High Dimensional Covariance Matrix Estimation

Clifford Lam*

Department of Statistics, London School of Economics and Political Science

Abstract

Covariance matrix estimation plays an important role in statistical analysis in many fields, including (but not limited to) portfolio allocation and risk management in finance, graphical modelling and clustering for genes discovery in bioinformatics, Kalman filtering and factor analysis in economics. In this paper, we give a selective review of covariance and precision matrix estimation when the matrix dimension can be diverging with, or even larger than the sample size. Two broad categories of regularization methods are presented. The first category exploits an assumed structure of the covariance or precision matrix for consistent estimation. The second category shrinks the eigenvalues of a sample covariance matrix, knowing from random matrix theory that such eigenvalues are biased from the population counterparts when the matrix dimension grows at the same rate as the sample size.

Key words and phrases. Structured covariance estimation, sparsity, low rank plus sparse, factor model, shrinkage.

[*] Clifford Lam is Associate Professor, Department of Statistics, London School of Economics. Email: C.Lam2@lse.ac.uk

1 Introduction

With tremendous technological advancement and increase in computational power over the past decade, it is easier than ever to obtain and analyse high dimensional data in different areas such as finance, economics, social science and health science. Various statistical procedures require the population covariance matrix $\mathbf{\Sigma}$, or the inverse $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ called the precision matrix, of a random sample (or stationary time series) of p-dimensional random vectors $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ as input. Assuming that $E(\mathbf{x}_i) = \mathbf{0}$ and $\text{var}(\mathbf{x}_t) = \mathbf{\Sigma}$ for $i = 1, \dots, n$, the sample covariance matrix is defined as $\mathbf{S} = n^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$, with $E(\mathbf{S}) = \mathbf{\Sigma}$. Despite its unbiasedness and simplicity, $\mathbf{S}$ is a poor estimator for $\mathbf{\Sigma}$ when p is large compared to the sample size n, in the sense that $p/n \to c \in (0, \infty)$. Marčenko and Pastur (1967) showed that for $\mathbf{\Sigma} = \mathbf{I}_p$, the $p \times p$ identity matrix, the empirical spectral density (ESD) − the distribution of the eigenvalues − of $\mathbf{S}$, does not converge to a single mass at 1 as we hoped for. Rather, it converges to a wildly different distribution, what we now called the Marčhenko-Pastur distribution. Moreover, the eigenvectors of $\mathbf{S}$, an important output from principal component analysis (PCA), can be far away from those of $\mathbf{\Sigma}$ (Johnstone and Lu, 2009, Ledoit and Péché, 2011).

The poor qualities of $\mathbf{S}$ lead searchers to look into various regularized covariance or precision matrix estimators in different applications. In general, structural assumptions on $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ are needed for consistent estimation. Methods include tapering (Furrer et al., 2006), banding (Bickel and Levina, 2008b), thresholding (Bickel and Levina, 2008a), penalization (Huang et al., 2006, Lam and Fan, 2009, Ravikumar et al., 2011, Rothman et al., 2008), modified cholesky decomposition (see 2.6) with regularization (Pan and Mackenzie, 2003, Pourahmadi, 2007, Rothman et al., 2010), graphical lasso (Friedman et al., 2008, Mazumder and Hastie, 2012), low rank plus sparse decomposition (Fan et al., 2008, 2013, Guo et al., 2017), to name but a few. Depending on applications, banded, sparse or low rank plus sparse assumptions on $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ can be realistic and useful in guiding us towards a good regularized estimator.

Another branch of estimators stems from assuming that $\mathbf{\Sigma}$ does not have diverging eigenvalues as $n, p \to \infty$. Focus is then not on estimators conforming to any structures on $\mathbf{\Sigma}$, but on shrinking the eigenvalues of $\mathbf{S}$. When $\mathbf{\Sigma} = \mathbf{I}_p$ and $p/n \to c > 0$, the smallest and largest eigenvalues of $\mathbf{S}$ are going to $\max(0, (1 - \sqrt{c})^2)$ and $(1 + \sqrt{c})^2$ respectively (Bai and Yin, 1993). This fact prompts us that shrinking the eigenvalues of $\mathbf{S}$ can be a good idea, especially when p is large compared to n in practice. Although not necessarily consistent for $\mathbf{\Sigma}$, shrinkage estimators are well-conditioned, and can improve drastically the performance of different procedures, such as portfolio allocation for example. In fact, the first shrinkage estimator of covariance matrix originates from Stein (1975, 1986). Ledoit and Wolf (2004) proposed a linear shrinkage estimator which shrinks the eigenvalues of $\mathbf{S}$ toward the identity matrix. Schäfer and Strimmer (2005) used the same shrinkage idea to shrink $\mathbf{S}$ to different known target matrices. Won et al. (2013) proposed an estimator which has the middle portion of the sample eigenvalues unchanged, but the more extreme eigenvalues winsorized at certain constants. Ledoit and Wolf (2012) proposed a nonlinear shrinkage formula for shrinking each eigenvalue in $\mathbf{S}$ nonlinearly so as to minimize a Frobenius error loss. Abadir et al. (2014) proposed a model free regularized estimator using a data splitting scheme, which Lam (2016) proved to be a nonparametric way in achieving the nonlinear shrinkage in Ledoit and Wolf (2012), and gave a theoretically supported data splitting scheme for asymptotic efficiency. Lam et al. (2017) and Lam and Feng (2018) used similar ideas to construct well-conditioned integrate volatility matrix estimators for intraday and high frequency tick-by-tick data respectively, demonstrating theoretically how the minimum variance portfolio can be benefitted. Donoho et al. (2018) proved that different loss functions can lead to completely different shrinkage formulae for the sample eigenvalues in a spiked covariance model, and worked out such formulae for various loss functions. Engle et al. (2017) proposed to use nonlinearly shrinkage technique to construct a dynamic covariance matrix estimator.

Review of high dimensional covariance matrix estimation has also been done in the past. See two nice reviews by Cai et al. (2016b) and Fan et al. (2016), with the former focused more on minimax adaptive estimations and related theoretical properties and bounds, while the latter focused on regularization methods leading to consistent estimation, including thresholding, penalized likelihood and factor-based methods, with discussions on robust estimation as well. The book by Pourahmadi (2013) adds an excellent account on many recent developments of the field, including shrinkage methods which will be discussed in this paper as well. Due to limited space, we do not include Bayesian related methods, which is a large field of study in its own right.

The rest of the paper is organized as follows. In Section 2 we give a selective review on some methods in structured covariance and precision matrix estimation, followed by shrinkage estimation in Section 3.

2 Structured Covariance Matrix Estimation

We present different estimators categorized by their structural assumptions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$. We denote x_{ki} the ith element of $\mathbf{x}_k$, $k = 1, \ldots, n,\ i = 1, \ldots, p$. We use $\mathbf{S}$, where

$$\mathbf{S} = n^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^{\mathrm{T}}, \quad \bar{\mathbf{x}} = n^{-1}\sum_{k=1}^{n}\mathbf{x}_k,$$

as the sample covariance matrix for the rest of the paper.

2.1 Receding off-diagonals

If there is a natural order in the elements in $\mathbf{x}_i$, for example $\mathbf{x}_i$ contains spatial or temporal variables, then it is natural that the off-diagonal elements in $\mathbf{\Sigma} = \text{var}(\mathbf{x}_i)$ are decreasing in magnitude as they are further from the main diagonal. These are also called bandable covariance matrices, since beyond a certain off-diagonal, elements are so small that we can regularize by setting the bands of those off-diagonals to 0. With this idea, for $\mathbf{\Sigma} = (\sigma_{ij})$, Bickel and Levina (2008b) introduced the following class of covariance matrices:

$$\begin{aligned}\mathcal{U}_\alpha(M_0, M) = \Big\{\mathbf{\Sigma} : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k > 0, \\ \text{and } 0 < M_0^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq M_0\Big\}, \qquad (2.1)\end{aligned}$$

where $\lambda_{\min}(\cdot)$, $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues of a matrix. They then propose to band the sample covariance matrix $\mathbf{S} = (s_{ij})$. Find k with $0 \leq k < p$, the banded estimator is

$$\widehat{\mathbf{\Sigma}}_k = B_k(\mathbf{S}) := (s_{ij}\mathbf{1}(|i-j| \leq k)), \qquad (2.2)$$

where $\mathbf{1}(\cdot)$ is an indicator function. They prove that if $k = k_n \asymp (n^{-1}\log p)^{-1/(2(\alpha+1))}$, then uniformly over the class $\mathcal{U}_\alpha$, for Gaussian or light-tailed data,

$$\big\|\widehat{\mathbf{\Sigma}}_{k_n} - \mathbf{\Sigma}\big\| = O_P\Big(\Big(\frac{\log p}{n}\Big)^{\alpha/(2(\alpha+1))}\Big) = \big\|\widehat{\mathbf{\Sigma}}_{k_n}^{-1} - \mathbf{\Sigma}^{-1}\big\|, \qquad (2.3)$$

where $\|\cdot\|$ denotes the spectral or L_2 norm of a matrix. The notation $a \asymp b$ means $a = O(b)$ and $b = O(a)$. They also introduce how to band the inverse and provided a parallel theorem, which is connected to banding the Cholesky factor from a modified Cholesky decomposition of $\mathbf{\Sigma}$. See their paper for more details, for instance how to determine k numerically.

Under the assumption of known rate of decay α in the class $\mathcal{U}_\alpha$ in (2.1), for Gaussian data, Cai et al. (2010) proposed a tapering estimator

$$\check{\mathbf{\Sigma}}_k = T_k(\mathbf{S}) := \Big(\frac{2s_{ij}}{k}\{(k - |i-j|)_+ - (k/2 - |i-j|)_+\}\Big),$$

where $(x)_+ = \max(x, 0)$. With $k \asymp n^{1/(2\alpha+1)}$, they showed that uniformly over $\mathcal{U}_\alpha$,

$$\big\|\check{\mathbf{\Sigma}}_k - \mathbf{\Sigma}\big\| = O_P\Big(n^{-\alpha/(2\alpha+1)} + \Big(\frac{\log p}{n}\Big)^{1/2}\Big),$$

which is always faster than the rate in (2.3). Moreover, they showed that this rate is minimax optimal over $\mathcal{U}_\alpha$.

To overcome the impracticality of assuming α is known, Cai and Yuan (2012) proposed to use a block thresholding scheme to achieve adaptive rate optimal estimation for Gaussian data over the parameter spaces $\mathcal{U}_\alpha$ for all $\alpha > 0$. For the detailed construction of such blocks, see their paper. With the blocks constructed, they then propose to threshold each block in $\mathbf{S}$, which has an adaptive level of threshold with a universal constant needed to be found. They then prove that their block-thresholded estimator $\widehat{\boldsymbol{\Sigma}}$ has

$$\sup_{\boldsymbol{\Sigma}\in\mathcal{U}_\alpha(M_0,M)} E\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|^2 \le C\min\left\{n^{-2\alpha/(2\alpha+1)}+\frac{\log p}{n},\frac{p}{n}\right\}$$

for all $\alpha > 0$, where C is a positive constant independent of n and p. This estimator is optimally rate adaptive over $\mathcal{U}_\alpha$ for all $\alpha > 0$.

