Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage

Young-Geun Choi[a,1], Johan Lim[b,*], Anindya Roy[c], Junyong Park[c]

[a]*Data R&D Center, SK Telecom, 65 Eulji-ro, Jung-Gu, Seoul 04539, Rep. of Korea*
[b]*Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Rep. of Korea*
[c]*Department of Mathematics and Statistics, University of Maryland Baltimore County, 1100 Hilltop Circle, Baltimore, MD 21250, United States*

Abstract

In this work, we study the positive definiteness (PDness) problem in covariance matrix estimation. For high dimensional data, many regularized estimators are proposed under structural assumptions on the true covariance matrix 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. To achieve the PDness, researchers consider additional regularizations (or constraints) on eigenvalues, which make both the asymptotic analysis and computation much harder. In this paper, we propose a simple modification of the regularized covariance matrix estimator to make it PD while preserving the support. We revisit the idea of linear shrinkage and propose to take a convex combination between the first-stage estimator (the regularized covariance matrix without PDness) and a given form of diagonal matrix. The proposed modification, which we denote as FSPD (Fixed Support and Positive Definiteness) estimator, is shown to preserve the asymptotic properties of the first-stage estimator, if the shrinkage parameters are carefully selected. It has a closed form expression and its computation is optimization-free, unlike existing PD sparse estimators. In addition, the FSPD is generic in the sense that it can be applied to any non-PD matrix including the precision matrix. The FSPD estimator is numerically compared with other sparse PD estimators to understand its finite sample properties as well as its computational gain. It is also applied to two multivariate procedures relying on the covariance matrix estimator — the linear minimax classification problem and the Markowitz portfolio optimization problem — and is shown to substantially improve the performance of both procedures.

Keywords: Covariance matrix, high dimensional estimation, linear minimax classification problem, linear shrinkage, portfolio optimization, precision matrix, positive definiteness
2010 MSC: Primary 62H12, Secondary 62F12

1. Introduction

Covariance matrix and its consistent estimation are involved in many multivariate statistical procedures, where the sample covariance matrix is popularly used. In recent, high dimensional data are prevalent everywhere for which the sample covariance matrix is known to be inconsistent [28]. To resolve the difficulty from high dimensionality, regularized procedures or estimators are 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 assumptions [1, 2, 7, 9–11, 33] or penalizes the Gaussian likelihood [3, 22]. The asymptotic theory of the regularized estimators are well understood and, particularly, they are shown to be consistent in estimating the true covariance matrix and its support (the positions of its non-zero elements).

The main interest of this paper is positive definiteness (PDness) of covariance matrix estimator. The PDness is an essential property for the validity of many multivariate statistical procedures. However, the regularized covariance

---

matrix estimators recently studied are often not PD in finite sample. This is because they more focus on the given structural assumptions and do not impose the PDness on their estimators. For example, the banding or thresholding method [2, 33] regularizes the sample covariance matrix in an elementwise manner and provides an explicit form of the estimator that satisfies the given assumptions. Nonetheless, the eigenstructure of the resulting estimator is completely unknown and, without doubt, the resulting covariance matrix estimate is not necessarily PD. A few efforts are made to find an estimator which attains both the sparsity and PDness by *incorporating* them in a single optimization problem [3, 22, 27, 32, 39]. In particular, the works by Rothman [32], Xue et al. [39], and Liu et al. [27] understand the soft thresholding of the sample covariance (correlation) matrix as a convex minimization problem and add a convex penalty (or constraint) to the problem in order to guarantee the PDness of solution. However, we remark that each of these incorporating approaches has to be customized to a certain regularization technique (e.g., soft thresholding estimator) and also requires us to solve a large-scale optimization problem.

Instead of simultaneously handling PDness and regularization, we propose a *separated* update of a given covariance matrix estimator. Our motivation is that the regularized estimators in the literature are already proven to be "good" in terms of consistency or rate-optimality in estimating their true counterparts. Thus, we aim to minorly modify them to retain the same asymptotic properties as well as be PD with the same support. To be specific, denote by $\widehat{\boldsymbol{\Sigma}}$ a given covariance matrix estimator. We consider a distance minimization problem

$$\underset{\widehat{\boldsymbol{\Sigma}}^*}{\text{minimize}} \left\{ \left\| \widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}} \right\| \ : \ \gamma_1(\widehat{\boldsymbol{\Sigma}}^*) \geq \epsilon, \ \text{supp}(\widehat{\boldsymbol{\Sigma}}^*) = \text{supp}(\widehat{\boldsymbol{\Sigma}}), \ \widehat{\boldsymbol{\Sigma}}^* = (\widehat{\boldsymbol{\Sigma}}^*)^\top \right\} \tag{1}$$

where $\|\cdot\|$ is a matrix norm, $\epsilon > 0$ is a pre-determined small constant and $\gamma_1(\widehat{\boldsymbol{\Sigma}}^*)$ denotes the smallest eigenvalue of $\widehat{\boldsymbol{\Sigma}}^*$. In solving (1), to make the modification simple, we further restrict the class of $\widehat{\boldsymbol{\Sigma}}^*$ to a family of linear shrinkage of $\widehat{\boldsymbol{\Sigma}}$ to the identity matrix that is

$$\widehat{\boldsymbol{\Sigma}}^* \in \left\{ \boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}) \equiv \alpha\widehat{\boldsymbol{\Sigma}} + (1-\alpha)\mu\mathbf{I} \ : \ \alpha \in [0,1], \ \mu \in \mathbb{R} \right\}. \tag{2}$$

The primary motivation for considering (2) is that shrinking $\widehat{\boldsymbol{\Sigma}}$ linearly to the identity enables us to handle the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^*$ easily while preserving the support of $\widehat{\boldsymbol{\Sigma}}$. We will reserve the term *fixed support PD* (FSPD) estimator to describe any estimator of the form (2) that solves the minimization problem (1), and refer to the process of modification as an FSPD procedure. The proposed FSPD estimator/procedure has several interesting and important properties. First, the calculation of $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ is optimization-free, since the choice of μ and α can be explicitly expressible with the smallest and largest eigenvalues of the initial estimator $\widehat{\boldsymbol{\Sigma}}$. Second, for suitable choices of μ and α in (2), the estimators $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ and $\widehat{\boldsymbol{\Sigma}}$ have the same rate of convergence to the true covariance matrix $\boldsymbol{\Sigma}$ under some conditions. Third, owing to the separating nature, the FSPD procedure is equally applicable to any (possibly) non-PD estimator of covariance matrix as well as precision matrix (the inverse of covariance matrix).

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 guarantee the PDness are also briefly reviewed. Our main results are presented in Section 3. The FSPD procedure is developed by solving the restricted distance minimization presented in (1) and (2). We not only derive statistical convergence rate of the resulting FSPD estimator, but also discuss some implementation issues for its practical use. In Section 4, we numerically show that FSPD estimator has comparable performances with recent PD sparse covariance matrix estimators introduced in Section 2, whereas ours are computationally much simpler and faster. In Section 5, we illustrate the usefulness of FSPD-updated regularized covariance estimators in two statistical procedures: the linear minimax classification problem [23] and Markowitz portfolio optimization with no short-sale [19]. Since the FSPD procedure is applicable to any covariance matrix estimators including precision 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 $p \times p$. Let $\boldsymbol{\Sigma}$, $\mathbf{S}$ and $\widehat{\boldsymbol{\Sigma}}$ be the true covariance matrix, the sample covariance matrix, and a generic covariance matrix estimator, respectively. For a symmetric matrix $\mathbf{A}$, the ordered eigenvalue will be denoted by $\gamma_1(\mathbf{A}) \leq \ldots \leq \gamma_p(\mathbf{A})$. In particular, we abbreviate $\gamma_i(\boldsymbol{\Sigma})$ to γ_i and $\gamma_i(\widehat{\boldsymbol{\Sigma}})$ to $\widehat{\gamma}_i$. The Frobenius norm of $\mathbf{A}$ is defined with scaling by $\|\mathbf{A}\|_F := \sqrt{\text{tr}(\mathbf{A}^\top\mathbf{A})/p}$. The spectral norm of $\mathbf{A}$ is $\|\mathbf{A}\|_2 := \sqrt{\gamma_p(\mathbf{A}^\top\mathbf{A})} \equiv \max_i |\gamma_i(\mathbf{A})|$. The norm without subscription, $\|\cdot\|$, will be used in the case both the spectral and Frobenius norm are applicable.

2

2. Covariance regularization and PDness

2.1. Non-PDness of regularized covariance matrix estimators

In this Section, we briefly review two most common regularized covariance matrix estimators and discuss their PDness.

2.1.1. Thresholding estimators

The thresholded sample covariance matrix (simply *thresholding estimator*) regards small elements in the sample covariance matrix as noise and set them to zero. For a fixed λ, the estimator is defined by

$$\widehat{\mathbf{\Sigma}}_\lambda^{\text{Thr}} := \left[T_\lambda(s_{ij}), 1 \le i, j \le p \right], \tag{3}$$

where $T_\lambda(\cdot) : \mathbb{R} \to \mathbb{R}$ is a thresholding function [33]. Some examples of the thresholding function are (i) (hard thresholding) $T_\lambda(s) = I(|s| \ge \lambda) \cdot s$, (ii) (soft thresholding) $T_\lambda(s) = \text{sign}(s) \cdot (|s| - \lambda)_+$, and (iii) (SCAD thresholding) $T_\lambda(s) = I(|s| < 2\lambda) \cdot \text{sign}(s) \cdot (|s| - \lambda)_+ + I(2\lambda < |s| \le a\lambda) \cdot \{(a-1)s - \text{sign}(s)a\lambda\}/(a-2) + I(|s| > a\lambda) \cdot s$ with $a > 2$. One may threshold only the off-diagonal elements of $\mathbf{S}$, in which case the estimator remains asymptotically equivalent to (3). The universal threshold λ in (3) can be adapted to each element; Cai and Liu [5] suggests an *adaptive thresholding estimator* that thresholds each s_{ij} with an element-adaptive threshold λ_{ij}. In Section 5, we will use an adaptive thresholding estimator with the soft thresholding function.

As we point out earlier, the elementwise manipulation in the thresholding estimators does not retain its PDness. The left panel of Figure 1 plots the minimum eigenvalues of thresholding estimators with various thresholding functions and tuning parameters, when the dataset of $n = 100$ and $p = 400$ are sampled from the multivariate t-distribution with 5 degrees of freedom and true covariance matrix $\mathbf{M}_1$ defined in Section 4. As shown in the figure, $\widehat{\mathbf{\Sigma}}_\lambda^{\text{Thr}}$ is not PD if λ is moderately small or selected via the five-fold cross-validation (CV) explained in Bickel and Levina [1].

