혹은 양정치성을 설명하기 위한 부가적인 정규화(혹은 제약 조건)가 고유치에 가해지게 되어 점근적 분석이나 계산을 더 어렵게 만든다. 본고에서는, 간단한 일단계 업데이트 절차를 제안하여 유한 표본에서 양정치성을 보장하지 않는 추정량들의 양정치성을 추가적으로 보장하고자 한다. 선형 축소 (Stein, 1956; Ledoit and Wolf, 2004) 논의에 착안하여, 여기서는 첫단계 공분산 행렬 추정량 (양정치성이 보장되지 않은 정규화된 공분산 행렬 추정량)과 대각행렬의 볼록 결합을 취할 것을 제안한다. 만약 볼록 결합의 계수가 잘 선택된다면, 이 일단계 보정 절차는 - LSPD라 명명하였다 - 첫단계 추정량의 점근적 성질을 보존함을 논증할 수 있다. 부가적으로 LSPD 절차는 기존에 제안된 양정치 정규화 추정량들과 달리 (Rothman, 2012; Xue et al., 2012) 닫힌 형태의 수식으로 표현이 가능하여 그 계산이 수치-

최적화 문제 해결을 필요로 하지 않는다. 본고에서는 수치계산을 통하여 다른 양정치 정규화 추정량들과 LSPD 추정량의 유한 표본 성질을 비교하며 LSPD 추정량의 계산속도 향상에 대하여도 논한다. 마지막으로 공분산 행렬 추정량에 의존하는 다변량 통계 절차들에 - 선형 최소최대 판별 문제 (Lanckriet et al., 2002) 및 포트폴리오 최적화 문제 - 본 방법론이 적용되는 경우 각 절차의 성능이 향상됨을 실제 자료로부터 예증한다.

주요어 : 공분산 행렬; 양정치성; 고차원 추정; 선형 축소.
학 번 : 2010-23063