Bien et al. (2016) argued that directly killing off blocks that are far from the off-diagonal as done in Cai and Yuan (2012) may not be data-adaptive enough. They proposed to use a hierarchical group lasso penalty for estimating $\boldsymbol{\Sigma}$. Define s_m to be the set of all pair of indices corresponding to the $p-m$th off-diagonals,i.e.,

$$s_m = \{(i,j) : |i-j| = p-m\}, \quad m = 1,\dots,p-1.$$

For an index set g, define $\boldsymbol{\Sigma}_g$ to be a vector of length $|g|$ of the corresponding elements in $\boldsymbol{\Sigma}$. Then, with $\|\cdot\|_F$ denoting the Frobenius norm of a matrix, Bien et al. (2016) proposed to solve

$$\min_{\boldsymbol{\Sigma}}\left\{\frac{1}{2}\left\|\boldsymbol{\Sigma}-\mathbf{S}\right\|_F^2+\sum_{\ell=1}^{p-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2\left\|\boldsymbol{\Sigma}_{s_m}\right\|^2}\right\}, \quad w_{\ell m} = \frac{\sqrt{2\ell}}{\ell-m+1},\ 1\le m\le \ell\le p-1.$$

They proposed to solve the above penalized hierarchical group lasso problem by solving its dual using the block coordinate descent. They also proved the Frobenius error rate of convergence as well as operator norm rate, with the Frobenius error rate being minimax adaptive up to multiplicative logarithmic factors over a class that generalizes approximate banded and K-banded matrices.

For robust estimation of bandable correlation matrices, Xue and Zou (2014) proposed to use a nonparanormal model for $\mathbf{x}$, with p monotonically increasing transformations for all p variables in $\mathbf{x}$ such that the resulting vector

$$\mathbf{f}(\mathbf{x}) = (f_1(x_1),\dots,f_p(x_p)) \sim N(\mathbf{0},\boldsymbol{\Sigma}_f), \tag{2.4}$$

where $\boldsymbol{\Sigma}_f$ is a correlation matrix. An important observation is that x_i and x_j are marginally independent if and only if $(\boldsymbol{\Sigma}_f)_{ij} = 0$, and hence a banded correlation matrix for $\mathbf{x}$ should results in a correlation matrix of the same banded pattern for $\mathbf{\Sigma}_f$. To estimate $\mathbf{\Sigma}_f$ (not the correlation of $\mathbf{x}$ itself), they proposed to use the Spearman's rank correlation coefficient $\widehat{r}_{ij}$. The key observation is that this $\widehat{r}_{ij}$ is the same for the (unknown) transformed data since the transformations are all monotonically increasing. In the end, they used a classical result from Kendall (1948) to arrive at $\widehat{\mathbf{R}}^s = (\widehat{r}^s_{ij})$, where $\widehat{r}^s_{ij} = 2\sin(\pi\widehat{r}_{ij}/6)$, as a first-step estimator for $\mathbf{\Sigma}_f$, which can be a poor estimator when the dimension p is large. They then proposed to estimate $\mathbf{\Sigma}_f$ by the regularization

$$\widehat{\mathbf{R}}^s_{gt} = (\widehat{r}^s_{ij} w_{ij})_{1\le i,j\le p}, \tag{2.5}$$

where $w_{ij} = 1$ for $|i-j| \le \lfloor k/2 \rfloor$, $w_{ij} = 0$ for $|i-j| > k$, and $0 \le w_{ij} \le 1$ for $\lfloor k/2 \rfloor < |i-j| \le k$. These are either tapering or banding weights, where k is chosen by cross-validation. They showed nice theoretical properties of $\widehat{\mathbf{R}}^s_{gt}$ as an estimator of $\mathbf{\Sigma}_f$.

Another robust estimation is proposed by Chen et al. (2018). Consider independent and identically distributed data vectors $\mathbf{x}_i$ from the ϵ-contamination model

$$P_{\epsilon,\mathbf{\Sigma},Q} = (1-\epsilon)\mathbf{P}_{\mathbf{\Sigma}} + \epsilon Q,$$

where $P_{\mathbf{\Sigma}} = N(\mathbf{0}, \mathbf{\Sigma})$, and Q is any distribution. Essentially, ϵ can be interpreted as the proportion of "outlying" data, so that the number of "outliers" from $P_{\mathbf{\Sigma}}$ is then $n\epsilon$ for a sample size of n. To estimate $\mathbf{\Sigma}$, they propose the concept of matrix depth which is inspired by the Tukey's median. The matrix depth of a positive semi-definite $\mathbf{\Gamma} \in \mathbb{R}^{p\times p}$ with respect to the distribution P restricted over a subset $\mathcal{U} \subset \mathcal{S}^{p-1}$ is defined as

$$\mathcal{D}_{\mathcal{U}}(\mathbf{\Gamma}, P) = \inf_{\mathbf{u}\in\mathcal{U}} \min\{P(|\mathbf{u}^{\mathrm{T}}\mathbf{x}|^2 \le \mathbf{u}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}), 1 - P(|\mathbf{u}^{\mathrm{T}}\mathbf{x}|^2 < \mathbf{u}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u})\}.$$

Clearly, the maximum value of depth is $1/2$ by the above definition. Chen et al. (2018) showed that in fact, for any $\mathcal{U} \in \mathcal{S}^{p-1}$, $\mathcal{D}_{\mathcal{U}}(\beta\mathbf{\Sigma}, P_{\mathbf{\Sigma}}) = 1/2$, where β is such that $\Phi(\sqrt{\beta}) = 3/4$. This inspires the authors to estimate $\mathbf{\Sigma}$ by

$$\begin{aligned}\widehat{\mathbf{\Sigma}} &= \arg\max_{\mathbf{\Gamma}\in\mathcal{F}} \mathcal{D}_{\mathcal{U}}(\mathbf{\Gamma}, \{\mathbf{x}_i\}_{i=1}^n)/\beta \\ &= \arg\max_{\mathbf{\Gamma}\in\mathcal{F}} \min_{\mathbf{u}\in\mathcal{U}} \min\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\mathbf{u}^{\mathrm{T}}\mathbf{x}_i|^2 \le \mathbf{u}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}\}, 1 - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\mathbf{u}^{\mathrm{T}}\mathbf{x}_i|^2 < \mathbf{u}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{u}\}\right\}/\beta,\end{aligned}$$

where $\mathcal{F}$ is a matrix class which can impose various structures on $\mathbf{\Sigma}$ for practical estimation. One class they considered is the bandable class $\mathcal{F}_k = \{\mathbf{\Sigma} \succeq 0 : (\mathbf{\Sigma})_{ij} = 0 \text{ if } |i-j| > k\}$, and they showed for $\epsilon < 1/5$, $0 < \delta < 1/2$,

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 \le C\left(\frac{k + \log p}{n} \vee \epsilon^2 + \frac{\log(1/\delta)}{n}\right)$$

for some constant $C > 0$, with $P_{\epsilon,\mathbf{\Sigma},Q}$-probability at least $1-2\delta$ uniformly over all Q and $\mathbf{\Sigma} \in \mathcal{F}_k$ such that $\mathbf{\Sigma}$ has uniformly bounded eigenvalues. These results are also available for sparse covariance estimation, and can even be extended to elliptical distribution with fat tails.

The topic on bandwidth selection for estimating bandable covariance matrix is also well-studied. Qiu and Chen (2012) proposed a criterion leading to consistent estimation of the banding parameter, while both Li and Zou (2016) and Li et al. (2018) analyzed the Stein's unbiased risk estimation (SURE) information criterion for a class of bandable covariance matrices, with minimizing such a criterion resulting in consistent estimation of the tuning parameter. Please refer to these papers for more details of their methods and theoretical results.

2.2 Sparsity

The assumption of receding off-diagonals in Section 2.1 is a special case of general sparsity of $\mathbf{\Sigma}$. Sparsity can also be in the precision matrix $\mathbf{\Omega}$, or even in other decomposed components of $\mathbf{\Sigma}$ using appropriate decompositions. In this section, we only present theoretical properties for covariance and precision matrix estimators that are not relating to support recovery of $\mathbf{\Omega}$, which is the core topic in graphical models for the next section.

Huang et al. (2006) proposed to use the modified Cholesky decomposition for penalized likelihood construction. The modified Cholesky decomposition of $\mathbf{\Sigma}$, with elements uniquely defined, is

$$\mathbf{T}\mathbf{\Sigma}\mathbf{T}^{\mathrm{T}} = \mathbf{D}, \tag{2.6}$$

where $\mathbf{D}$ is diagonal, and $\mathbf{T}$ is a unit lower-triangular matrix having ones on its diagonal. If $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ has $\mathrm{var}(\mathbf{x}) = \mathbf{\Sigma}$, then (2.6) means that we can always decompose

$$x_t = \sum_{j=1}^{t-1} \phi_{tj} x_j + \epsilon_t, \tag{2.7}$$

where $-\phi_{tj}$ is the (t, j)-th element of $\mathbf{T}$ for $2 \le t \le p$ and $j = 1, \ldots, t-1$. Also, $\mathrm{var}(\boldsymbol{\epsilon}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2) = \mathbf{D}$, with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^{\mathrm{T}}$ the vector of successive prediction errors. The above means that $\mathbf{T}\mathbf{x} = \boldsymbol{\epsilon}$, and taking variance on both sides we get back (2.6).

The idea of sparsity in the Cholesky factor $\mathbf{T}$ in (2.6) comes from that if the elements in the data vector $\mathbf{x}$ are ordered, then in the face of (2.7), ϕ_{tj} should be close to 0 as t and j are further apart. With

(2.6), Huang et al. (2006) proposed to minimize the penalized likelihood

$$n\text{log}|\mathbf{D}| + \sum_{i=1}^{n} \mathbf{x}_i \mathbf{T}^{\mathrm{T}} \mathbf{D}^{-1} \mathbf{T} \mathbf{x}_i + \lambda p(\{\phi_{tj}\}),$$

where $p(\cdot)$ is a penalizing function to be set by the user. Selection of λ is by cross-validation and a practical algorithm is discussed, but no theoretical results on the estimators are given.

Bien and Tibshirani (2011) studied sparse estimation of $\mathbf{\Sigma}$ through

$$\min_{\mathbf{\Sigma} \succ 0} \log\det(\mathbf{\Sigma}) + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) + \lambda \big\|\mathbf{P} \circ \mathbf{\Sigma}\big\|_1.$$

where $\mathbf{P}$ is an arbitrary constant matrix and $\circ$ is the elementwise multiplication. The choice of $\mathbf{P}$ is problem dependent. They propose a majorization-minimization approach to solving the above, with the idea to convert solving a non-convex minimization problem to solving a series of simpler convex minimization problems. See their paper for more details.