2.1.2. Banding estimators

The banded covariance matrix arises when the variables of data are ordered and serially dependent as in the data of time series, climatology, or spectroscopy. Bickel and Levina [2] proposes a class of *banding estimators*

$$\widehat{\mathbf{\Sigma}}_h^{\text{Band}} = \left[s_{ij} \cdot w_{|i-j|} \right],$$

where banding weight w_m $(m = 0, \dots, p - 1)$ is proposed by $w_m = I(m \le h)$ in the paper for a fixed bandwidth h. Roughly speaking, $\widehat{\mathbf{\Sigma}}_h^{\text{Band}}$ discards the sample covariance s_{ij} if corresponding indices i and j are distant. Cai et al. [10] considers a *tapering estimator* which smooths the banding weight in $\widehat{\mathbf{\Sigma}}_h^{\text{Band}}$ as

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

We could see that both banding estimators are again from elementwise operations on the sample covariance matrix and, as in the thresholding estimators, do not retain the PDness of the sample estimator. The right panel of Figure 1 tells that this is indeed. The panel is based on the same simulated data set for the left panel and shows that the two estimators result in non-PD estimates regardless of the size of bandwidth.

2.2. Adding PDness to structural (sparse) regularization

The PDness of regularized covariance matrix estimator recently get attention of researchers and a limited number of works are done in the literature; see [27, 32, 39]. The three approaches, namely PD sparse estimators, are all based on the fact that the soft thresholding estimator (the thresholding estimator equipped with the soft thresholding function) can be obtained by minimizing the following convex function:

$$\widehat{\mathbf{\Sigma}}^{\text{Soft}}(\lambda) = \underset{\mathbf{\Sigma}}{\text{argmin}} \; \|\mathbf{\Sigma} - \mathbf{S}\|_{\text{F}}^2 + \lambda \sum_{i<j} |\sigma_{ij}|. \tag{4}$$
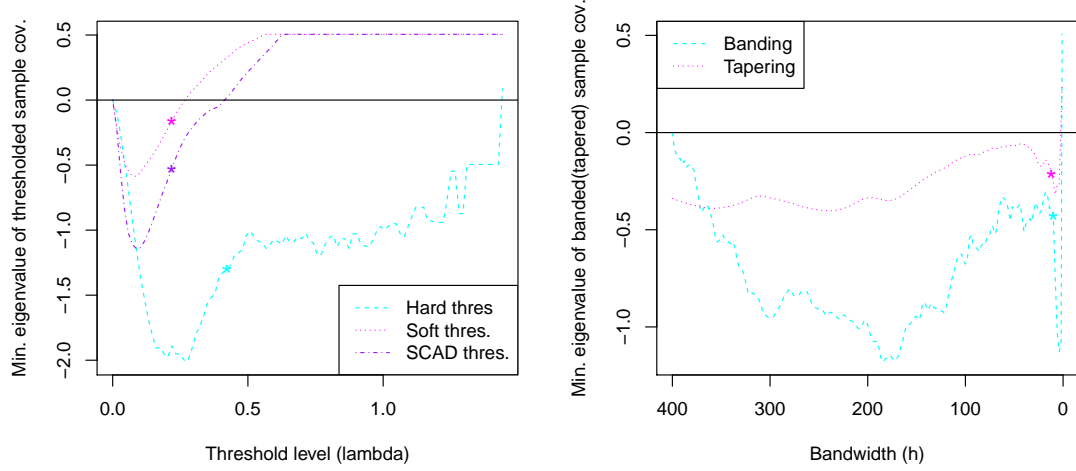
3

Figure 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 a five-fold cross-validation (CV). Note that all the CV-selected estimates are not PD.

This expression coincides (3) when $T_\lambda(s) = \text{sign}(s) \cdot (|s| - \lambda)$ with the diagonal elements not being regularized. The PD sparse estimators add a constraint or penalty function to (4) to encourage the solution to be PD. For example, Rothman [32] considers to add a log-determinant penalty:

$$\widehat{\boldsymbol{\Sigma}}^{\text{logdet}}(\lambda) := \underset{\boldsymbol{\Sigma}}{\text{argmin}} \ \|\boldsymbol{\Sigma} - \mathbf{S}\|_{\text{F}}^2 - \tau \log \det(\boldsymbol{\Sigma}) + \lambda \sum_{i<j} |\sigma_{ij}|, \tag{5}$$

where $\tau > 0$ is fixed a small value. The additional term behaves a convex barrier that naturally ensures the PDness of the solution and preserves the convexity the objective function. In the paper, the author solves the normal equations with respect to $\boldsymbol{\Sigma}$, where column vectors of the current estimate $\boldsymbol{\Sigma}$ are alternatingly updated by solving ($p - 1$)-variate lasso regressions. We note that, even each lasso regression can be calculated fast, repeating it for every column leads to $O(p^3)$ flop computations for the entire matrix to be updated. On the other hand, Xue et al. [39] proposes to solve

$$\widehat{\boldsymbol{\Sigma}}^{\text{EigCon}}(\lambda) := \underset{\boldsymbol{\Sigma} \, : \, \boldsymbol{\Sigma} \succeq \epsilon \mathbf{I}}{\text{argmin}} \ \|\boldsymbol{\Sigma} - \mathbf{S}\|_{\text{F}}^2 + \lambda \sum_{i<j} |\sigma_{ij}|, \tag{6}$$

where $\boldsymbol{\Sigma} \succeq \epsilon \mathbf{I}$ means that $\boldsymbol{\Sigma} - \epsilon \mathbf{I}$ is positive semidefinite. Since the additional constraint $\{\boldsymbol{\Sigma} \, : \, \boldsymbol{\Sigma} \succeq \epsilon \mathbf{I}\}$ is a convex set, (6) still has a global optimum. The objective function is optimized via an alternating direction method of multipliers (ADMM) algorithm [4], in which each iteration contains the eigenvalue decomposition and thresholding of a p by p matrix. The computational cost of the algorithm is, mainly due to eigenvalue decomposition, $O(p^3)$ flops per iteration which may be as demanding as Rothman's method. Similarly Liu et al. [27] also considers (6) in which $\mathbf{S}$ is replaced by the sample correlation matrix.

We conduct a small simulation study on the above PD-sparse covariance matrix estimators to understand their estimation performance and computational hardness. Table 1 lists empirical risks (estimation errors evaluated in matrix ℓ_1, spectral, and Frobenius norms) and computation times of $\widehat{\boldsymbol{\Sigma}}^{\text{logdet}}$, $\widehat{\boldsymbol{\Sigma}}^{\text{EigCon}}$, and $\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}$, for a fixed tuning parameter, based on 100 replications from the same simulation settings as that used in Figure 1. Both of the PD sparse covariance estimators show comparable performance to that of the soft thresholding estimator. However, the computation times of PD sparse estimators are substantially increased. According to the table, when the five-fold cross validations are done for tuning parameter selection with 100 candidates of λ, it takes more than 2,500 seconds to get either $\widehat{\boldsymbol{\Sigma}}^{\text{logdet}}$ or $\widehat{\boldsymbol{\Sigma}}^{\text{EigCon}}$, while $\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}$ only costs 6 seconds.

4

	Matrix ℓ_1	Spectral	Frobenius	#(PD)	Comp. Time (sec)
Soft thresholding	19.34	7.47	25.94	0/100	0.01
Eigen. constraint	18.92	7.45	25.92	100/100	4.93
log-det barrier	18.90	7.45	25.99	100/100	2.33

Table 1: Averaged empirical risks (estimation errors evaluated in matrix ℓ_1, spectral, and Frobenius norms), frequency of positive definite estimates, and computing time (in seconds), for one fixed tuning parameter under 100 replications from the same simulation settings as that used in Figure 1. Computational convergence criteria is set as the relative error be smaller than 10^{-7}. ϵ and τ are set as 10^{-2} for both the eigenvalue constraint method and the log-determinant barrier method.

We finally remark that the above optimization-based PD sparse (regularized) estimators are only applicable to a specific type of regularized estimator, that is expressible as the minimum of a convex function. Note that not all regularized estimators have the convex expression; for example, the hard-thresholding estimator or SCAD estimator does not have the convex expression. In addition, even though an initial regularized estimator can be written as the solution to a convex optimization problem, the modified objective function with the PD constraint (or penalty) strongly depends on the initial convex expression. Thus, the corresponding PD regularized estimator should re-defined; the algorithm to solve the modified optimization problem should re-developed; and the statistical property of the resulting estimator should also be re-investigated, as a new.

3. The linear shrinkage for fixed support positive definiteness (FSPD)

We now explain how the FSPD procedure is applied to any covariance matrix estimator that possibly lacks PDness. Let $\widehat{\boldsymbol{\Sigma}}$ be a given initial estimator. Recall that we propose a class of linear shrinkage as a modification of $\widehat{\boldsymbol{\Sigma}}$,

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

where $\alpha \in [0,1]$ and $\mu \in \mathbb{R}$. Our aim is to minimize the distance between $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ and $\widehat{\boldsymbol{\Sigma}}$, while keeping the minimum eigenvalue of $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ positive. We state this constrained minimization problem quantitatively in Section 3.1 and derive the minimum for α while fixing μ as a constant in Section 3.2. Subsequently, we determine the condition of μ that minimizes the distances for the spectral and Frobenius norms in Section 3.3. The statistical convergence rate of $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ for carefully chosen μ and α are established in Section 3.4. Moreover, we discuss the fast computation of the FSPD procedure in Section 3.5.