Bickel and Levina (2008a) presented a class of sparse covariance matrix,

$$\mathcal{C}_q(c_0(p), M, M_0) = \Big\{\mathbf{\Sigma} : \sigma_{ii} \le M, \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p), \text{ for all } i, \ \lambda_{\min}(\mathbf{\Sigma}) \ge M_0 > 0\Big\}, \tag{2.8}$$

where $0 \le q < 1$. If $q = 0$, with the convention $0^0 = 0$, it is a class of exactly sparse matrix. The thresholded matrix estimator is defined as

$$T_\lambda(\mathbf{S}) := (s_{ij}\mathbf{1}(|s_{ij}| \ge \lambda)). \tag{2.9}$$

They showed that uniformly on $\mathcal{C}_q$, if M' is sufficient large and $n^{-1}\text{log}\, p = o(1)$, then

$$\big\|T_{\lambda_n}(\mathbf{S}) - \mathbf{\Sigma}\big\| = O_P\Big(c_0(p)\Big(\frac{\log p}{n}\Big)^{(1-q)/2}\Big) = \big\|(T_{\lambda_n}(\mathbf{S}))^{-1} - \mathbf{\Omega}\big\|, \quad \lambda_n = M'\sqrt{\frac{\log p}{n}}.$$

The above estimator is obtained using a universal threshold λ_n for all elements in $\mathbf{S}$.

Cai and Liu (2011) proposed to use an adaptive threshold λ_{ij} for their estimator $\widehat{\mathbf{\Sigma}}^* = (\widehat{\sigma}^*_{ij})$, with

$$\widehat{\sigma}^*_{ij} = g_{\lambda_{ij}}(s_{ij}), \tag{2.10}$$

where $g_\lambda(\cdot)$ is a general thresholding function introduced in Rothman et al. (2009). The hard-thresholding function used in Bickel and Levina (2008a) is a special case of $g_\lambda(\cdot)$. The adaptive threshold is defined as

$$\lambda_{ij} = \delta\sqrt{\frac{\widehat{\theta}_{ij}\log p}{n}},$$

where δ is a tuning parameter, and $\widehat{\theta}_{ij}$ is proposed as

$$\widehat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\Big[(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)-s_{ij}\Big]^2, \quad \bar{x}_i = \frac{1}{n}\sum_{k=1}^{n}x_{ki},$$

which is an estimator of

$$\theta_{ij} = \mathrm{var}[(x_i - Ex_i)(x_j - Ex_j)].$$

Define a class $\mathcal{C}_q^*$ which is larger than $\mathcal{C}_q$ in (2.8),

$$\mathcal{C}_q^* = \Big\{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} \text{ positive definite, } \max_i \sum_{j=1}^{p}(\sigma_{ii}\sigma_{jj})^{(1-q)/2}|\sigma_{ij}|^q \le c_0(p)\Big\}.$$

Then for $\delta \ge 2$ and $0 \le q < 1$, they showed that $\widehat{\boldsymbol{\Sigma}}^*$ has a rate of convergence of $c_0(p)(n^{-1}\log p)^{(1-q)/2}$ uniformly over $\mathcal{C}_q^*$, and the data can even have polynomial-type tails.

Although thresholding the Cholesky factor $\mathbf{T}$ in the modified Cholesky decomposition in (2.6) guarantees positive definiteness of the estimator, the variables in $\mathbf{x}$ does need a certain kind of ordering for (2.7) to produce a sparse $\mathbf{T}$. At the same time, the estimators in Bickel and Levina (2008b) and Cai and Liu (2011) may not even be positive semi-definite for finite sample. In the face of this problem, Xue et al. (2012) proposed an alternating direction algorithm for solving

$$\min_{\boldsymbol{\Sigma}\succeq\epsilon\mathbf{I}_p} \frac{1}{2}\big\|\boldsymbol{\Sigma}-\mathbf{S}\big\|_F^2 + \lambda\big\|\boldsymbol{\Sigma}\big\|_1,$$

where $\boldsymbol{\Sigma}\succeq\epsilon\mathbf{I}_p$ means that $\boldsymbol{\Sigma}-\epsilon\mathbf{I}_p$ is positive semi-definite, and $\big\|\boldsymbol{\Sigma}\big\|_1$ denotes the sum of absolute values of $\boldsymbol{\Sigma}$.

More recent works on sparse estimation of high dimensional covariance matrix attempt to bridge patterned sparsity (like bandedness) and non-patterned sparsity. Bien (2019) proposed a graph-guided banding estimator with global or local bandwidth. The idea is to view a covariance matrix as a linear combination of matrices with “graph-guided” sparsity patterns. Interested readers are referred to the respective paper for further details.

2.3 Graphical model

This is related to the fact that if the data $\mathbf{x} = (x_1, \ldots, x_p)$ is Gaussian with $\mathbf{\Omega} = (\omega_{ij})$, then $\omega_{ij} = 0$ if and only if x_i is conditionally independent of x_j given all the remaining variables in $\mathbf{x}$. In a graph, it means that x_i is only connected to x_j through other variables but not directly. In this sense, a sparse graph means a sparse $\mathbf{\Omega}$ and vice versa. Hence in graphical models, the most important aspect of an estimator $\widehat{\mathbf{\Omega}}$ of $\mathbf{\Omega}$ is that the connectedness represented in $\widehat{\mathbf{\Omega}}$ is as close to that in $\mathbf{\Omega}$ as possible. In mathematical terms, we want ideally probability 1 for the event $\widehat{\omega}_{ij} = 0$ when $\omega_{ij} = 0$ and $\widehat{\omega}_{ij} \neq 0$ when $\omega_{ij} \neq 0$ for all i, j. In this section, theoretical results presented are hence focused on this event with probability going to 1, which sometimes is termed as graph selection consistency. If papers are only concerned with other consistency results of $\widehat{\mathbf{\Omega}}$ such as the Frobenius or spectral norms consistency, they will not be presented in this section.

Suppose $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a Gaussian random sample with $E(\mathbf{x}_i) = \mathbf{0}$ and $\text{var}(\mathbf{x}_i) = \mathbf{\Sigma}$. In the paper by Meinshausen and Bühlmann (2006), they proposed to estimate the graph implied by $\mathbf{\Omega}$ using the lasso, by regressing the ith variable on the rest of them, for each i, penalized by a tuning parameter λ. If the estimated coefficient for the jth variable on i, or the estimated coefficient for the i variable on j, is non-zero, the (i, j)th entry in $\mathbf{\Omega}$ is estimated as non-zero (they also consider requiring both to be non-zero). They prove the method to be consistent in estimating the set of zeros in $\mathbf{\Omega}$ under certain assumptions, in the sense that for each node of a graph a, defining ne_a to be the true neighbourhood of a (i.e., set of variables that are conditionally dependent on a), and $\widehat{\text{ne}}_a^\lambda$ the estimated neighbourhood using parameter λ, then

$$P(\widehat{\text{ne}}_a^\lambda \subseteq \text{ne}_a) = 1 - O(\exp(-cn^\epsilon)) = P(\text{ne}_a \subseteq \widehat{\text{ne}}_a^\lambda),$$

where $c > 0$ is a constant, λ has the same order as $n^{-(1-\epsilon)/2}$ for some bounded ϵ.

Yuan and Lin (2007) proposed to solve the following restricted L_1 regularized negative log-likelihood problem:

$$\min_{\mathbf{\Theta} \succ 0} -\log\det(\mathbf{\Theta}) + \text{tr}(\mathbf{S\Theta}) + \lambda \big\|\mathbf{\Theta}\big\|_1. \tag{2.11}$$

They considered solving (2.11) by an interior point method called maxdet algorithm in convex maximization. They also proved a result on the asymptotic distribution of the estimator itself for the above lasso-type estimator.

For consistent neighbourhood estimation, they proposed to solve for a nonnegative garrote-type estimator $\widehat{\mathbf{\Theta}}$ by considering

$$\min_{\mathbf{\Theta} \succ 0} -\log\det(\mathbf{\Theta}) + \text{tr}(\mathbf{S\Theta}) + \lambda \sum_{i \neq j} \frac{\theta_{ij}}{\widetilde{\theta}_{ij}},$$

where $\widetilde{\mathbf{\Theta}}$ is an initial estimator. They consider $\widetilde{\mathbf{\Theta}} = \mathbf{S}^{-1}$, implicitly assuming $p < n$. With $\mathbf{S}^{-1}$ as the initial estimator, and p fixed while $n \to \infty$, they proved that $P(\widehat{\theta}_{ij} = 0) \to 1$ if $\omega_{ij} = 0$, and other elements of $\mathbf{\Omega}$ have the same limiting distribution as the maximum likelihood estimator on the true graph structure. Friedman et al. (2008) proposed an efficient algorithm, called the graphical lasso, to solve (2.11). They use a framework developed by Banerjee et al. (2008) which considers the dual problem of (2.11) as a start, and arrive at solving an L_1 penalized regression problem.

Since a graphical model needs normality to infer a sparse graph from a sparse $\mathbf{\Omega}$, Liu et al. (2009) used the nonparanormal model in (2.4), and proposed to estimate those transformations f_i from data. Then they replace $\mathbf{S}$ in (2.11) by the sample covariance matrix of the transformed data, and use the graphical lasso to solve for an estimator of $\mathbf{\Omega}$, which is in fact an estimator of $\mathbf{\Omega}_f = \mathbf{\Sigma}_f^{-1}$ in the notation of (2.4). The key observation is that $\mathbf{\Omega}_f$ retains the sparsity pattern of $\mathbf{\Omega}$, i.e., $\text{sign}((\mathbf{\Omega}_f)_{ij}) = \text{sign}((\mathbf{\Omega})_{ij})$, and hence we can then infer the graph of $\mathbf{x}$. They prove that under certain assumptions, the estimator achieves sign consistency, in the sense that

$$P(\text{sign}(\widehat{\mathbf{\Omega}}_f)_{ij} = \text{sign}(\mathbf{\Omega}_f)_{ij}) \geq 1 - o(1), \quad \lambda \asymp \sqrt{\frac{\log p \log^2 n}{n^{1/2}}},$$

where λ is the penalization parameter used in (2.11).

Lam and Fan (2009) proposed Gaussian penalized quasi-likelihood in the form

$$q(\mathbf{\Omega}) = \text{tr}(\mathbf{S\Omega}) - \log|\mathbf{\Omega}| + \sum_{i \neq j} p_\lambda(\omega_{ij}).$$

This is essentially the same as (2.11) apart from the penalty function being the SCAD penalty introduced in Fan and Li (2001), which is a nonconvex penalty designed to overcome the general bias problems associated with the L_1 penalty in lasso. The resulting estimator by minimizing $q(\mathbf{\Omega})$ with respect to $\mathbf{\Omega}$ is proved to be consistent to the true sparse precision matrix in Frobenius norm under certain conditions. Further theoretical results give rates of convergence and sparsistency – zero elements are estimated as 0 with probability going to 1. This is not complete graph selection consistency but at least non-existence of edges in a graph is identified. One highlight of the paper however, is that in using the L_1 penalty when the elements under penalization are in fact large, convergence of estimators (in Frobenius norm) is only guaranteed when $\mathbf{\Omega}$ is very sparse. On the other hand, using "unbiased" penalty functions like hard-thresholding or SCAD, $\mathbf{\Omega}$ does not need to be as sparse. For detailed rates and algorithm please refer to the paper itself.

Cai et al. (2011) proposed the constrained L_1 minimization for inverse matrix estimation (CLIME), which considered the problem

$$\min \big\|\boldsymbol{\Omega}\big\|_1 \text{ subject to } \big\|\mathbf{S}\boldsymbol{\Omega} - \mathbf{I}_p\big\|_{\max} \le \lambda_n, \tag{2.12}$$

where $\|\cdot\|_{\max}$ denotes the maximum absolute element of a matrix. The above can be decomposed into p convex optimization problems by solving for $\boldsymbol{\Omega}$ column by column, using e.g. linear programming. They spelt out the rate of convergence of the resulting estimator $\widehat{\boldsymbol{\Omega}}$ under spectral, Frobenius and ℓ_∞ norms for both exponential-type and polynomial-type tails for the data as well. Relating to graph selection consistency, they also proved that under the $s_0(p)$-sparse precision matrix class

$$\mathcal{U} = \{\boldsymbol{\Omega} : \boldsymbol{\Omega} \succ 0, \big\|\boldsymbol{\Omega}\big\|_1 \le M, \max_{1\le i\le p} \sum_{j=1}^{p} \mathbf{1}\{\omega_{ij} \ne 0\} \le s_0(p)\},$$

the thresholded estimator $T_{\tau_n}(\widehat{\boldsymbol{\Omega}})$ (see (2.9)), where $\tau_n \ge C\sqrt{\log p/n}$ for some $C > 0$, achieves sign consistency with rate $1 - O(n^{-\delta/8} + p^{-\tau/2})$ for some constants δ and τ. Cai et al. (2016a) improved the method to ACLIME, with an adaptive threshold, and proved minimax optimal rate of convergence uniformly over large classes of approximately sparse precision matrices. They also proposed a thresholded estimator like $T_{\tau_n}(\widehat{\boldsymbol{\Omega}})$ that achieves graph selection consistency, but is not presented as formal theorem.

Xue and Zou (2012) proposed to use the nonparanormal model (2.4) for $\mathbf{x}$. The technique is the same as the tapering estimation described before (2.5), so that using the Spearman's rank correlation coefficient $\widehat{r}_{ij}$, they first estimate $\boldsymbol{\Sigma}_f$ by $\widehat{\mathbf{R}}^s = (\widehat{r}^s_{ij})$, where $\widehat{r}^s_{ij} = 2\sin(\pi\widehat{r}_{ij}/6)$. Then the sparse $\boldsymbol{\Omega}_f$ can be estimated with, e.g., CLIME as in (2.12) with $\mathbf{S}$ replaced by $\widehat{\mathbf{R}}^s$. Theoretical results are given as well, showing that a properly chosen λ in the CLIME step help achieve sign consistency. See also Liu et al. (2012) which used the same nonparanormal model idea and proposed estimating $\boldsymbol{\Omega}_f$ using CLIME as well as Dantzig selector after obtaining $\widehat{\mathbf{R}}^s$.

Chandrasekaran et al. (2012) assumed that there are unknown number of latent variables and we only have the observed data. They proposed to split the precision matrix for the observed variables into a "sparse minus low rank" representation, and proposed estimators for each of them using a form of penalized log-likelihood for both. The sparse representation is in fact an estimator for the inverse of covariance of the observed variables conditional on the latent ones. Theoretical results are also presented. Ma et al. (2013) proposed to solve the optimization problem in Chandrasekaran et al. (2012) using two alternating direction methods, with global convergence proved. A more general low rank plus sparse representation for covariance matrix is presented in the next section.

2.4 Low rank plus sparse

We focus on covariance structure induced by a factor model in this paper. Ross (1976) introduced the strict factor model to use a small number of factors to explain a large number of returns. Write

$$\mathbf{x}_i = \mathbf{A}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2.13}$$

where $\mathbf{A}$ is a $p \times r$ factor loadings matrix, $\mathbf{f}_i$ is an $r \times 1$ vector of factors and $\boldsymbol{\epsilon}_i$ is a vector of (idiosyncratic) noise. Assuming r to be much smaller than p, the dynamics of the p components in $\mathbf{x}_i$ can then be summarized by the dynamics of the small number of factors $\mathbf{f}_i$. Assuming uncorrelatedness between $\mathbf{f}_i$ and $\boldsymbol{\epsilon}_i$, the covariance matrix for $\mathbf{x}_i$ is then

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{x}_i) = \mathbf{A}\boldsymbol{\Sigma}_f\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_\epsilon, \quad \boldsymbol{\Sigma}_f = \mathrm{var}(\mathbf{f}_i), \quad \boldsymbol{\Sigma}_\epsilon = \mathrm{var}(\boldsymbol{\epsilon}_i). \tag{2.14}$$

A strict factor model in Ross (1976) assumes that $\boldsymbol{\Sigma}_\epsilon$ is diagonal. Hence $\boldsymbol{\Sigma}$ in (2.14) is of a low rank ($\mathbf{A}\boldsymbol{\Sigma}_f\mathbf{A}^{\mathrm{T}}$, of rank r) plus sparse ($\boldsymbol{\Sigma}_\epsilon$) structure. Chamberlain and Rothschild (1983) relaxed the strict factor model to an approximate one, where $\boldsymbol{\Sigma}_\epsilon$ is sparse rather than diagonal. These papers are not focused on covariance estimation though.

With known factors, Fan et al. (2008) estimated $\mathbf{A}$ using least squares, and $\boldsymbol{\Sigma}_\epsilon$ is estimated using the estimated residuals $\widehat{\boldsymbol{\epsilon}}_i$, with diagonalized $\widehat{\boldsymbol{\Sigma}}_\epsilon = n^{-1}\sum_{i=1}^n \mathrm{diag}(\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}})$. Under the prospect of both $n, p \to \infty$, they established rates of convergence with respective to various loss functions, including Frobenius, spectral, Stein, and a re-scaled quadratic loss. Still with known factors, Fan et al. (2019) proposed first to estimate

$$\boldsymbol{\Sigma}_z := \mathrm{var}[(\mathbf{x}_i^{\mathrm{T}}, \mathbf{f}_i^{\mathrm{T}})^{\mathrm{T}}] = \begin{pmatrix} \mathbf{A}\boldsymbol{\Sigma}_f\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_\epsilon & \mathbf{A}\boldsymbol{\Sigma}_f \\ \boldsymbol{\Sigma}_f\mathbf{A}^{\mathrm{T}} & \boldsymbol{\Sigma}_f \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Because of potential heavy-tailed distributions for $\mathbf{x}_i$ and $\mathbf{f}_i$, each element in $\boldsymbol{\Sigma}_z$ is estimated as

$$(\widehat{\boldsymbol{\Sigma}}_z)_{ij} := \underset{x}{\mathrm{argmin}} \sum_{t=1}^n l_\alpha(z_{it}z_{jt} - x), \tag{2.15}$$

where $l_\alpha(x)$ is the Huber loss, defined by

$$l_\alpha(x) = \begin{cases} 2\alpha|x| - \alpha^2, & |x| > \alpha; \\ x^2, & |x| \le \alpha. \end{cases} \tag{2.16}$$

With $\widehat{\boldsymbol{\Sigma}}_z$, $\mathbf{A}\boldsymbol{\Sigma}_f\mathbf{A}^{\mathrm{T}}$ can be estimated by $\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21}$, and $\widehat{\boldsymbol{\Sigma}}_\epsilon$ can be obtained by diagonalizing or thresholding $\widehat{\mathbf{\Sigma}}_z - \widehat{\mathbf{\Sigma}}_{12}\widehat{\mathbf{\Sigma}}_{22}^{-1}\widehat{\mathbf{\Sigma}}_{21}$. An appropriately diverging α results in good properties of the estimators, with rates of convergence provided.

When the factors are unknown, the asymptotic PCA method is to solve (Bai and Ng, 2002)

$$\min_{\mathbf{A}, f_i} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{A}\mathbf{f}_i \right\|^2, \quad \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r.$$

The (possibly not unique) solution is $\widehat{\mathbf{A}}$ being the first r (column) eigenvectors of $\mathbf{S}$ corresponding to the first r largest eigenvalues, and $\widehat{\mathbf{f}}_i = \mathbf{A}^{\mathrm{T}}\mathbf{y}_t$. The space spanned by the columns in $\widehat{\mathbf{A}}$ is unique though, called the (estimated) factor loading space. This is asymptotic PCA in the sense that they need $p \to \infty$ in order for the estimated factor loading space to converge to the true factor loading space, since the effect of $\mathbf{\Sigma}_\epsilon$ can then be diluted as $p \to \infty$ (treated relatively as going to the zero matrix, in a sense). In view of this, Lam et al. (2011) proposed a "statistical" factor model, where $\mathbf{A}\mathbf{f}_i$ is viewed as signal and $\boldsymbol{\epsilon}_i$ is viewed as noise, with the key assumption that $\{\boldsymbol{\epsilon}_i\}$ is a vector white noise. With this, observe that for $k > 0$, if we assume $\mathbf{\Sigma}_{\epsilon x}(k) := \mathrm{cov}(\boldsymbol{\epsilon}_t, \mathbf{x}_{t-k}) = \mathbf{0}$ (e.g., $\{\boldsymbol{\epsilon}_t\}$ is an innovation series), then

$$\mathbf{\Sigma}_x(k) = \mathrm{cov}(\mathbf{x}_t, \mathbf{x}_{t-k}) = \mathbf{A}(\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}} + \mathbf{\Sigma}_{x\epsilon}(k)),$$

so that the product $\mathbf{M}_K = \sum_{k=1}^{K} \mathbf{\Sigma}_x(k)\mathbf{\Sigma}_x(k)^{\mathrm{T}}$ is sandwiched between $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$. Hence an eigenanalysis of $\mathbf{M}_K$ would result in an estimator of $\mathbf{A}$ without the need for $p \to \infty$, and we can use the sample autocovariance matrices for estimating $\mathbf{M}_K$. Theoretical results are given in both papers, but they are not focused on covariance matrix estimation.

Fan et al. (2013) proposed to estimate $\mathbf{A}$ using asymptotic PCA as described before, and given r, estimate the low rank part by

$$\widehat{\mathbf{\Sigma}}_R = \sum_{j=1}^{r} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j^{\mathrm{T}}, \quad \text{with } \widehat{\mathbf{A}} = (\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_r), \tag{2.17}$$

and $\widehat{\lambda}_j$ the jth largest eigenvalue of $\mathbf{S}$, which is proved to be a good estimator of λ_j for $\mathbf{\Sigma}$ for a factor model, $j = 1, \ldots, r$. Then $\mathbf{\Sigma}_\epsilon$ is estimated by thresholding $\mathbf{S} - \widehat{\mathbf{\Sigma}}_R$. The method is abbreviated as POET. They proved nice asymptotic properties of the POET estimator and argue that sparseness for $\mathbf{\Sigma}_\epsilon$ is more relaxed than a strict factor model, and is more likely to be satisfied in applications like finance.