3.1. Distance minimization

We set a small cut-point $\epsilon > 0$ to determine whether $\widehat{\boldsymbol{\Sigma}}$ is PD; we will modify $\widehat{\boldsymbol{\Sigma}}$ if $\widehat{\gamma}_1 < \epsilon$, otherwise we do not need to update it. Let $\widehat{\gamma}_1 < \epsilon$. We solve the following minimization problem:

$$\underset{\mu,\alpha\in\mathbb{R}}{\text{minimize}} \quad \left\|\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}) - \widehat{\boldsymbol{\Sigma}}\right\| \tag{7}$$
$$\text{subject to} \quad \begin{aligned}&\alpha\widehat{\gamma}_1 + (1-\alpha)\mu \geq \epsilon;\\ &\alpha \in [0,1).\end{aligned}$$

In (7), the first constraint enforces the minimum eigenvalue of $\boldsymbol{\Phi}_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}})$ to be at least ϵ. The second constraint specifies the range of α, intensity of linear shrinkage, where $\alpha = 0$ corresponds to the complete shrinkage to $\mu\mathbf{I}$ and $\alpha = 1$ denotes no shrinkage. By the assumption $\widehat{\gamma}_1 < \epsilon$, the two constraints also imply that $\mu \geq \epsilon$, because

$$\mu \geq \frac{1}{1-\alpha}\epsilon - \frac{\alpha}{1-\alpha}\widehat{\gamma}_1 > \frac{1}{1-\alpha}\epsilon - \frac{\alpha}{1-\alpha}\epsilon = \epsilon$$

for any $\alpha \in [0,1)$.

5

3.2. The choice of α

Let $\mu \in [\epsilon, \infty)$ be fixed. We have $\widehat{\gamma}_1 < \epsilon \le \mu$ since we are assuming $\widehat{\gamma}_1 < \epsilon$. Then the first constraint of (7) yields

$$1 - \alpha \ge \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}. \tag{8}$$

Observe that the objective function satisfies $\|\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\| = (1 - \alpha)\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\|$ and then achieves its minimum when $(1 - \alpha)$ touches the lower bound in (8). Thus, we have the following lemma.

Lemma 1. *Let $\widehat{\mathbf{\Sigma}}$ be given and assume $\epsilon > \widehat{\gamma}_1$. Then for any given $\mu \in [\epsilon, \infty)$, the problem (7), with respect to α, is minimized at*

$$\alpha^* := \alpha^*(\mu) = \frac{\mu - \epsilon}{\mu - \widehat{\gamma}_1}. \tag{9}$$

Lemma 1 also implies that the smallest eigenvalue of the updated estimator is always ϵ, because $\gamma_1\left\{\mathbf{\Phi}_{\mu,\alpha}(\widehat{\mathbf{\Sigma}})\right\} = \alpha^*\widehat{\gamma}_1 + (1 - \alpha^*)\mu = \epsilon$.

3.3. The choice of μ

By substituting (9) into (7), we have a reduced problem, which depends only on μ,

$$\underset{\mu \,:\, \mu > \epsilon}{\text{minimize}} \quad \left\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\right\| = \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|. \tag{10}$$

The solution of (10) differs according to how the matrix norm $\|\cdot\|$ is defined. We consider two most popular matrix norms in below, the spectral norm $\|\cdot\|_2$ and the (scaled) Frobenius norm $\|\cdot\|_{\text{F}}$.

Lemma 2 (Spectral norm). *If $\widehat{\mathbf{\Sigma}}$ is given and $\epsilon > \widehat{\gamma}_1$, then*

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\right\|_2 \ge \epsilon - \widehat{\gamma}_1 \quad \textit{for all } \mu \ge \epsilon,$$

and the minimum of (10) is achieved for any $\mu \ge \mu_{\text{S}} := \max\left\{\epsilon, (\widehat{\gamma}_p + \widehat{\gamma}_1)/2\right\}$.

Proof. By the definition of the spectral norm,

$$\begin{aligned}
\frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|_2 &= \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\max_i\left|\mu - \widehat{\gamma}_i\right| \\
&= \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\max\left\{|\mu - \widehat{\gamma}_p|, |\mu - \widehat{\gamma}_1|\right\}.
\end{aligned}$$

Consider $\psi_2(t) := \max\{|t - a|, |t - b|\}/(t - a)$ with $a < b$, $t \in (a, \infty)$. One can easily verify that $\psi_2(t) \equiv 1$ when $\mu \ge (a + b)/2$, and $\psi_2(t) > 1$ when $a < t < (a + b)/2$. Now substitute $t \leftarrow \mu$, $a \leftarrow \widehat{\gamma}_1$, and $b \leftarrow \widehat{\gamma}_p$. $\qquad\square$

Lemma 2 indicates that whenever we take μ such that $\mu \ge \mu_{\text{S}}$, the spectral-norm distance between the updated and initial estimators is exactly equal to the difference between their smallest eigenvalues — recall that the smallest eigenvalue of $\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}})$ is ϵ.

Lemma 3 (Scaled Frobenius norm). *Assume that $\widehat{\mathbf{\Sigma}}$ is given and $\epsilon > \widehat{\gamma}_1$. If $\widehat{\mathbf{\Sigma}} \ne c\mathbf{I}$ for any scalar c, then*

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\right\|_{\text{F}} \ge (\epsilon - \widehat{\gamma}_1)\sqrt{\frac{\sum_{i=1}^p(\widehat{\gamma}_i - \overline{\overline{\gamma}})^2}{\sum_{i=1}^p(\widehat{\gamma}_i - \widehat{\gamma}_1)^2}} \quad \textit{for all } \mu > \epsilon, \tag{11}$$

where $\overline{\overline{\gamma}} := \sum_i \widehat{\gamma}_i/p$. The equality holds if and only if

$$\mu = \mu_{\text{F}} := \frac{\sum_{i=1}^p(\widehat{\gamma}_i - \widehat{\gamma}_1)^2}{\sum_{i=1}^p(\widehat{\gamma}_i - \widehat{\gamma}_1)}.$$

6

In particular,

$$\left\|\mathbf{\Phi}_{\mu,\alpha^*}\!\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \le \epsilon - \widehat{\gamma}_1 \quad \textit{for all } \mu \ge \mu_{\mathrm{F}}. \tag{12}$$

Proof. We have that

$$
\begin{aligned}
\frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\left\|\mu\mathbf{I} - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} &= \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1}\sqrt{p^{-1}\sum_{i=1}^{p}(\mu - \widehat{\gamma}_i)^2} \\
&= \frac{\epsilon - \widehat{\gamma}_1}{\sqrt{p}}\sqrt{\frac{\sum_{i=1}^{p}\{(\mu - \widehat{\gamma}_1) - (\widehat{\gamma}_i - \widehat{\gamma}_1)\}^2}{(\mu - \widehat{\gamma}_1)^2}} \\
&= \frac{\epsilon - \widehat{\gamma}_1}{\sqrt{p}}\sqrt{\sum_{i=1}^{p}(\theta t_i - 1)^2}
\end{aligned}
$$

where $\theta = \theta(\mu) = (\mu - \widehat{\gamma}_1)^{-1}$ and $t_i = \widehat{\gamma}_i - \widehat{\gamma}_1$. The minimum of $h(\theta) = \sum_{i=1}^{p}(\theta t_i - 1)^2$ occurs at $\theta = (\sum_{i=1}^{p} t_i^2)/(\sum_{i=1}^{p} t_i)$ (equivalently $\mu = \mu_{\mathrm{F}}$), and is equal to

$$\frac{p\sum_{i=1}^{p}(t_i - \bar{t})^2}{\sum_{i=1}^{p} t_i^2},$$

which proves (11). In addition, that $\widehat{\mathbf{\Sigma}} \ne c\mathbf{I}$ ensures that $\sum_{i=1}^{p} t_i \ne 0$ and $\sum_{i=1}^{p} t_i^2 \ne 0$. To check (12), observe that each $(\theta t_i - 1)^2$, as a function of μ, is strictly increasing on $(\mu_{\mathrm{F}}, \infty)$ and converges to 1 as $\mu \to \infty$. $\qquad\square$

The assumption $\widehat{\mathbf{\Sigma}} \ne c\mathbf{I}$ implies that the initial estimator should not be trivial.

Finally, as a summary of Lemma 2 and 3, we have the following theorem.

Theorem 1 (Distance between the initial and FSPD estimators)**.** *Assume $\widehat{\mathbf{\Sigma}}$ and $\epsilon > 0$ are given. Set*

$$\alpha^* := \alpha^*(\mu) = \begin{cases} 1 & \textit{if } \widehat{\gamma}_1 \ge \epsilon \\ 1 - \frac{\epsilon - \widehat{\gamma}_1}{\mu - \widehat{\gamma}_1} & \textit{if } \widehat{\gamma}_1 < \epsilon \end{cases}; \qquad \mu \in [\mu_{\mathrm{SF}}, \infty),$$

where $\mu_{\mathrm{SF}} := \max\{\mu_{\mathrm{S}}, \mu_{\mathrm{F}}\}$ with μ_{S} and μ_{F} defined in Lemma 2 and 3, respectively. Then

1. *$\|\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\|_2$ exactly equals to $(\epsilon - \widehat{\gamma}_1)_+$ for any μ ;*

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

$$(\epsilon - \widehat{\gamma}_1)\sqrt{\frac{\sum_{i=1}^{p}(\widehat{\gamma}_i - \overline{\widehat{\gamma}})^2}{\sum_{i=1}^{p}(\widehat{\gamma}_i - \widehat{\gamma}_1)^2}} \le \left\|\mathbf{\Phi}_{\mu,\alpha^*}\!\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \le (\epsilon - \widehat{\gamma}_1)_+.$$

From the theorem, for any μ in $[\mu_{\mathrm{SF}}, \infty)$, the distance between the initial estimator $\widehat{\mathbf{\Sigma}}$ and its FSPD-updated version is less than $(\epsilon - \widehat{\gamma}_1)_+$. Indeed, the Monte Carlo experiments in Section 4 show that the estimating performance of $\mathbf{\Phi}_{\mu,\alpha^*}(\widehat{\mathbf{\Sigma}})$ is robust to the choice of μ. However, $\mu = \mu_{\mathrm{SF}}$ would be a preferable choice if one wants to minimize the Frobenius-norm distance from $\widehat{\mathbf{\Sigma}}$ while maintaining the spectral-norm distance as minimal. We remark that a special case of the proposed FSPD estimator has been briefly discussed by Cai et al. [8, Section 5.2]. In that study, the authors propose to modify the initial estimator $\widehat{\mathbf{\Sigma}}$ to $\widehat{\mathbf{\Sigma}} + (\epsilon - \widehat{\gamma}_1)\mathbf{I}$. This coincides with our FSPD procedure with $\alpha = \alpha^*$ and $\mu \to \infty$.

3.4. Statistical properties of the FSPD estimator

The convergence rate of FSPD estimator to the true $\mathbf{\Sigma}$ is based on the triangle inequality

$$
\begin{aligned}
\left\|\mathbf{\Phi}\!\left(\widehat{\mathbf{\Sigma}}\right) - \mathbf{\Sigma}\right\| &\le \left\|\mathbf{\Phi}\!\left(\widehat{\mathbf{\Sigma}}\right) - \widehat{\mathbf{\Sigma}}\right\| + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \\
&\le (\epsilon - \widehat{\gamma}_1)_+ + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|. \tag{13}
\end{aligned}
$$

7

To establish convergence rate for the modified estimator, we need to assume

(A1): The smallest eigenvalue of $\mathbf{\Sigma}$ is such that $\epsilon < \gamma_1$.