To accommodate the possibility of heavy-tailed data, Fan et al. (2018) assumed the data can be elliptically distributed:

$$\mathbf{x} = \boldsymbol{\mu} + \zeta\mathbf{B}\mathbf{U},$$

where $\mathbf{U}$ is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^q$, ζ is a non-negative scalar random variable independent of $\mathbf{U}$, and $\mathbf{B} \in \mathbb{R}^{p\times q}$ is deterministic such that $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$. They showed that as long as $\mathbf{S}$ can be replaced by $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{A}}$ by $\widehat{\mathbf{\Gamma}}$ and $\widehat{\lambda}_j$ by $\widehat{\eta}_j$ such that

$$\big\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|_{\max} = O_P(\sqrt{\log p/n}) = \max_{j=1,\dots,r} |\widehat{\eta}_j - \lambda_j|\lambda_j^{-1}, \quad \big\|\widehat{\mathbf{\Gamma}} - \mathbf{A}\big\|_{\max} = O_P(\sqrt{\log p/(np)}),$$

then the POET estimator can be obtained as in (2.17) and the description thereafter, with guaranteed rates of convergence. To obtain $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{D}}\widehat{\mathbf{R}}\widehat{\mathbf{D}}$, where $\widehat{\mathbf{D}}$ is the diagonal matrix with the estimated variances, they proposed to estimate each variances in $\widehat{\mathbf{D}}$ similar to the estimator in (2.15) using the Huber loss (2.16) for an appropriate α (can use the same trick to estimate the mean $\widehat{\boldsymbol{\mu}}$ first if $\boldsymbol{\mu}$ is not $\mathbf{0}$). Then they estimate the correlation matrix $\widehat{\mathbf{R}} = (\widehat{r}_{ij})$ by using the Kendall rank correlation $\widehat{r}_{ij}^{(K)}$ (a.k.a. Kendall's tau) and the formula $\widehat{r}_{ij} = \sin(\pi\widehat{r}_{ij}^{(K)}/2)$. The estimator $\widehat{\eta}_j$ can then be taken as the leading eigenvalues for $\widehat{\mathbf{\Sigma}}$. For $\widehat{\mathbf{\Gamma}}$, it would be taken as the eigenvectors from the U-statistic (spatial Kendall's tau)

$$\frac{2}{n(n-1)} \sum_{i<j} \frac{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}}{\big\|\mathbf{x}_i - \mathbf{x}_j\big\|^2}.$$

The key to why the above statistic works is that the summand can be proved to be independent of the distribution of ζ.

3 Shrinkage Covariance Matrix Estimation

While structured covariance matrix estimation is very useful in many applications, structural assumptions on $\mathbf{\Sigma}$ can require prior information on the data or $\mathbf{\Sigma}$ itself that is not available at times. A class of covariance matrix estimator built on the idea of shrinkage, perhaps first published in Stein (1975), has the eigenvalues of the sample covariance matrix shrunk explicitly according to a certain formula, while the eigenvectors are unchanged. Since then there are a number of attempts in shrinkage estimation. For instance, Daniels and Kass (2001) proposed several Bayesian estimators which shrink the eigenvalues of the sample covariance matrix, possibly towards a structure. Unfortunately some estimators need MCMC computations which can be computationally expensive, and some are not guaranteed to be positive semi-definite. Moreover, asymptotic results are only for fixed p.

3.1 Linear shrinkage

A major breakthrough comes in the method of linear shrinkage in Ledoit and Wolf (2004). They introduced an estimator

$$\widehat{\mathbf{\Sigma}}_{\mathrm{LS}} = \widehat{\phi}_1 \mathbf{I}_p + \widehat{\phi}_2 \mathbf{S} = \frac{\beta^2}{\alpha^2+\beta^2}\mu \mathbf{I}_p + \frac{\alpha^2}{\alpha^2+\beta^2}\mathbf{S}, \tag{3.1}$$

which is the solution to minimizing the expected quadratic loss $E\big\|\phi_1\mathbf{I}_p + \phi_2\mathbf{S} - \mathbf{\Sigma}\big\|_F^2$ with respect to ϕ_1, ϕ_2, with $\mu = \mathrm{tr}(\mathbf{S})/p$, $\alpha^2 = \big\|\mathbf{\Sigma} - \mu\mathbf{I}_p\big\|_F^2$ and $\beta^2 = E\big\|\mathbf{S} - \mathbf{\Sigma}\big\|_F^2$. Simple estimators for μ, α and β are also proposed and analyzed, with asymptotic framework $p/n \to c > 0$ for some finite constant c. The estimator cannot be consistent under this framework without further structural assumptions, but optimality under expected Frobenius loss is proved, while the estimator is always positive definite. Observe that

$$\widehat{\mathbf{\Sigma}}_{\mathrm{LS}} = \mathbf{P}\Big(\frac{\beta^2}{\alpha^2+\beta^2}\mu\mathbf{I}_p + \frac{\alpha^2}{\alpha^2+\beta^2}\mathbf{D}\Big)\mathbf{P}^{\mathrm{T}},$$

where $\mathbf{S} = \mathbf{PDP}^{\mathrm{T}}$, with $\mathbf{P}$ a matrix of eigenvectors and $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ the corresponding diagonal matrix of eigenvalues of $\mathbf{S}$. Hence $\widehat{\mathbf{\Sigma}}_{\mathrm{LS}}$ retains the eigenvectors but shrinks the eigenvalues of $\mathbf{S}$ towards a constant multiple of $\mathbf{I}_p$. This falls into the so-called rotation-equivariant class of estimators used in Stein (1975).

The simplicity of $\widehat{\mathbf{\Sigma}}_{\mathrm{LS}}$ attracted a lot of similar studies. Schäfer and Strimmer (2005) proposed a number of different target population covariance matrices instead of just a constant multiple of $\mathbf{I}_p$, and proposed corresponding optimality analyses in minimizing the expected quadratic loss. Warton (2008) proposed linear shrinkage to shrink the sample correlation matrix towards the identity, and proposed to use a K-fold cross-validation to estimate the shrinkage parameter.

Recently, Huang and Fryzlewicz (2018) introduced the NOVELIST ("NOVEL Integration of the Sample and Thresholded covariance estimators"), which combines linear shrinkage and sparse estimators, in the form

$$\widehat{\mathbf{\Sigma}}_{\mathrm{nv}} = (1-\delta)\mathbf{S} + \delta\mathbf{T}_\lambda(\mathbf{S}),$$

where $\mathbf{T}_\lambda(\mathbf{S})$ is a thresholded estimator of $\mathbf{S}$ with parameter λ introduced in (2.9), and δ controls if $\widehat{\mathbf{\Sigma}}_{\mathrm{novelist}}$ is closer to $\mathbf{S}$ or the sparse estimator $\mathbf{T}_\lambda(\mathbf{S})$. Compared to (3.1), NOVELIST is a linear shrinkage estimator with target matrix a sparsely estimated covariance matrix, which can also be replaced by other structured covariance matrix estimators like the POET introduced in Section 2.4.

With the class of sparse covariance matrix $\mathcal{C}_q$ introduced in (2.8), and assuming $\int_0^\infty \exp(\gamma t) dG_j(t) < \infty$ for γ on a bounded interval around 0, where G_j is the cumulative distribution function of the jth variable of a generic data vector $\mathbf{x}$, they proved that

$$\|\widehat{\mathbf{\Sigma}}_{\mathrm{nv}} - \mathbf{\Sigma}\| = O_P\left((1-\delta)p\sqrt{\frac{\log p}{n}} + \delta c_0(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right) = \|\widehat{\mathbf{\Sigma}}_{\mathrm{nv}}^{-1} - \mathbf{\Sigma}^{-1}\|,$$

where $\lambda = M'\sqrt{\log p/n}$ for sufficiently large M' with $\log p/n = o(1)$. If $p = o(n)$ and $\mathbf{u}^{\mathrm{T}}\mathbf{x}$ has Gaussian tails for all unit vector $\mathbf{u}$, then the above result still holds with $p\sqrt{\log p}$ replaced by $\sqrt{p + \log n}$ on the left term of the rate. The above rate however is assuming that δ is known. See Huang and Fryzlewicz (2018) for more details on how to choose λ and δ.

3.2 Nonlinear shrinkage and others

The shrinkage estimator proposed in Stein (1975) is shrinking the sample eigenvalues nonlinearly, but without a proper loss function associated. Won et al. (2013) proposed to maximize the normal log-likelihood of the data but impose a condition number constraint on the estimator, resulting in winsorized eigenvalues while retaining the sample eigenvectors $\mathbf{P}$. The estimator is also proved to have lower entropy loss than $\mathbf{S}$, but is not proved to be optimally nonlinearly shrunk with respect to such loss.

Nonlinear shrinkage comes with the important parallel development of random matrix theory, which fast-tracked the study of many other powerful statistical procedures and their corresponding theoretical analysis. We refer interested readers to two review papers Paul and Aue (2014) and Johnstone and Paul (2018) for more technical details related to random matrix theory, which will not be covered in the description of nonlinear shrinkage below.

With respect to minimizing the Frobenius loss $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2$, Ledoit and Péché (2011) showed that with a class of rotation-equivariant estimators $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_p)$ is the matrix of eigenvectors for $\mathbf{S}$ and $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$ is a diagonal matrix to be determined, the solution is $d_i = \mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}\mathbf{p}_i$ for $i = 1, \ldots, p$. It means that if $\mathbf{P}$ is used as eigenvectors, it is not necessary that the true eigenvalues are giving optimality, since d_i is not converging to the corresponding true eigenvalue when $p/n \to c > 0$. This also represents an ideal shrinkage formula when Frobenius error is concerned. Under $p/n \to c > 0$ (excluding $c = 1$ for technical reasons) and using random matrix theory (mainly the Stieltjes transformation as a technical tool), Ledoit and Péché (2011) developed explicit formulae for estimating d_i which involves the so-called generalized Marčhenko-Pastur equation. Ledoit and Wolf (2012) proposed how to use data to estimate such a nonlinear transformation, thereby resulting in the nonlinear shrinkage estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$. They have proved asymptotic efficiency and convergence of $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$ to the "ideal" estimator of the form

$$\mathbf{\Sigma}_{\mathrm{Ideal}} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{P})\mathbf{P}^{\mathrm{T}}. \tag{3.2}$$

This ideal estimator is the theoretical optimal estimator that minimizes the Frobenius loss. Hence $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$, being convergent to $\mathbf{\Sigma}_{\mathrm{Ideal}}$, is asymptotically optimal under the framework $p/n \to c > 0$ with respect to the class of rotation-equivariant estimators. Ledoit and Wolf (2017) applied $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$ to portfolio allocation and also developed further a portfolio allocation strategy using a mean target return. Engle et al. (2017) implement $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$ into its dynamic conditional correlation framework, such that the updating equation of the dynamic correlation matrix is based on the large estimated correlation matrix derived from $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$.

Abadir et al. (2014) proposed to split the data into two parts, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ has size $p \times n_1$ and $\mathbf{X}_2$ is $p \times n_2$ with $n = n_1 + n_2$. Defining $\widetilde{\mathbf{\Sigma}}_i = n_i^{-1}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$, $i = 1, 2$, and $m = n_1$ to be the split location, they propose

$$\check{\mathbf{\Sigma}}_m = \mathbf{P}\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}^{\mathrm{T}}, \tag{3.3}$$

where $\mathbf{P}_i$ is the matrix of eigenvectors for $\widetilde{\mathbf{\Sigma}}_i$. With independent observations in $\mathbf{X}$, we can permute the data and form the above estimator again from the split data $\mathbf{X} = (\mathbf{X}_1^{(j)}, \mathbf{X}_2^{(j)})$, $j = 1, \ldots, M$, ultimately leading to the grand average estimator

$$\check{\mathbf{\Sigma}} = \mathbf{P}\bigg(M^{-1}\sum_{j=1}^{M}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\bigg)\mathbf{P}^{\mathrm{T}}, \tag{3.4}$$

where $\widetilde{\mathbf{\Sigma}}_i^{(j)} = n_i^{-1}\mathbf{X}_i^{(j)}\mathbf{X}_i^{(j)\mathrm{T}} = \mathbf{P}_{ij}\mathbf{D}_{ij}\mathbf{P}_{ij}^{\mathrm{T}}$, $i = 1, 2$. They show that when $p < n - m$ and $p/n \to 0$, a split such that $m/n \to \gamma \in (0, 1)$ will make $\check{\mathbf{\Sigma}}_m$ optimal with respect to the expected element-wise L_1 or L_2 loss.

Using the same data splitting idea, Lam (2016) proposed the NERCOME,

$$\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}. \tag{3.5}$$

This estimator is designed to minimize $\big\|\mathbf{P}_1\mathbf{D}\mathbf{P}_1 - \mathbf{\Sigma}\big\|_F^2$, and is proved to be convergent to $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal,1}} = \mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{\Sigma}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$, the ideal estimator with $\mathbf{P}_1$ replacing $\mathbf{P}$, under the spectral norm with $p/n \to c > 0$ when $\sum_{n\geq 1} p(n-m)^{-5} < \infty$, including the case $c = 1$. The estimator $\widehat{\mathbf{\Sigma}}_m$ is also asymptotically as efficient as the ideal estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal}}$ (the one using $\mathbf{P}$) in estimating $\mathbf{\Sigma}$ with respect to the Frobenius loss when we also have $m/n \to 1$ and $n - m \to \infty$. Importance about $\widehat{\mathbf{\Sigma}}_m$ is that the convergence to $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal,1}}$ means that $\widehat{\mathbf{\Sigma}}_m$ is also a nonlinear shrinkage estimator like $\widehat{\mathbf{\Sigma}}_{NS}$, with only $\mathbf{P}$ replaced by $\mathbf{P}_1$. Its calculation involves data splitting and eigenanalysis, which can be faster than the algorithm in Ledoit and Wolf (2012) when p is small to moderate in size (e.g. p of order of hundreds). Practically, $c = 1$ poses no problems for $\widehat{\mathbf{\Sigma}}_m$, while $\widehat{\mathbf{\Sigma}}_{\mathrm{NS}}$ can have problems from the QuEST package proposed in Ledoit and Wolf (2012). At the same time, $m/n \to 1$ is needed but not $m/n \to \gamma \in (0, 1)$, since in the analysis in Lam (2016), p is growing as fast as n, while Abadir et al. (2014) considered p to be growing slower than n.

While using $\mathbf{P}_1$ as the matrix of eigenvectors do not fully utilize all data like $\mathbf{P}$, the averaged estimator

$$\widehat{\mathbf{\Sigma}} = M^{-1} \sum_{j=1}^{M} \mathbf{P}_{1j} \text{diag}(\mathbf{P}_{ij}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2^{(j)} \mathbf{P}_{1j}) \mathbf{P}_{1j}^{\mathrm{T}} \tag{3.6}$$

can be performing better than $\check{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}_{\text{NS}}$ as demonstrated numerically in Lam (2016). This estimator is also proved to be asymptotically as efficient as $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ in estimating $\mathbf{\Sigma}$ with respect to the Frobenius loss when $p/n \to c > 0$, while its inverse is asymptotically as efficient as $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}^{-1}$ in estimating $\mathbf{\Sigma}^{-1}$ with respect to the inverse Stein's loss under $p/n \to c > 0$. Lam (2016) also proved these asymptotic properties when the data follows a factor model, so that no estimation of the number of factors is necessary while such properties are kept.

Beyond Frobenius loss, nonlinear shrinkage can have very different formulas compared to those proposed in Ledoit and Wolf (2012) even for the rotation-equivariant class, since the solution to the optimization problem $\min_{\mathbf{D}} L(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}, \mathbf{\Sigma})$, where $L(\cdot,\cdot)$ is a general loss function, can be very different from $d_i = \mathbf{p}_i^{\mathrm{T}} \mathbf{\Sigma} \mathbf{p}_i$. With normality of data assumed, under $p/n \to c \in (0, 1]$ and a spiked covariance model where $\mathbf{\Sigma}$ has r fixed top-eigenvalues followed by all ones, Donoho et al. (2018) derived optimal shrinker for a wide variety of loss functions, showing that optimality is very loss-function, and hence application, dependent.

4 Applications

Depending on the application, an estimated covariance or precision matrix can be used for many different purposes. From a stepping stone for further data analysis to being the highlight in its own right, we give several applications of covariance matrix estimation in this section, comparing a number of different procedures introduced in the process.

4.1 Principal component analysis (PCA)

This is a perfect example of a very common statistical procedure where the population covariance matrix $\mathbf{\Sigma}$ is of central importance to the problem, but optimization should be carried out with other quantities in mind, namely the first r largest eigenvalues and their corresponding eigenvectors for $\mathbf{\Sigma}$, where r is usually the number of "factors" that explain most of the variance of the data. Depending on the application, there may not be distinguishable "factors" though, in the sense that all eigenvalues of $\mathbf{\Sigma}$ are of the same (constant) order.

In the PCA literature, it is often of interest to study a spike model for $\mathbf{\Sigma}$. As in Shen et al. (2016) for instance, a multiple-component spike model is defined as

$$\lambda_j = \begin{cases} c_j p^{\alpha}, & j \leq m; \\ 1, & j > m, \end{cases} \quad \alpha \geq 0,$$

where m is a finite integer and the constants c_j are positive with $c_j > c_{j+1} > 1$ for $j = 1, \ldots, m-1$. This is equivalent to defining

$$\mathbf{\Sigma} = \mathbf{Q}\text{diag}(c_1 p^{\alpha} - 1, \ldots, c_m p^{\alpha} - 1)\mathbf{Q}^{\mathrm{T}} + \mathbf{I}_p, \tag{4.1}$$

where $\mathbf{Q} \in \mathbb{R}^{p \times m}$ is such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_m$. A more general model for $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \mathbf{Q}\text{diag}(c_1 p^{\alpha}, \ldots, c_m p^{\alpha})\mathbf{Q}^{\mathrm{T}} + \mathbf{\Sigma}_e, \tag{4.2}$$

where $\mathbf{\Sigma}_e$ has uniformly bounded eigenvalues. Both (4.1) and (4.2) are associated with the factor model for the data,

$$\mathbf{x}_i = \mathbf{Q}\mathbf{f}_i + \mathbf{e}_i, \ \ i = 1, \ldots, n, \tag{4.3}$$

where $\mathbf{Q}$ is as defined in (4.1) or (4.2), $\mathbf{f}_i$ is independent of $\mathbf{e}_i$, with $\text{var}(\mathbf{f}_i) = \text{diag}(c_1 p^{\alpha}, \ldots, c_m p^{\alpha})$ and $\text{var}(\mathbf{e}_i) = \mathbf{\Sigma}_e$. Compare this model to (2.13) in Section 2.4. In financial econometrics literature, (4.1) is called a strict factor model while (4.2) is called an approximate factor model. The index α can be considered the signal strength of the model. If α is large, then consistent estimation of the eigenvectors of $\mathbf{\Sigma}$ (i.e., the principal component directions) is easier to achieve through an eigenanalysis of $\mathbf{S}$, the sample covariance matrix of the data.

Consider a very simple one-factor model ($m = 1$),

$$\mathbf{x}_i = \mathbf{A}u_i + \mathbf{e}_i, \ \ i = 1, \ldots, n,$$

where $\mathbf{A} = (a_i)$ is a column vector of constants such that $0 < c_{\min} \leq |a_i| \leq c_{\max} < \infty$ for $c_{\min}$ and $c_{\max}$ some universal constants, and $\text{var}(u_i) < \infty$ uniformly as $n, p \to \infty$. Then we can rewrite

$$\mathbf{x}_i = \frac{\mathbf{A}}{\|\mathbf{A}\|} \cdot \|\mathbf{A}\| u_i + \mathbf{e}_i,$$

so that we can take $\mathbf{Q} = \mathbf{A}/\|\mathbf{A}\|$ and $f_i = \|\mathbf{A}\| u_i$ for model (4.3), since by assumption of $\mathbf{A}$, $\|\mathbf{A}\|$ has order $p^{1/2}$, so that $\text{var}(f_i) = \|\mathbf{A}\|^2 \text{var}(u_i)$ has order p. This means that $\alpha = 1$. The factor u_i is then called a pervasive factor in the jargon of financial econometrics literature. It means that the dynamics of u_i is affecting almost all of the variables in $\mathbf{x}_i$.

For a general r-factor model with r pervasive factors (i.e., $\alpha = 1$ for the first r eigenvalues of $\boldsymbol{\Sigma}$), Fan et al. (2013) showed that with sparse $\boldsymbol{\Sigma}_e$, their POET method can consistently estimate $\mathbf{Q}$ when $n = o(p^2)$, even when p is growing exponentially fast relative to n. In fact, sparsity of $\boldsymbol{\Sigma}_e$ is not needed for just consistent estimation of $\mathbf{Q}$ as long as $\boldsymbol{\Sigma}_e$ has all eigenvalues uniformly bounded above. Hence in terms of PCA analysis, POET can achieve consistent estimation of the principal component directions with $\alpha = 1$. This actually means that an eigen-analysis of the sample covariance matrix $\mathbf{S}$ is already enough for consistent estimation of the first r principal component directions in such a pervasive r-factor model, since POET used these directions for the construction of their estimator. It also means that all rotation-equivariant shrinkage estimators mentioned in Section 3.2 are also fine for extracting the first r principal component directions since they utilize all eigenvectors of $\mathbf{S}$ in their construction.

However, Shen et al. (2016) showed for a very wide range of (p, n, α) that, in general, when p is large, α is small (say $\alpha = 0$) or n is small, it is more difficult for the sample eigenvectors to be consistent. See their paper for detailed theoretical results with rates of convergence, and the references therein. Since $\boldsymbol{\Sigma}$ is not the ultimate aim of PCA, but rather the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$, methods are proposed for structural estimation of the eigenvectors of $\boldsymbol{\Sigma}$. The estimation of eigenvalues is the study of the spectrum of $\boldsymbol{\Sigma}$, which falls in the spectrum estimation of high dimensional covariance matrix. The techniques heavily involve random matrix theory again. Interested readers are referred to Ledoit and Wolf (2015) and the references therein.

For structural estimation of the eigenvectors, a very popular choice is to assume that the eigenvectors themselves are sparse, in the sense that many elements in the eigenvectors are very small, except for a few. Sparse PCA (SPCA) is to extract principal component directions assuming sparsity of the eigenvectors of $\boldsymbol{\Sigma}$. This enhances interpretability of principal component directions, in the sense that it indicates only a few variables among p of them are important in a particular principal component direction. Instead of reviewing different SPCA methods, see a nice review paper such as Zou and Xue (2018) for more details on how SPCA can help obtain consistency in the eigenvectors estimation again under high dimension. For a general PCA review, see Abdi and Williams (2010) and Johnstone and Paul (2018) and the references therein.