The assumption (A1) is the same with that assumed in Xue et al. [39]. It means that the cut-point ϵ, determining the PDness of the covariance matrix estimator, is set as smaller than the smallest eigenvalue of the true covariance matrix. For this choice of ϵ, we claim that the convergence rate of the FSPD estimator is at least equivalent to that of the initial estimator in terms of spectral norm.

Theorem 2. *Let $\widehat{\mathbf{\Sigma}}$ be any estimator of the true covariance matrix $\mathbf{\Sigma}$. Suppose ϵ satisfies* (A1)*, $\alpha = \alpha^*$ and $\mu \in [\mu_{\mathrm{SF}}, \infty)$, where α^* and μ_{SF} are defined in Theorem 1. Then, we have*

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\!\left(\widehat{\mathbf{\Sigma}}\right) - \mathbf{\Sigma} \right\|_2 \;\le\; 2 \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_2 .$$

Proof. For ϵ satisfying (A1), we have

$$(\epsilon - \widehat{\gamma}_1)_+ \;\le\; (\gamma_1 - \widehat{\gamma}_1)_+ \;\le\; \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_2 . \tag{14}$$

The first inequality in (14) is simply from $\epsilon \le \gamma_1$ and the second inequality follows from the Weyl's perturbation inequality, which states that if $\mathbf{A}$ and $\mathbf{B}$ are general symmetric matrices, then

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

Combining (13) and (14) completes the proof. $\qquad\square$

We remark that the arguments made for Theorem 2 are completely deterministic and there are no assumptions on how the true covariance matrix is structured or on how the estimator is defined. Thus, the convergence rate of FSPD estimator can be easily adapted to any given initial estimator. In terms of spectral norm, Theorem 2 shows that the FSPD estimator has convergence rate at least equivalent to that of the initial estimator, which implies that $\mathbf{\Phi}_{\mu,\alpha^*}\!\left(\widehat{\mathbf{\Sigma}}\right)$ is a consistent or minimax rate-optimal estimator if $\widehat{\mathbf{\Sigma}}$ is. Examples include the banding or tapering estimator [10], the adaptive block thresholding estimator [9], the universal thresholding estimator [11], and the adaptive thresholding estimator [5].

To discuss the convergence rate of the FSPD estimator in Frobenius norm, we further assume

$$\textbf{(A2):} \quad \frac{|\widehat{\gamma}_1 - \gamma_1|}{\sqrt{\sum_{i=1}^{p}(\widehat{\gamma}_i - \gamma_i)^2 / p}} = O_p(1).$$

The assumption (A2) implies that, in the initial estimator, the estimation error of the smallest eigenvalue has the same asymptotic order with the averaged mean-squared error over all eigenvalues. It ensures that the rate of convergence of the estimated eigenvalues are comparable across the spectrum. Since our PD modification is being done to all eigenvalues based on the error in estimation of the minimum eigenvalue, this type of assumption is needed to obtain consistency in terms of (scaled) Frobenius norm. With this additional assumption (A2), we have the following theorem.

Theorem 3. *Let $\widehat{\mathbf{\Sigma}}$ be any estimator of the true covariance matrix $\mathbf{\Sigma}$. Suppose $\alpha = \alpha^*$ and $\mu \in [\mu_{\mathrm{SF}}, \infty)$, where α^* and μ_{SF} are defined in Theorem 1. Then, under the assumptions* (A1) *and* (A2)*, we have*

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\!\left(\widehat{\mathbf{\Sigma}}\right) - \mathbf{\Sigma} \right\|_{\mathrm{F}} \;=\; \left\{ 1 + O_p(1) \right\} \cdot \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_{\mathrm{F}} .$$

8

Proof. Under the assumptions **(A1)** and **(A2)**, the inequality (13) shows

$$
\begin{aligned}
\left\|\mathbf{\Phi}_{\mu,\alpha^*}\big(\widehat{\mathbf{\Sigma}}\big) - \mathbf{\Sigma}\right\|_{\mathrm{F}} &\leq (\epsilon - \widehat{\gamma}_1)_+ + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\mathrm{F}} \\
&\leq (\gamma_1 - \widehat{\gamma}_1)_+ + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\mathrm{F}} \\
&= O_p(1)\sqrt{\sum_{i=1}^{p}(\widehat{\gamma}_i - \gamma_i)^2/p} + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\mathrm{F}} \\
&\leq \left\{1 + O_p(1)\right\}\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\mathrm{F}} \quad\quad (15)
\end{aligned}
$$

where the last inequality (15) is from Wielandt-Hoffman inequality: for any symmetric **A** and **B**,

$$
\sum_{i=1}^{p}\{\gamma_i(\mathbf{A}) - \gamma_i(\mathbf{B})\}^2 \leq p\,\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2.
$$

$\square$

We remark that the inequality $\left\|\mathbf{\Phi}_{\mu,\alpha^*}\big(\widehat{\mathbf{\Sigma}}\big) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \leq (\epsilon - \widehat{\gamma}_1)_+$ can be very conservative. When μ is close to μ_{F}, Theorem 1 implies

$$
\left\|\mathbf{\Phi}_{\mu,\alpha^*}\big(\widehat{\mathbf{\Sigma}}\big) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}} \approx (\epsilon - \widehat{\gamma}_1)_+ \sqrt{\frac{\sum_{i=1}^{p}(\widehat{\gamma}_i - \overline{\overline{\gamma}})^2}{\sum_{i=1}^{p}(\widehat{\gamma}_i - \widehat{\gamma}_1)^2}},
$$

where $\overline{\overline{\gamma}} = \sum_{i=1}^{p}\widehat{\gamma}_i/p$. Given a fixed sequence of $\widehat{\gamma}_i$, the term $\sum_{i=1}^{p}(\widehat{\gamma}_i - \overline{\overline{\gamma}})^2/\sum_{i=1}^{p}(\widehat{\gamma}_i - \widehat{\gamma}_1)^2$ could be considerably smaller than one. In sequel, the Frobenius norm $\left\|\mathbf{\Phi}_{\mu,\alpha^*}\big(\widehat{\mathbf{\Sigma}}\big) - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}}$ becomes much smaller than $(\epsilon - \widehat{\gamma}_1)_+$, and possibly comparable to $\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}}$. Thus, the assumption **(A2)** is only a sufficient condition, and the conclusion of Theorem 3 could be true in much more generality.

3.5. Computation

The proposed FSPD estimator has by itself a great computational advantage over the existing estimators that incorporate both structural regularization and PDness constraint in optimization problem [27, 32, 39]. It is primarily because the FSPD procedure is optimization-free and runs eigenvalue decomposition only once. Moreover, in practice, it can be implemented much more quickly without eigenvalue decomposition. Recall that the FSPD estimator depends on the four functionals of $\widehat{\mathbf{\Sigma}}$: $\widehat{\gamma}_p$, $\widehat{\gamma}_1$, $\sum_{i=1}^{p}(\widehat{\gamma}_i - \overline{\overline{\gamma}})^2$ and $\sum_{i=1}^{p}(\widehat{\gamma}_i - \widehat{\gamma}_1)^2$ which in turn can be written in terms of $\widehat{\gamma}_p$, $\widehat{\gamma}_1$, $\overline{\overline{\gamma}}$ and $\mathrm{V}(\widehat{\gamma}) := \sum_{i=1}^{p}(\widehat{\gamma}_i - \overline{\overline{\gamma}})^2$. For $\widehat{\gamma}_p$ and $\widehat{\gamma}_1$, the largest and smallest eigenvalue of a symmetric matrix can be independently calculated from the whole spectrum by the Krylov subspace method (see Chapter 7 of Demmel [13] or Chapter 10 of Golub and Van Loan [18] for examples), which is faster than the usual eigenvalue decomposition of large-scale sparse matrices. Some of these algorithms are implemented using public and commercial software; for example, the built-in MATLAB function `eigs()`, which is based on Lehoucq and Sorensen [25] and Sorensen [34], or a user-defined MATLAB function `eigifp()` [17], which is available online from Q. Ye's personal webpage (http://www.ms.uky.edu/~qye/). During the simulation, we use `eigifp()` instead of `eigs()`, as `eigs()` fails to converge in the smallest eigenvalue computations for some matrices. The other functionals are written as

$$
\overline{\overline{\gamma}} = \frac{1}{p}\sum_{i=1}^{p}\widehat{\gamma}_i = \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}), \qquad \mathrm{V}(\widehat{\gamma}) = \frac{1}{p}\sum_{i=1}^{p}\widehat{\gamma}_i^2 - \mathrm{M}(\widehat{\gamma})^2 = \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^2) - \left\{\frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}})\right\}^2,
$$

and can be evaluated without the computation of the entire spectrum.

9

4. Simulation study

We numerically compare the finite sample performances of the FSPD estimator and other existing estimators. For the comparison, we use the soft thresholding estimator with universal threshold in (3) as an initial regularized estimator. We show that the FSPD estimator induced by the soft thresholding estimator is comparable in performance to the existing PD sparse covariance matrix estimators and is computationally much faster.

4.1. Empirical risk

We generate samples of p-variate random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from (i) multivariate normal distribution or (ii) multivariate t distribution with 5 degrees of freedom. We consider two types of matrices as true $\boldsymbol{\Sigma}$ following Rothman [32] and Xue et al. [39]:

1. ("Linearly tapered Toeplitz" matrix) $[\mathbf{M}_1]_{ij} := (1 - |i - j|/10)_+$ (when $[\mathbf{A}]_{ij}$ denotes the (i, j)-th element of $\mathbf{A}$);

2. ("Overlapped block-diagonal" matrix) $[\mathbf{M}_2]_{ij} := I(i = j) + 0.4\, I((i, j) \in (I_k \cup \{i_k + 1\}) \times (I_k \cup \{i_k + 1\}))$, where the row (column) index $\{1, \ldots, p\}$ is partitioned into $K := p/20$ subsets, which are non-overlapping and of equal size, and i_k denotes the maximum index in I_k.