4.2 Covariance estimation for cosmological data

The data comes from a mock cosmological survey analyzed in Joachimi (2016). To avoid as much technical terms in cosmological study as possible, the data consists of $p = 120$ and $n_r = 2000$ independent and identically distributed data vectors, recording “two-point correlation functions of cosmic weak lensing”. They are observed deep into a part of the universe, and if the region of the universe we observe grow bigger, then so does p. We want to estimate the population covariance matrix $\mathbf{\Sigma}$ of the observations, to be further used in statistical analysis of some astronomical models. Ultimately, we want to get as little bias and variance in the parameters estimated in those models as possible, meaning a good covariance matrix estimator is essential. The precision matrix is particularly important, since likelihood functions of those models involve the precision matrix.

Each of the n_r realizations is actually independent simulation of the universe from big bang, and hence involve hugely expensive computational cost and needed mainframe supercomputers to finish. Hence the main aim of the study in Joachimi (2016) is to discover if a regularized estimator of $\mathbf{\Sigma}$, instead of the sample covariance matrix, can achieve good performance for the estimation of those astronomical parameters relative to the population covariance matrix, but with much reduced sample size, ideally much less than $n_r = 2000$. If this is achievable, it means that we can maintain the standard of the estimated parameters, but with much less realizations (and hence much less computational cost) for estimating $\mathbf{\Sigma}$.

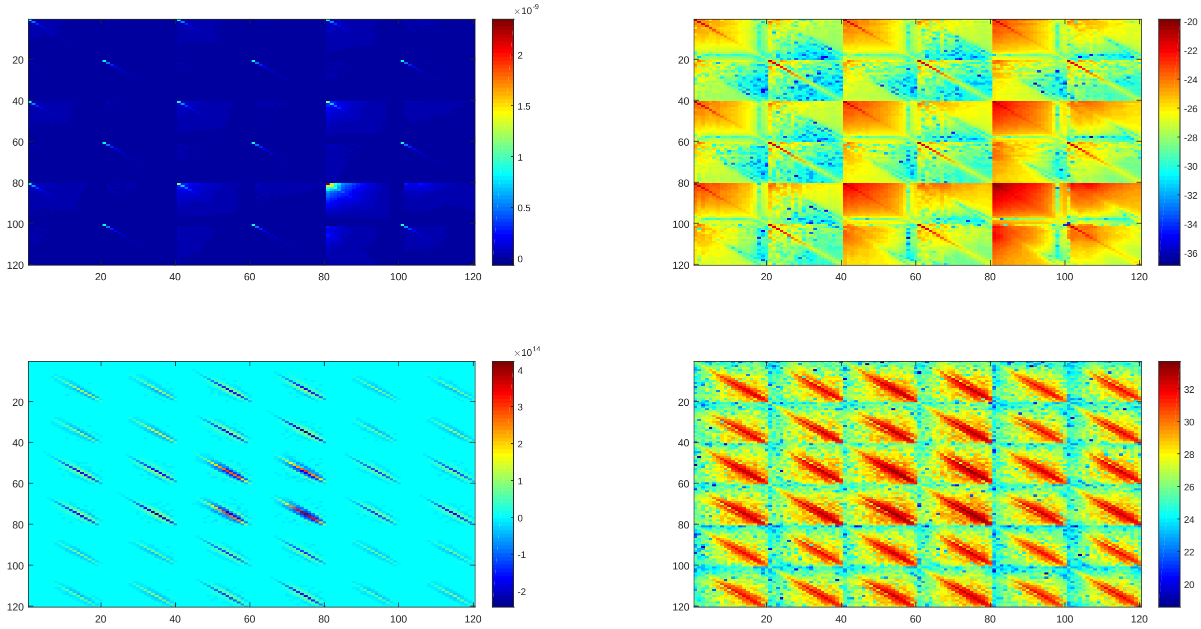

Figure 1: Upper left: $\mathbf{\Sigma}$, $p = 120$. Upper right: $|\mathbf{\Sigma}|$, in log-scale. Lower left: $\mathbf{\Sigma}^{-1}$. Lower right: $|\mathbf{\Sigma}^{-1}|$, in log-scale.

Figure 1 shows $\mathbf{\Sigma}$, estimated by the sample covariance using all $n_r = 2000$ realizations. The log-scale plot reveal many fine structures of the covariance and the precision matrices, although the original scale plot displays many close-to-zero elements. Certainly we do not know if $\mathbf{\Sigma}$ is sparse or not if we do not have all $n_r = 2000$ realizations, but it would not be too difficult to see that there could be sparse elements in both $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ even from looking at the heat map of a sample covariance matrix of a much smaller subsample, say $n = 80$. This prompts us to use sparse estimation of covariance and precision matrix. Since precision matrix plays a more important role, we would want to put more focus on the performance of an estimator for $\mathbf{\Sigma}^{-1}$.

Joachimi (2016) has used $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ as a signal-to-noise ratio, where $\boldsymbol{\mu}$ is the true mean of the realizations. We perform a simulation experiment to compare the estimated signal-to-noise ratio to the true one, using the bias $\widehat{\boldsymbol{\mu}}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ to gauge performance. On top of this, we use the Frobenius error $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$ to illustrate the last point made in Section 3.2, namely, different criteria can give different optimal estimators. We randomly draw $n = 80$ realizations from the pool of 2000 for each simulation. We compare 5 estimators: Graphical lasso (GLASSO) from Section 2.3, the NOVELIST estimator by Huang and Fryzlewicz (2018), NERCOME by Lam (2016), nonlinear shrinkage (Nonlin) estimator by Ledoit and Wolf (2012), and finally the grand average (Grand Avg.) estimator by Abadir et al. (2014). The last 3 estimators are introduced in Section 3.2, while the NOVELIST is introduced in Section 3.1. The tuning parameter for the graphical lasso is pre-set so that it can estimate the signal-to-noise ratio best. Since $\mathbf{\Sigma}^{-1}$ is sparse as seen in Figure 1, we expect the graphical lasso to perform well since it encourages sparsity in the precision matrix.

We run the simulations 200 times. Table 1 shows, as expected, that the graphical lasso is the best for estimating the signal-to-noise ratio which involve $\mathbf{\Sigma}^{-1}$. Changing to the Frobenius loss, the best estimator becomes NERCOME. All in all, it is of much importance to determine what kind of criterion to use, and what knowledge/structure we can assume on $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-1}$, before we determine what estimators to use.

	NERCOME	Nonlin	Grand Avg.	NOVELIST	GLASSO	**S**
$\widehat{\boldsymbol{\mu}}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$	$46.8_{(35.8)}$	$55.5_{(34.9)}$	$10.3_{(21.3)}$	$39.1_{(101.9)}$	$-3.8_{(9.5)}$	-
$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$	$2.7_{(1.1)}$	$3.4_{(4.0)}$	$3.0_{(2.0)}$	$3.1_{(3.5)}$	$16.6_{(1.4)}$	$3.3_{(4.0)}$

Table 1: Bias and Frobenius error for 5 estimators, with mean and standard deviation (in bracket) reported. The last column is the sample covariance matrix, which is always singular in this experiment.

4.3 Risk management and portfolio allocation

In finance, portfolio management is important for risk and return control. Assuming $\mathbf{x}_i$ is an observed daily/weekly/monthly log-return vector of p assets, with $\mathbf{m} = E(\mathbf{x}_i)$, $\mathrm{var}(\mathbf{x}_i) = \mathbf{\Sigma}$ and they are assumed to be stationary for a period of time. The classical Markowitz portfolio allocation theory solves the problem

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w} \;\text{ subject to }\; \mathbf{m}^{\mathrm{T}}\mathbf{w} \geq \mu, \; \mathbf{w}^{\mathrm{T}}\mathbf{1}_p = 1, \tag{4.4}$$

where $\mathbf{1}_p$ is a vector of p ones, and μ is a "target return" which the portfolio hope to achieve on average. In words, we want to minimize the "risk" of the portfolio $\mathbf{w}$, which is defined as the variance of the associated return, $\mathrm{var}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$, subject to the mean return $\mathbf{m}^{\mathrm{T}}\mathbf{w}$ being larger than the target return. Without the target return constraint, the *minimum variance portfolio* is the solution

$$\mathbf{w}_{\mathrm{mv}} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_p}. \tag{4.5}$$

If $\mathbf{m}^{\mathrm{T}}\mathbf{w}_{\mathrm{mv}} \geq \mu$, then $\mathbf{w}_{\mathrm{mv}}$ is also the solution to (4.4). Otherwise, if $\mathbf{m}$ is linearly independent of $\mathbf{1}_p$, then the solution to (4.4) is

$$\mathbf{w}_{\mathrm{opt}} = (1-\alpha)\mathbf{w}_{\mathrm{mv}} + \alpha\mathbf{w}_{\mathrm{mkt}}, \quad \mathbf{w}_{\mathrm{mkt}} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{m}}{\mathbf{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m}}, \quad \alpha = \frac{\mu(\mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m})(\mathbf{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}) - (\mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e})^2}{(\mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m})(\mathbf{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}) - (\mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e})^2}. \tag{4.6}$$

We can see that the explicit solution always involve the precision matrix $\boldsymbol{\Sigma}^{-1}$, and hence it is important to have a good estimator of $\boldsymbol{\Sigma}^{-1}$. At the same time, the estimated risk of any given portfolio $\mathbf{w}$ is $\mathbf{w}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{w}$, so that we want to have a good estimator for $\boldsymbol{\Sigma}$ when it comes to risk assessment. This is on top of the need for estimating the high dimensional vector $\mathbf{m}$, which is often not being stationary after a certain (supposedly short) time period.

Focusing on estimating $\mathbf{w}_{\mathrm{mv}}$, an ideal criterion will be to minimize $\|\mathbf{w}_{\mathrm{mv}} - \widehat{\mathbf{w}}_{\mathrm{mv}}\|$, where $\widehat{\mathbf{w}}_{\mathrm{mv}}$ has $\boldsymbol{\Sigma}$ replaced by an estimator $\widehat{\boldsymbol{\Sigma}}$. As far as we know there are no estimators that aim to minimize this. Fan et al. (2013) instead assumed a factor model (2.13) and proposed POET for $\widehat{\boldsymbol{\Sigma}}$ as described in Section 2.4. This makes sense as return data usually has at least a market factor that is pervasive.

In an empirical study, Lam (2016) considered risk minimization for a portfolio of $p = 100$ stocks, which is also considered in section 7.2 of Fan et al. (2013). The data consists of 2640 annualized daily excess returns $\{r_t\}$ for the period January 1st 2001 to December 31st 2010 (22 trading days each month). Five portfolios are created at the beginning of each month using five different methods in estimating the covariance matrix of returns. A typical setting here is $n = 264, p = 100$, that is, one year of past returns to estimate a covariance matrix of 100 stocks. Each portfolio has weights given by

$$\widehat{\mathbf{w}} = \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_p},$$

where $\widehat{\mathbf{\Sigma}}^{-1}$ is an estimator of the $p \times p$ precision matrix for the stock returns, using strict factor model (i.e., (4.3) with $\text{var}(\mathbf{e}_i) = \sigma^2 \mathbf{I}_p$, abbreviated as SFM), POET from Section 2.4, and grand average, NERCOME and nonlinear shrinkage (Nonlin) from Section 3.2 respectively. At the end of each month, for each portfolio, we compute the total excess return, the out-of-sample variance and the mean Sharpe ratio, given respectively by (see also Demiguel and Nogales (2009)):

$$\widehat{\mu} = \sum_{i=12}^{119} \sum_{t=22i+1}^{22i+22} \mathbf{w}^{\mathrm{T}} \mathbf{r}_t, \;\; \widehat{\sigma}^2 = \frac{1}{2376} \sum_{i=12}^{119} \sum_{t=22i+1}^{22i+22} (\mathbf{w}^{\mathrm{T}} \mathbf{r}_t - \widehat{\mu}_i)^2, \;\; \widehat{\text{sr}} = \frac{1}{108} \sum_{i=12}^{119} \frac{\widehat{\mu}_i}{\widehat{\sigma}_i^2}.$$

	SFM	POET	NERCOME	Grand Avg.	Nonlin
Total excess return	153.9	109.5	128.0	127.9	124.8
Out-of-sample variance	.312	.267	.264	.264	.264
Mean Sharpe Ratio	.224	.197	.212	.211	.205

Table 2: *Performance of different methods. SFM represents the strict factor model, with diagonal covariance matrix.*

Table 2 shows the results. Clearly, the out-of-sample variance, which is a measure of risk, is the smallest for NERCOME, grand average and Nonlin, while the strict factor model has the highest risk. One highlight of NERCOME, which is proved in Lam (2016) but not in Abadir et al. (2014) or Ledoit and Wolf (2012) for the grand average and Nonlin respectively, is that we do not even need to estimate the number of factors for the underlying factor model for the data, which is a crucial input in the case for strict factor model and POET.

4.3.1 Intraday data

Previously we consider $\mathbf{x}_i$ to be *low-frequency* return data recorded at most daily. When *intraday data* is concerned, we consider the log-price processes of p assets using a diffusion model

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \mathbf{\Theta}_t d\mathbf{W}_t, \quad t \in [0, 1],$$

where $\mathbf{W}_t$ is a p-dimensional standard Brownian motion. With this model, the integrated covariance matrix $\mathbf{\Sigma} = \int_0^1 \mathbf{\Theta}_t \mathbf{\Theta}_t^{\mathrm{T}} dt$ then plays the central role for portfolio allocation. For a portfolio $\mathbf{w}$, $\mathbf{w}^T \mathbf{\Sigma} \mathbf{w} = \int_0^1 \mathbf{w}^{\mathrm{T}} \mathbf{\Theta}_t \mathbf{\Theta}_t^{\mathrm{T}} \mathbf{w} dt$ can be considered as an accumulation of instantaneous risk $\mathbf{w}^{\mathrm{T}} \mathbf{\Theta}_t \mathbf{\Theta}_t^{\mathrm{T}} \mathbf{w}$ at time t over the whole period $[0, 1]$. The literature is rich (yet still young) in how to estimate this matrix under high dimension for intraday data, or even truly high-frequency data (prices within 5 minutes interval, or even tick-by-tick data), where the price is contaminated by the so-called microstructure noise. See for instance Wang and Zou (2010), Dao et al. (2017) and Fan and Kim (2017) and the references therein.

Since Lam (2016) has shown that nonlinear shrinkage can be attained using a data splitting scheme without explicit transformation formulas, Lam et al. (2017) proposed to estimate $\mathbf{\Sigma}$ assuming constant correlation matrix process assumptions, allowing us to write $\mathbf{\Theta}_t = \gamma_t \mathbf{\Lambda}$, so that $\mathbf{\Sigma} = \int_0^1 \gamma_t^2 dt \cdot \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$. This leads to a two-part estimation procedure, with $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ to be estimated using the self-normalized return $\mathbf{r}_\ell = p^{1/2}\Delta\mathbf{X}_\ell / \|\Delta\mathbf{X}_\ell\|$, where $\Delta\mathbf{X}_\ell = \mathbf{X}_{\tau_\ell} - \mathbf{X}_{\tau_{\ell-1}}$ and τ_ℓ represents the ℓth synchronous observation time, $\ell = 1, \ldots, n$. When p is fixed, the sample covariance matrix of $\mathbf{r}_\ell$, $\mathbf{S}_r = n^{-1}\sum_{\ell=1}^n \mathbf{r}_\ell \mathbf{r}_\ell^{\mathrm{T}}$ works well. Under the framework $p/n \to c > 0$, we are tempted to use nonlinear shrinkage on $\mathbf{S}_r$ for regularization, but $\mathbf{r}_\ell$ is not of the form $\mathbf{A}\mathbf{z}_\ell$ for some constant matrix $\mathbf{A}$ and random vector $\mathbf{z}_t$ with independent and identical standardized elements, meaning that the formulas in Ledoit and Wolf (2012) are not applicable. The NERCOME estimator in Lam (2016) can be applied to the data $\mathbf{r}_\ell$, and the corresponding asymptotic optimal properties are presented in Lam et al. (2017). Lam and Feng (2018) even applied nonlinear shrinkage on tick-by-tick data which is contaminated by market microstructure noise with non-synchronous trading times, and hence the returns are not even independent. They constructed a nonlinear shrinkage integrated volatility matrix estimator that is proved to be converging to an ideal estimator with a specific rate of convergence while the constructed matrix is positive definite in probability. They also proved that the minimum variance portfolio using such an estimator has some nice exposure bounds that are the same as the theoretical minimum variance portfolio. See simulations and portfolio exercises carried out in Lam and Feng (2018) which compared a number of state-of-the-art alternatives, including methods that directly regularizes on the portfolio weight in Fan et al. (2012) and DeMiguel et al. (2009).

5 Conclusion

Estimation of covariance matrix in high dimension is difficult because the sample covariance matrix simply fails miserably, and we have to impose regularization explicitly. Two branches of regularization are presented in this paper. One branch is to impose particular structure in our estimation procedure, and another branch to shrink the extreme eigenvalues of the sample covariance matrix. The appropriate method to use depends heavily on applications as well. If scientific knowledge indicates particular structures in the population covariance matrix, then we want to use regularization that enhances those structures. Otherwise, shrinkage is not a bad idea in the absence of *a priori* information on the data and the structure of the population covariance matrix. Shrinkage estimators are usually associated with particular loss functions. Different loss functions can result in different shrinkage formulas for the eigenvalues of the sample covariance matrix.

There are still many open challenges. Like Section 4.3 mentioned, an ideal criterion for optimization does not mean it is easy to derive the corresponding optimizer, when ideal criterion is also problem-dependent. And for time series data, how can we estimate conditional covariance matrix efficiently? Partial solutions are offered in finance, for example, in Engle et al. (2019) where they proposed a large dynamic covariance matrix estimator, but a useful estimator may be different in other scientific fields. We also mentioned robust estimation in Section 2.1 and accommodation of heavy-tailed data in Section 2.4, which are both very important topics in covariance matrix estimation. Finally, if a data vector can be naturally formed into an array for each observation (for example, a matrix), a covariance matrix for the data vector may then be generalized to a higher order tensor structure, depending on applications. How can we perform regularization effectively then?

References

Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165 – 180.

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 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. *Ann. Statist.*, 36(1):199–227.

Bien, J. (2019). Graph-guided banding of the covariance matrix. *Journal of the American Statistical Association*, 114(526):782–792.

Bien, J., Bunea, F., and Xiao, L. (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association*, 111(514):834–845.

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

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. T., Liu, W., and Zhou, H. H. (2016a). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016b). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.*, 10(1):1–59.

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

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

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under hubers contamination model. *Ann. Statist.*, 46(5):1932–1960.

Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184.

Dao, C., Lu, K., and Xiu, D. (2017). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. *Chicago Booth Research Paper No. 17-02.*

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Demiguel, V. and Nogales, F. J. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Donoho, D., Gavish, M., and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.*, 46(4):1742–1778.

Engle, R. F., Ledoit, O., and Wolf, M. (2017). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 0(0):1–13.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J. and Kim, D. (2017). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association*. Forthcoming.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high- frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

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.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.*, 46(4):1383–1414.

Fan, J., Wang, W., and Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5 – 22. Special Issue on Financial Engineering and Risk Management.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Guo, S., Box, J. L., and Zhang, W. (2017). A dynamic structure for high-dimensional covariance matrices and its application in portfolio allocation. *Journal of the American Statistical Association*, 112(517):235–253.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Huang, N. and Fryzlewicz, P. (2018). Novelist estimator of large correlation and covariance matrices and their inverses. *TEST*.

Joachimi, B. (2016). Non-linear shrinkage estimation of large-scale structure covariance. *Monthly Notices of the Royal Astronomical Society: Letters*, 466(1):L83–L87.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693. PMID: 20617121.

Johnstone, I. M. and Paul, D. (2018). Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292.

Kendall, M. (1948). *Rank correlation methods.* Griffin, London.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist.*, 44(3):928–953.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278.

Lam, C. and Feng, P. (2018). A nonparametric eigenvalue-regularized integrated covariance matrix estimator for asset return data. *Journal of Econometrics*, 206(1):226 – 257.

Lam, C., Feng, P., and Hu, C. (2017). Nonlinear shrinkage estimation of large integrated covariance matrices. *Biometrika*, 104(2):481–488.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264.

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

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360 – 384.

Ledoit, O. and Wolf, M. (2017). Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388.

Li, D., Xue, L., and Zou, H. (2018). Applications of peter hall's martingale limit theory to estimating and testing high dimensional covariance matrices. *Statistica Sinica*, 28:2657–2670.

Li, D. and Zou, H. (2016). Sure information criteria for large covariance matrix estimation and their asymptotic properties. *IEEE Transactions on Information Theory*, 62(4):2153–2169.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328.

Ma, S., Xue, L., and Zou, H. (2013). Alternating direction methods for latent variable gaussian graphical model selection. *Neural Comput.*, 25(8):2172–2198.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electron. J. Statist.*, 6:2125–2149.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Pan, J. and Mackenzie, G. (2003). On modelling meancovariance structures in longitudinal studies. *Biometrika*, 90(1):239–244.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1 – 29.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika*, 94(4):1006–1013.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation With High-Dimensional Data.* Wiley Series in Probability and Statistics. Wiley Interscience.

Qiu, Y. and Chen, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *Ann. Statist.*, 40(3):1285–1314.

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

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341 – 360.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515.

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.

Rothman, A. J., Levina, E., and Zhu, J. (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Shen, D., Shen, H., and Marron, J. S. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, 17:1–29.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34(1):1373–1403.

Wang, Y. and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *Ann. Statist.*, 38(2):943–978.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103(481):340–349.

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.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571.

Xue, L. and Zou, H. (2014). Rank-based tapering estimation of bandable correlation matrices. *Statistica Sinica*, 24:83–100.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106(8):1311–1320.