We generate 100 datasets for each of the 36 possible combinations: distribution $\in$ {multivariate normal, multivariate-t}, $\boldsymbol{\Sigma} \in \{\mathbf{M_1}, \mathbf{M_2}\}$, $n \in \{100, 200, 400\}$, and $p \in \{100, 200, 400\}$.

We first compute the sample covariance matrix $\mathbf{S} := (n - 1)^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\top}$, where $\bar{\mathbf{x}} := n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$, and the soft thresholding estimator $\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}(\lambda^*)$. The optimal threshold λ^* is chosen from the five-fold cross-validation (CV) to minimize (an empirical approximation of) the predictive risk:

$$\text{PR}(\lambda) = \|\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}_{\text{Train}}(\lambda) - \mathbf{S}_{\text{Test}}\|_{\text{F}}^2,$$

where the soft thresholding estimator $\widehat{\boldsymbol{\Sigma}}^{\text{Soft}}_{\text{Train}}(\lambda)$ is obtained using the training dataset with a fixed tuning parameter λ and the sample covariance matrix $\mathbf{S}_{\text{Test}}$ is calculated using the testing dataset. The candidate set of λ is $\{(k/100) \cdot \max_{i<j} |s_{ij}| : k = 0, \ldots, 100\}$, a grid search from zero to the maximum mangnitude of sample covariances.

First, we investigate the spectra of the soft thresholding estimator to understand how often and in what magnitude it violates PDness. The frequencies of non-PD estimates and the proportion of negative eigenvalues of the soft thresholding estimator are empirically studied by Xue et al. [39]. Here, we further investigate the magnitudes/distributions of the smallest eigenvalues under normal and t-distributions; the empirical results from a messier distribution (t-distribution) beyond the normal are new here. Table 2 summarizes the negative eigenvalues of the soft thresholding estimates from the simulated datasets. For both $\mathbf{M}_1$ and $\mathbf{M}_2$, we find that the soft thresholding estimator easily lacks PDness. The frequency of non-PDness increases as p increases as reported in Xue et al. [39]. Further, the smallest eigenvalues of the estimates become more negatively away from zero when the distribution has heavier tail (t-distribution) or the dimension p is larger. The results suggest that PD-modification procedures are more frequently mandatory in more high-dimensional or heavier tailed data.

p		$\boldsymbol{\Sigma} = \mathbf{M}_1$			$\boldsymbol{\Sigma} = \mathbf{M}_2$		
		Min. eig.	#(Neg. eig.)/p	#(PD)	Min. eig.	#(Neg. eig.)/p	#(PD)
100	$\mathcal{N}$	-0.035 (0.004)	0.017 (0.001)	16/100	0.257 (0.003)	0.000 (0.000)	100/100
	t	-0.066 (0.006)	0.020 (0.001)	9/100	0.292 (0.012)	0.000 (0.000)	100/100
200	$\mathcal{N}$	-0.061 (0.003)	0.017 (0.001)	2/100	0.138 (0.003)	0.000 (0.000)	100/100
	t	-0.114 (0.020)	0.018 (0.001)	4/100	0.133 (0.018)	0.001 (0.000)	87/100
400	$\mathcal{N}$	-0.086 (0.003)	0.016 (0.001)	0/100	-0.039 (0.003)	0.006 (0.000)	7/100
	t	-0.270 (0.026)	0.017 (0.001)	0/100	-0.150 (0.026)	0.014 (0.001)	7/100

Table 2: Non-PDness of the soft thresholding estimator. The columns give the minimum eigenvalue (Min. eig), the proportion of negative eigenvalues (#(Neg. eig.)/p), and the number of cases that the estimates are PD (#(PD)) over 100 replications. Corresponding standard errors are presented in parenthesis.

In order to compare the empirical risks, we consider the following four PD covariance matrix estimators:

1. ("FSPD(μ_{SF})") FSPD estimator $\mathbf{\Phi}_{\mu,\alpha^*}\{\widehat{\mathbf{\Sigma}}^{\text{Soft}}(\lambda^*)\}$ with $\mu = \mu_{\text{SF}}$

2. ("FSPD(∞)") FSPD estimator $\mathbf{\Phi}_{\mu,\alpha^*}\{\widehat{\mathbf{\Sigma}}^{\text{Soft}}(\lambda^*)\}$ with $\mu = \infty$

3. ("EigCon") Xue et al. [39]'s eigenvalue constraint estimator, $\widehat{\mathbf{\Sigma}}^{\text{EigCon}}(\lambda^*)$ in (6)

4. ("log-det") Rothman [32]'s log-determinant barrier estimator $\widehat{\mathbf{\Sigma}}^{\text{logdet}}(\lambda^*)$ in (5)

In applying the respective methods, we set $\epsilon = 10^{-2}$ for the two FSPD estimators and the eigenvalue constraint estimator and $\tau = 10^{-2}$ for the log-determinant barrier estimator, which are larger than Rothman [32] and Xue et al. [39]'s choices ($\epsilon = 10^{-5}$ and $\tau = 10^{-4}$), so as to reduce the computation times for both estimators. In theory, smaller values on ϵ and τ are preferred by (**A1**). But as Table 3 shows, our choice of $\epsilon = 10^{-2}$ and $\tau = 10^{-2}$ result in almost the same empirical risks with the initial soft-thresholding estimator.

Table 3 lists the empirical risks (estimation errors) measured by three popular matrix norms (the matrix ℓ_1, spectral, and unscaled Frobenius norms) for the four estimators considered. The table does not include the results for $p = 100$, 200 of $\mathbf{M}_2$, because the initial soft thresholding estimates are mostly PD and empirical risks are almost identical across the considered estimators.

Now, we compare the empirical risks between the estimators. For the soft thresholding and FSPD approaches, The estimation errors of both FSPD(μ_{SF}) and FSPD(∞) are lower in the spectral norm and higher in both the matrix ℓ_1 and Frobenius norms than those of soft thresholding estimates; however, the differences do not exceed 4%. FSPD(μ_{SF}) demonstrates up to 2% greater empirical risk than FSPD(∞) in all the three norms for normally distributed datasets but lower empirical risk (within 5%) for t-distributed datasets.

A comparison of the FSPD estimator and the two optimization-based estimators show that in all simulated cases, FSPD(μ_{SF}) has lower risks compared to both the eigenvalue constraint and log-determinant methods in the matrix ℓ_1 and spectral norms, but higher risks in the Frobenius norm; however, the difference never exceeds 4%. In addition, FSPD(∞) has a higher risk than the two optimization-based methods in both the matrix ℓ_1 and Frobenius norms, except for $\mathbf{M}_2$ under $p = 400$, and lower risk in spectral norm. Similarly, the difference of risks are less than 4%.

In summary, the empirical risks associated with the inspected estimators are different within approximately 4%; furthermore, the standard errors of these risks show that the risks of the other methods are within the confidence interval of each of the estimators. Therefore, we can conclude that the empirical errors of the proposed FSPD estimators are comparable to the errors of the soft thresholding estimator as well as the two optimization-based PD sparse estimators when the error is measured by the matrix ℓ_1, spectral, and Frobenius norms.

4.2. Computation time

We now numerically show the proposed FSPD estimator is much faster and simpler than optimization-based PD sparse estimators. We record the computation time of the four PD covariance matrix estimators listed in Section 4.1 as well as the soft thresholding estimator for $n = 100$ and $p = 400, 1200, 3600$. The distribution is multivariate normal with the true covariance matrix $\mathbf{\Sigma} = \mathbf{M}_1$ and $\mathbf{M}_2$. Here, λ^* is selected as in Section 4.1. The calculations in this Section are performed using MATLAB running on a computer with an Intel Core i7 CPU (3.4 GHz) and 16 GB RAM. Since the codes for both eigenvalue constraint and log-determinant methods are not available in MATLAB, we write the codes by ourselves. As those two optimization-based estimators - the eigenvalue constraint estimator and the log-determinant barrier estimator - are solved iteratively, and the convergence criteria is set as $\|\widehat{\mathbf{\Sigma}}^{\text{(New)}} - \widehat{\mathbf{\Sigma}}^{\text{(Old)}}\|/\|\widehat{\mathbf{\Sigma}}^{\text{(Old)}}\| < 10^{-7}$. We remark that, due to the time issue of the log-determinant approach as pointed out in Xue et al. [39], we have coded its inner loop in C. Thus, the comparison of computation times between log-det and EigCon approach is not anymore meaningful, while it is worth to compare the times between our FSPD and the two existing approaches.

The results are summarized in Table 4. It is not very surprising that the two FSPD estimates are calculated extremely faster than the eigenvalue constraint and the log-determinant barrier estimates, because both the latter methods require iterative computations of $O(p^3)$ flops as noted in Sections 2.2 and 3.5. In addition, FSPD(∞) is always faster than FSPD(μ_{SF}), since FSPD(μ_{SF}) computes $\widehat{\gamma}_1, \widehat{\gamma}_p, \overline{\overline{\gamma}}$, and V$(\widehat{\gamma})$ whereas FSPD(∞) calculates $\widehat{\gamma}_1$ only. We also note that the empirical risks from both FSPD(μ_{SF}) and FSPD(∞) are consistent even when $n = 100$ and $p = 1200, 3600$, which is omitted here to save space.

In summary, the FSPD estimator induced by the soft thresholding estimator has empirical risk comparable to the two existing PD sparse covariance matrix estimators, but it is computationally much faster and simpler than them.

11

	Multivariate normal			Multivariate t		
	Matrix ℓ_1	Spectral	Frobenius	Matrix ℓ_1	Spectral	Frobenius
			$\mathbf{\Sigma} = \mathbf{M}_1$, $p = 100$			
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)
FSPD(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)
FSPD(∞)	6.20 (0.11)	3.56 (0.05)	7.21 (0.07)	9.23 (0.35)	5.04 (0.13)	10.41 (0.17)
EigCon	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)
			$\mathbf{\Sigma} = \mathbf{M}_1$, $p = 200$			
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)
FSPD(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)
FSPD(∞)	7.05 (0.08)	4.16 (0.04)	11.40 (0.05)	13.51 (0.78)	6.24 (0.21)	17.40 (0.45)
EigCon	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)
			$\mathbf{\Sigma} = \mathbf{M}_1$, $p = 400$			
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)
FSPD(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)
FSPD(∞)	7.93 (0.08)	4.62 (0.03)	17.86 (0.06)	19.60 (1.15)	7.57 (0.34)	26.87 (0.57)
EigCon	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)
			$\mathbf{\Sigma} = \mathbf{M}_2$, $p = 400$			
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)
FSPD(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)
FSPD(∞)	11.73 (0.10)	5.59 (0.05)	19.32 (0.07)	19.83 (0.60)	7.39 (0.09)	28.32 (0.40)
EigCon	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)

Table 3: Empirical risks (estimation errors evaluated in matrix ℓ_1, spectral, and Frobenius norms) of the PD covariance matrix esimators and soft thresholding estimator based on 100 replications. Standard errors are presented in parenthesis.

	$\mathbf{\Sigma} = \mathbf{M}_1$			$\mathbf{\Sigma} = \mathbf{M}_2$		
	$p = 400$	$p = 1200$	$p = 3600$	$p = 400$	$p = 1200$	$p = 3600$
Soft thres.	0.00	0.02	0.23	0.00	0.03	0.24
FSPD(μ_{SF})	0.01	0.12	0.66	0.01	0.13	0.82
FSPD(∞)	0.01	0.09	0.50	0.01	0.09	0.58
EigCon	4.93	190.68	7757.47	2.42	106.13	4470.47
log-det	2.33	99.14	3157.80	2.30	97.28	3156.08

Table 4: Computation time (in seconds) of the four PD estimators and the soft thresholding estimator.

5. Two applications

In this Section, we apply the FSPD approach to two statistical procedures from the literature, linear minimax classification and constrained Markowitz's portfolio allocation, which require PD estimation of covariance matrix. Both applications are illustrated with real data examples. The linear minimax classifier is illustrated with an example from speech recognition [36], and Markowitz's portfolio allocation is illustrated with a Dow Jones stock return example [38].

5.1. Linear minimax classifier applied to speech recognition

5.1.1. Linear minimax probability machine

The *linear minimax probability machine* (LMPM) proposed by Lanckriet et al. [23] constructs a linear binary classifier with no distributional assumptions given the mean vector and covariance matrix. Let $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ be a pair of the population mean vector and covariance matrix for group 0 and 1, respectively. The LMPM finds a separating hyperplane that minimizes the maximum probability of misclassification over all distributions for given $\boldsymbol{\mu}$s and $\boldsymbol{\Sigma}$s:

$$\max_{\alpha, \mathbf{a} \neq 0, b} \quad \text{s.t.} \quad \inf_{\mathbf{x} \sim (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)} \Pr\left(\mathbf{a}^\top \mathbf{x} \leq b\right) \geq \alpha \tag{16}$$

$$\inf_{\mathbf{y} \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)} \Pr\left(\mathbf{a}^\top \mathbf{x} \geq b\right) \geq \alpha$$

The solution to (16) does not have a closed-form formula and should be numerically computed, although it is expressed by the minimizer of a convex program. The PDness of the covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ (or their estimators) is a necessary condition for both the convexity of the problem and the existence of a convergent algorithm to the solution.

In practice, the true $\boldsymbol{\mu}$s and $\boldsymbol{\Sigma}$s are unknown and we must plug their estimators into the LMPM formulation (16). Lanckriet et al. [23] uses sample mean and covariance matrices when they are well defined. In case the sample covariance matrix $\mathbf{S}$ is singular, they suggest using $\mathbf{S} + \delta\mathbf{I}$ with a given constant δ rather than $\mathbf{S}$.

5.1.2. Example: Speech recognition

We illustrate the performance of the LMPM with various PD covariance matrix estimators using voice data [36]. The data are available on the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). The dataset comprises 126 signals representing the pronunciation of vowel /a/ by patients with Parkinson's disease. Each signal was pre-processed into 309 features and labeled as "acceptable" (42 signals) or "unacceptable" (84 signals) by an expert. The dataset is in the form of a 126×309 matrix with binary labels. Rothman [32] also considers similar speech recognition examples – our analysis example is different in two aspects. First, Rothman [32] solves quadratic discriminant analysis, while we optimize the linear minimax probability machine. Second, Rothman [32] uses a dataset with $p = 22$ provided by Little et al. [26], while we use a more recent and larger-dimensional dataset with $p = 309$ provided by the same research team [36], which has more often non-PD estimates for thresholding estimators (detailed results are omitted to save space). Thus, we believe that the dataset of Tsanas et al. [36] would more clearly illustrate the advantage of our PD modification method.

In order to measure classification accuracy, we randomly split the dataset 100 times into 90% training samples and 10% testing samples; the LMPM is constructed using the training samples and classification accuracy is measured using the testing samples. In building the LMPM, the true mean vectors are estimated from the sample means and covariance matrices are estimated from the following PD estimators: (1) "Sample," the sample covariance matrix added by $\delta\mathbf{I}$, where $\delta = 10^{-2}$; (2) "Diagsample," the diagonal matrix of the sample variances; (3) "LedoitWolf," the linear shrinkage estimator by Ledoit and Wolf [24]; (4) "CondReg," the condition-number-regularized estimator by Won et al. [38]; (5) "Adap.+EigCon," the eigenvalue constraint estimator by Xue et al. [39] based on the adaptive thresholding estimator by Cai and Liu [5] with the soft thresholding; and (6) "Adap.+FSPD," the proposed FSPD estimate ($\alpha = \alpha^*, \mu = \mu_{\text{SF}}$) induced by 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 marginal variances of the given data can be unequal over variables, in which case the adaptive thresholding is known to perform better than the universal thresholding [5]. The tuning parameter of the adaptive thresholding estimator is selected using the five-fold CV as in Cai and Liu [5]. We find that none among the 100 adaptive thresholding estimates (from 100 random splitting of training and testing samples) are PD and, thus, the PD modification is an important step to facilitate the use of the adaptive thresholding estimates in the LMPM. The pre-determined constant ϵ for Adap.+EigCon and Adap.+FSPD is set as 10^{-5}, which is the same with that in Xue et al. [39]. For implementation, we code the Adap.+FSPD and the Adap.+EigCon by ourselves. The MATLAB codes for the LMPM procedure and LedoitWolf estimator are available in the authors' personal webpages (e.g., http://gertlanckriet.com/ and http://www.ledoit.net), while those for the CondReg procedure was provided by the courtesy of J.-H. Won.

13

The average and standard deviation of classification accuracy over the 100 random partitions are reported in Table 5. All the regularization methods (items 3–6) show better classification accuracy than the naive sample covariance matrix (item 1). In addition, Adap.+FSPD has the highest accuracy with average 89.3% and standard deviation 9.4%. This record is highly competitive to the results reported in the original paper [36], which is based on a random forest and support vector machine after a feature selection algorithm named LOGO by Sun et al. [35]. Finally, we note that the entire process of Adap.+FSPD, from covariance matrix estimation to LMPM construction, takes only 0.68 seconds.

Sample	Diagsample	LedoitWolf	CondReg	Adap.+EigCon	Adap.+FSPD
73.8 (12.4)	76.4 (12.6)	86.9 (10.7)	75.3 (13.6)	81.0 (19.5)	89.3 (9.4)

Table 5: The average classification accuracies (standard deviations in parenthesis) for the LMPM with selected PD covariance matrix estimators based on 100 random partitions. The abbreviations of the estimators are introduced in the main body of the Section.

5.2. Markowitz portfolio optimization

5.2.1. Minimum-variance portfolio allocation and short-sale

In finance, portfolio refers to a family of (risky) assets held by an institution or private individual. If there are multiple assets to invest in, a combination of assets is considered and it becomes an important issue to select an optimal allocation of portfolio. *Minimum variance portfolio* (MVP) optimization is one of the well-established strategies for portfolio allocation [12]. The author proposes to choose a portfolio that minimizes risk, standard deviation of return. Since the MVP problem may yield an optimal allocation with short-sale or leverage, Jagannathan and Ma [19] proposes to add a no-short-sale constraint to the MVP optimization formula.

We briefly introduce the settings for the constrained MVP optimization. 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 constituting portfolio ($j = 1, \ldots, p$). Denote by $\mathbf{\Sigma} := \text{Var}(\mathbf{r})$ the unknown covariance matrix of assets. A p-by-1 vector $\mathbf{w}$ represents an allocation of the investor's wealth in such a way that each w_j stands for the weight of the j-th asset and $\sum_{j=1}^p w_j = 1$. Then, the MVP optimization by Chan [12] is formulated as

$$\underset{\mathbf{w}}{\text{minimize}} \ \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \ \text{ subject to } \ \mathbf{w}^\top \mathbf{1} = 1. \tag{17}$$

Note that (17) allows its solution to have a weight that is negative (short-sale) or greater than 1 (leverage). Jagannathan and Ma [19] points out that short-sale and leveraging are sometimes impractical because of legal constraints and considers the MVP optimization problem with no short-sale constraint:

$$\underset{\mathbf{w}}{\text{minimize}} \ \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \ \text{ subject to } \ \mathbf{w}^\top \mathbf{1} = 1, \ \mathbf{w} \geq \mathbf{0}, \tag{18}$$

where $\mathbf{w} \geq \mathbf{0}$ is defined component-wise. The combination of two constraints in (18) ensures that the resulting optimal weights are restricted to $[0, 1]$. Jagannathan and Ma [19] empirically and analytically illustrate that (18) could have a smaller risk than (17) even if the no-short-sale constraint is wrong. We note that the paper handles only the case in which the sample covariance matrix is nonsingular and plugged into the unknown $\mathbf{\Sigma}$ in (18). In principle, $\mathbf{\Sigma}$ can be replaced by any suitable PD estimator.

5.2.2. Example: Dow Jones stock return

The aim of this analysis is not only to reproduce with different data the finding of Jagannathan and Ma [19] that the no-short-sale constraint does not affect the risk of (17), but also to investigate empirically whether the same conclusion can hold for another choice of PD covariance matrix estimator. We compare two portfolio optimization schemes: simple MVP (17) and no-short-sale MVP (18) The unknown covariance matrix $\mathbf{\Sigma}$ is estimated with seven PD covariance matrix estimators. Five estimators, (1) "Sample," (2) "LedoitWolf," (3) "CondReg," (4) "Adap.+EigCon," and (5) "Adap.+FSPD," are already introduced in Section 5.1.2. To these, we add (6) "POET+EigCon," an eigenvalue constraint estimator based on the POET estimator proposed by Fan et al. [14] (see Xue and Zou's discussion on pp. 672 of the cited paper), and (7) "POET+FSPD," the FSPD estimator induced by the POET estimator. The reason why we additionally consider (6) and (7) is that stock return data are believed to have a factor structure that can

14

supposedly be reflected through the POET estimation. The source code for POET is available in J. Fan's personal webpage (http://orfe.princeton.edu/~jqfan/). We remark that Fan et al. [14] propose to additionally tune the thresholding parameter to achieve the PDness of the estimator. Here, in this example, we use the POET estimator without the additional tuning to understand the performance of the FSPD modification. The pre-determined constant ϵ for (4)–(7) is set to 10^{-5} as in the previous data example. The example of no-short-sale MVP is also presented in Xue et al. [39]. However, their example is limited to the comparison between the sample covariance matrix and the EigCon based on the universal thresholding, while our study here exhaustively compare most of estimators in the literature, that are applied to no-short-sale MVP. We believe this provides a good support for our proposal.

As data, we use the 30 stocks that constituted the Dow Jones Industrial Average in July 2008; previously, these were used in Won et al. [38]. The dataset contains the daily closing prices from December 1992 to June 2008, adjusting for splits and dividend distributions. The portfolios are constructed as follows. For covariance matrix estimation, we use the stock returns of the past 60 or 240 trading days (approximately 3 or 12 months, respectively). As Condreg, Adap, and POET require the selections of tuning parameters, they are done via five-fold CV in which the returns of each day are treated as independent samples. Once the portfolios are established by solving (17) and (18) with covariance matrices plugged in by the corresponding estimators, we hold each for 60 trading days. This process begins on February 18, 1994, and is continually repeated until July 6, 2008, producing 60 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 *the Sharpe ratio*. Here, the realized return and realized risk of a portfolio are defined as the average and standard deviation of daily returns from that portfolio, respectively. The Sharpe ratio is a risk-adjusted performance index, defined as {(realized return) - (risk-free rate)}/(realized risk). The risk-free rate is set at 5% per annum.

	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	24.63	4.17	4.17	4.73	4.71
Adap.+FSPD	20.39	21.16	3.53	3.07	4.37	5.26
Adap.+EigCon	20.51	21.18	3.54	3.07	4.39	5.26
POET+FSPD	19.30	20.69	3.14	3.04	4.55	5.16
POET+EigCon	19.07	20.60	3.15	3.04	4.47	5.13
(Portfolio rebalancing based on 240 previous trading days)						
Sample	22.42	23.06	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.+FSPD	21.13	22.91	2.93	2.94	5.51	6.09
Adap.+EigCon	21.13	22.91	2.93	2.94	5.51	6.09
POET+FSPD	21.81	23.33	2.94	2.95	5.72	6.21
POET+EigCon	22.16	23.36	2.97	2.95	5.77	6.22

Table 6: 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 results summarized in Table 6. We first compare the realized risks of the simple and no-short-sale MVPs; the third and fourth columns of the table show that the differences in the realized risks of the respective MVPs are competitive or lower in no-short-sale MVPs. This appears to reproduce the findings of Jagannathan and Ma [19]; the paper finds that no-short-sale MVPs have a smaller risk than simple MVPs. Interestingly, the results of realized returns in the first and second columns show that no-short-sale MVPs produce higher realized returns than all simple MVPs except for Condreg, showing a relative improvement of the Sharpe ratios in all cases except for Condreg.

Next, we compare the 60- and 240-day training results for the construction of portfolios (rows 1–6 and 7–12).The LedoitWolf, the Adap.s and the POETs show higher realized returns and Sharpe ratios for the 240-day training, where

the POETs show the largest improvements. This is true for both simple and no-short-sale MVPs. We conjecture that the factored structure of the POET estimation captures the latent structures of stocks better in long-history data than short-history data.

CondReg performs somewhat differently from other regularized covariance matrix estimators; the realized return of CondReg dominates those of the other methods. This confirms the results in Won et al. [38], which illustrates that an MVP with CondReg produces the highest wealth growth. This seems to be empirical evidence for "high risk, high return," as CondReg shows the highest realized risk.

Unlike the previous example, the FSPD estimators (Adap.+FSPD and POET+FSPD) and the eigenvalue constraint estimators (Adap.+EigCon and POET+EigCon) perform almost equally. This is because the PD modifications are effective in only a few times. In details, when training data set is composed of the records from previous 60 trading days at the time, both Adap. and POET estimates are PD in 53 cases and PD modifications (both FSPD and EigCon) are applied to only 7 remaining non-PD cases, For the 240-day training (training with the records from 240 previous trading days), the PD updates are applied for only 14 case of the POET estimates which are not PD. As the FSPD estimators are far simpler and faster than the eigenvalue constraint estimators, we claim that they are more desirable for practical use.

6. Estimation of positive definite precision matrices

In this Section, we emphasize that the FSPD approach can be applied to estimating PD precision matrix estimators as well. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ be the unknown true precision matrix and $\widehat{\mathbf{\Omega}}$ be one of its estimators, which possibly lacks PD. In the theory presented in Section 3, the initial estimator $\widehat{\mathbf{\Sigma}}$ can be treated as a generic symmetric matrix and thus Theorem 1 is valid for $\widehat{\mathbf{\Omega}}$. Therefore, we have the following theorem.

Theorem 4. *Let $\widehat{\mathbf{\Omega}}$ be any estimator of the true precision matrix $\mathbf{\Omega}$. Suppose $\alpha = \alpha^*$ and $\mu \in [\mu_{\mathrm{SF}}, \infty)$, where α^* and μ_{SF} are defined in Theorem 1 with $\widehat{\gamma}_i$ being understood by the i-th smallest eigenvalue of $\widehat{\mathbf{\Omega}}$. Then, under the assumption* (**A1**)*, we have*

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Omega}}\right) - \mathbf{\Omega} \right\|_2 \leq 2 \left\| \widehat{\mathbf{\Omega}} - \mathbf{\Omega} \right\|_2.$$

Further, if (**A2**) *holds additionally,*

$$\left\| \mathbf{\Phi}_{\mu,\alpha^*}\left(\widehat{\mathbf{\Omega}}\right) - \mathbf{\Omega} \right\|_{\mathrm{F}} = \left(1 + O_p(1)\right) \cdot \left\| \widehat{\mathbf{\Omega}} - \mathbf{\Omega} \right\|_{\mathrm{F}}.$$

Theorem 4 implies that, as in the estimation of the PD covariance matrix, the FSPD-updated precision matrix estimator can preserve both the support and convergence rate.

Several regularization methods are proposed for sparse precision matrices, and, like regularized covariance matrix estimation, some estimators do not guarantee PDness in finite sample. Examples include CLIME by Cai et al. [6], the neighborhood selection by Meinshausen and Bühlmann [30], SPACE by Peng et al. [31], the symmetric lasso by Friedman et al. [16], and CONCORD by Khare et al. [20]. In addition, penalized Gaussian likelihood methods, whose very definition ensures PDness of the solution, may have a solution lacking PDness, since the relevant optimization algorithms approximate solutions by a non-PD matrix. For instance, Mazumder and Hastie [29] reports that the graphical lasso algorithm by Friedman et al. [15] and Witten et al. [37] solves the problem in an indirect way and could return non-PD solutions.

Despite the possibility of non-PDness, we empirically find that the precision matrix estimators listed above are likely to be PD; thus, they suffer minimally from potential non-PDness. Let $\widehat{\mathbf{\Omega}}(\lambda)$ be a regularized precision estimator with tuning parameter λ. Although non-PDness often arises in $\widehat{\mathbf{\Omega}}(\lambda)$ when λ is not large and $\widehat{\mathbf{\Omega}}(\lambda)$ is dense, we observe that, with optimal selection of λ^* through CV, $\widehat{\mathbf{\Omega}}(\lambda^*)$ is PD in most cases. The case $\mathbf{\Omega} = \mathbf{M}_1$ (tapered Toeplitz matrix) is the only one in which we find non-PDness in some precision matrix estimators (CONCORD and symmetric lasso). Table 6 summarizes the spectral information of these estimators.

7. Concluding remarks

In this study, we propose a two-stage approach with covariance matrix regularization and the conversion toward PDness considered to be two separate steps. Because of its two-staged nature, the proposed FSPD procedure can be

16

		CONCORD [20]		The symmetric lasso [16]	
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 7: Spectral information of selected sparse precision matrix estimators when the true precision matrix is $\mathbf{M}_1$ (tapered Toeplitz matrix).

combined with any regularized covariance or precision matrix estimator. The procedure considers a convex combination involving only the initial regularized estimator and the related parameters can be selected in explicit form. Thus, the FSPD procedure is optimization-free and can be quickly computed. Despite its simplicity, the FSPD estimator enjoys theoretical advantages, in that it can preserve *both the sparse structure and convergence rate* of a given initial estimator. Our procedure is also applicable to non-PD correlation matrix estimator, although the updated matrix may not have its diagonal terms being equal to one. The FSPD-updated correlation matrix is useful for statistical procedures that require only the PDness and a good convergence rate of correlation matrix estimator [21].

The FSPD procedure finds a PD matrix close to the initial estimator subject to the class of linear shrinkage estimators. Here, we conclude the paper with a discussion on the linear shrinkage constraint in the FSPD. Consider three classes of covariance matrices, $\mathcal{S}_0 = \{\widehat{\boldsymbol{\Sigma}}^* : \widehat{\boldsymbol{\Sigma}}^* = (\widehat{\boldsymbol{\Sigma}}^*)^\top, \gamma_1(\widehat{\boldsymbol{\Sigma}}^*) \geq \epsilon\}$, $\mathcal{S}_1 = \{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0 : \mathrm{supp}(\widehat{\boldsymbol{\Sigma}}^*) = \mathrm{supp}(\widehat{\boldsymbol{\Sigma}})\}$, and $\mathcal{S}_2 = \{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0 : \widehat{\boldsymbol{\Sigma}}^* = \alpha\widehat{\boldsymbol{\Sigma}} + (1-\alpha)\mu\mathbf{I}, \ \alpha \in [0,1], \ \mu \in \mathbb{R}\}$. In (1), we originally aim to solve $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_1} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|$ given $\widehat{\boldsymbol{\Sigma}}$. However, in (2), we solve $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|$ (FSPD) instead of (1), because its solution can be written explicitly. It would be of interest to know the price to pay of solving in the restricted class — to be specific, the difference between $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|$ and $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_1} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|$.

First, we find no cost in view of the spectral norm. Suppose that $\widehat{\gamma}_1$ and $\widehat{\gamma}_1^*$ are the minimum eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}^*$, respectively. We know that $\|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \geq \widehat{\gamma}_1^* - \widehat{\gamma}_1 \geq \epsilon - \widehat{\gamma}_1$ for all $\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0$, and thus $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \geq \epsilon - \widehat{\gamma}_1$. Recall that Lemma 2 implies that $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \leq \epsilon - \gamma_1$, and

$$\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \leq \min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_1} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \leq \min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 \leq \epsilon - \gamma_1$$

because $\mathcal{S}_2 \subseteq \mathcal{S}_1 \subseteq \mathcal{S}_0$. Therefore, $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 = \min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_1} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_2 = \epsilon - \widehat{\gamma}_1$. Thus, no additional cost is given by using the linear shrinkage constraint in spectral norm.

Second, in view of the Frobenius norm, we recall that Lemma 3 tells $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_2} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_F = (\epsilon - \widehat{\gamma}_1) \cdot \sqrt{\sum_{i=1}^p (\widehat{\gamma}_i - \overline{\widehat{\gamma}})^2} / \sqrt{\sum_{i=1}^p (\widehat{\gamma}_i - \widehat{\gamma}_1)^2}$, whereas, in general, $\min_{\widehat{\boldsymbol{\Sigma}}^* \in \mathcal{S}_0} \|\widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}\|_F = \sqrt{\sum_{i=1}^p (\epsilon - \widehat{\gamma}_i)_+^2 / p}$. The two bounds (minimums) depend on the eigenvalues of the true covariance matrix and the initial estimator in a complex way and hard to understand their closeness. For example, suppose we consider the FSPD estimator based on the soft thresholding estimator. A numerical investigation, not reported in this paper, shows the difference between two bounds decreases as p increases in all the scenarios of $\boldsymbol{\Sigma} = \mathbf{M}_2$ in Section 4. However, we do not find any special pattern in the scenarios of $\boldsymbol{\Sigma} = \mathbf{M}_1$.

Acknowledgement

References

[1] P. J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Stat. 36 (2008) 2577–2604. URL: http://projecteuclid.org/euclid.aos/1231165180. doi:10.1214/08-AOS600.

[2] P. J. Bickel, E. Levina, Regularized estimation of large covariance matrices, Ann. Stat. 36 (2008) 199–227. URL: http://projecteuclid.org/euclid.aos/1201877299. doi:10.1214/009053607000000758.

[3] J. Bien, R. J. Tibshirani, Sparse estimation of a covariance matrix, Biometrika 98 (2011) 807–820. doi:10.1093/biomet/asr054.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (2010) 1–122. URL: https://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=8186925. doi:10.1561/2200000016.

[5] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Amer. Stat. Assoc. 106 (2011) 672–684. URL: http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm10560. doi:10.1198/jasa.2011.tm10560.

[6] T. Cai, W. Liu, X. Luo, A constrained l1 minimization approach to sparse precision matrix estimation, J. Amer. Stat. Assoc. 106 (2011) 594–607. URL: http://www.tandfonline.com/doi/abs/10.1198/jasa.2011.tm10155. doi:10.1198/jasa.2011.tm10155. arXiv:1102.2233.

[7] T. T. Cai, M. Low, A framework for estimation of convex functions, Stat. Sinica 25 (2015) 423–456. URL: http://www3.stat.sinica.edu.tw/statistica/J25N2/J25N21/J25N21.html. doi:10.5705/ss.2013.279.

[8] T. T. Cai, Z. Ren, H. H. Zhou, Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation, Electron. J. Stat. 10 (2016) 1–59. URL: http://projecteuclid.org/euclid.ejs/1455715952. doi:10.1214/15-EJS1081.

[9] T. T. Cai, M. Yuan, Adaptive covariance matrix estimation through block thresholding, Ann. Stat. 40 (2012) 2014–2042. URL: http://projecteuclid.org/euclid.aos/1351602535. doi:10.1214/12-AOS999.

[10] T. T. Cai, C.-H. Zhang, H. H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Stat. 38 (2010) 2118–2144. URL: http://projecteuclid.org/euclid.aos/1278861244. doi:10.1214/09-AOS752.

[11] T. T. Cai, H. H. Zhou, Optimal rates of convergence for sparse covariance matrix estimation, Ann. Stat. 40 (2012) 2389–2420. URL: http://projecteuclid.org/euclid.aos/1359987525. doi:10.1214/12-AOS998.

[12] L. Chan, On portfolio optimization: forecasting covariances and choosing the risk model, Rev. Financial Stud. 12 (1999) 937–974. doi:10.1093/rfs/12.5.937.

[13] J. W. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia, PA, 1997.

[14] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Series B Stat. Methodol. 75 (2013) 603–680. URL: http://doi.wiley.com/10.1111/rssb.12016. doi:10.1111/rssb.12016.

[15] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2008) 432–441. URL: https://academic.oup.com/biostatistics/article/9/3/432/224260. doi:10.1093/biostatistics/kxm045.

[16] J. Friedman, T. Hastie, R. Tibshirani, Applications of the lasso and grouped lasso to the estimation of sparse graphical models, Technical Report, Stanford University, Stanford, CA, 2010. URL: statweb.stanford.edu/~tibs/ftp/ggraph.pdf.

[17] G. H. Golub, Q. Ye, An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems, SIAM J. Sci. Comput. 24 (2002) 312–334. URL: https://doi.org/10.1137/S1064827500382579. doi:10.1137/S1064827500382579.

[18] G. H. Golub, C. F. Van Loan, Matrix Computations, fourth ed., Johns Hopkins University Press, Baltimore, MD, 2012.

[19] R. Jagannathan, T. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, J. Finance 58 (2003) 1651–1683. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6261.00580. doi:10.1111/1540-6261.00580.

[20] K. Khare, S.-Y. Oh, B. Rajaratnam, A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees, J. R. Stat. Soc. Series B Stat. Methodol. 77 (2015) 803–825. URL: http://doi.wiley.com/10.1111/rssb.12088. doi:10.1111/rssb.12088. arXiv:1307.5381.

[21] Y. Kwon, Y.-G. Choi, T. Park, A. Ziegler, M. C. Paik, Generalized estimating equations with stabilized working correlation structure, Comput. Stat. Data Anal. 106 (2017) 1–11. URL: http://dx.doi.org/10.1016/j.csda.2016.08.016. doi:10.1016/j.csda.2016.08.016.

[22] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Ann. Stat. 37 (2009) 4254–4278. URL: https://projecteuclid.org/euclid.aos/1256303543. doi:10.1214/09-AOS720.

[23] G. R. Lanckriet, L. El Ghaoui, C. Bhattacharyya, M. I. Jordan, A robust minimax approach to classification, J. Mach. Learn Res. 3 (2002) 555–582. URL: http://www.jmlr.org/papers/volume3/lanckriet02a/lanckriet02a.pdf. doi:10.1162/153244303321897726.

[24] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. 88 (2004) 365–411. URL: http://www.sciencedirect.com/science/article/pii/S0047259X03000964. doi:10.1016/S0047-259X(03)00096-4.

[25] R. B. Lehoucq, D. C. Sorensen, Deflation techniques for an implicitly restarted arnoldi iteration, SIAM J. Matrix Anal. Appl. 17 (1996) 789–821. URL: https://doi.org/10.1137/S0895479895281484. doi:10.1137/S0895479895281484.

[26] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, Suitability of dysphonia measurements for telemonitoring of parkinson's disease, IEEE Trans. Biomed. Eng. 56 (2009) 1015–1022. URL: http://ieeexplore.ieee.org/document/4636708/. doi:10.1109/TBME.2008.2005954.

[27] H. Liu, L. Wang, T. Zhao, Sparse covariance matrix estimation with eigenvalue constraints, J. Comput. Graph. Stat. 23 (2014) 439–459. URL: http://www.tandfonline.com/doi/abs/10.1080/10618600.2013.782818. doi:10.1080/10618600.2013.782818.

[28] V. A. Marcenko, L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sbornik 1 (1967) 457–483. doi:10.1070/SM1967v001n04ABEH001994.

[29] R. Mazumder, T. Hastie, The graphical lasso: New insights and alternatives, Electron. J. Stat. 6 (2012) 2125–2149. URL: https://projecteuclid.org/euclid.ejs/1352470831. doi:10.1214/12-EJS740. arXiv:1111.5479.

[30] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Stat. 34 (2006) 1436–1462. URL: http://projecteuclid.org/euclid.aos/1152540754. doi:10.1214/009053606000000281. arXiv:0608017.

[31] J. Peng, P. Wang, N. Zhou, J. Zhu, Partial correlation estimation by joint sparse regression models, J. Amer. Stat. Assoc. 104 (2009) 735–746. doi:10.1198/jasa.2009.0126.

18

[32] A. J. Rothman, Positive definite estimators of large covariance matrices, Biometrika 99 (2012) 733–740. URL: https://academic.oup.com/biomet/article-abstract/99/3/733/359725. doi:10.1093/biomet/ass025.

[33] A. J. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, J. Amer. Stat. Assoc. 104 (2009) 177–186. URL: http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0101. doi:10.1198/jasa.2009.0101.

[34] D. C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, SIAM J. Matrix Anal. Appl. 13 (1990) 357–385. URL: https://epubs.siam.org/doi/10.1137/0613025. doi:10.1137/0613025.

[35] Y. Sun, S. Todorovic, S. Goodison, Local-learning-based feature selection for high-dimensional data analysis, IEEE Trans. Pattern. Anal. Mach. Intell. 32 (2010) 1610–1626. URL: https://dx.doi.org/10.1109/TPAMI.2009.190. doi:10.1109/TPAMI.2009.190.

[36] A. Tsanas, M. a. Little, C. Fox, L. O. Ramig, Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease, IEEE Trans. Neural. Syst. Rehabil. Eng. 22 (2014) 181–190. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6678640. doi:10.1109/TNSRE.2013.2293575.

[37] D. M. Witten, J. H. Friedman, N. Simon, New insights and faster computations for the graphical lasso, J. Comput. Graph. Stat. 20 (2011) 892–900. URL: http://www.tandfonline.com/doi/abs/10.1198/jcgs.2011.11051a. doi:10.1198/jcgs.2011.11051a.

[38] J.-H. Won, J. Lim, S.-J. Kim, B. Rajaratnam, Condition number regularized covariance estimation, J. R. Stat. Soc. Series B Stat. Methodol. 75 (2013) 427–450. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2012.01049.x/abstract. doi:10.1111/j.1467-9868.2012.01049.x.

[39] L. Xue, S. Ma, H. Zou, Positive-definite l1-penalized estimation of large covariance matrices, J. Amer. Stat. Assoc. 107 (2012) 1480–1491. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.725386. doi:10.1080/01621459.2012.725386.