Electronic Journal of Statistics
Vol. 10 (2016) 1–59
ISSN: 1935-7524
DOI: 10.1214/15-EJS1081

Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation*

T. Tony Cai†

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: tcai@wharton.upenn.edu

Zhao Ren

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: zren@pitt.edu

and

Harrison H. Zhou‡

Department of Statistics, Yale University, New Haven, CT 06511, USA
e-mail: huibin.zhou@yale.edu

Abstract: This is an expository paper that reviews recent developments on optimal estimation of structured high-dimensional covariance and precision matrices. Minimax rates of convergence for estimating several classes of structured covariance and precision matrices, including bandable, Toeplitz, sparse, and sparse spiked covariance matrices as well as sparse precision matrices, are given under the spectral norm loss. Data-driven adaptive procedures for estimating various classes of matrices are presented. Some key technical tools including large deviation results and minimax lower bound arguments that are used in the theoretical analyses are discussed. In addition, estimation under other losses and a few related problems such as Gaussian graphical models, sparse principal component analysis, factor models, and hypothesis testing on the covariance structure are considered. Some open problems on estimating high-dimensional covariance and precision matrices and their functionals are also discussed.

MSC 2010 subject classifications: Primary 62H12; secondary62F12, 62G09.
Keywords and phrases: Adaptive estimation, banding, block thresholding, covariance matrix, factor model, Frobenius norm, Gaussian graphical model, hypothesis testing, minimax lower bound, operator norm, optimal rate of convergence, precision matrix, Schatten norm, spectral norm, tapering, thresholding.

Received August 2014.

*Discussed in 10.1214/15-EJS1018, 10.1214/15-EJS1081A, 10.1214/15-EJS1006 and 10.1214/15-EJS1019; rejoinder at 10.1214/15-EJS1081REJ.

†The research of Tony Cai was supported in part by NSF Grant DMS-1208982 and NIH Grant R01 CA 127334-05.

‡The research of Harrison Zhou was supported in part by NSF Grant DMS-1209191.

Contents

1. Introduction

Driven by a wide range of applications in many fields, from medicine to signal processing to climate studies and social science, high-dimensional statistical inference has emerged as one of the most important and active areas of current research in statistics. There have been tremendous recent efforts to develop new methodologies and theories for the analysis of high-dimensional data, whose dimension p can be much larger than the sample size n. The methodological

and theoretical developments in high-dimensional statistics are mainly driven by the important scientific applications, but also by the fact that some of these high-dimensional problems exhibit new features that are very distinct from those in the classical low-dimensional settings.

Covariance structure plays a particularly important role in high-dimensional data analysis. A large collection of fundamental statistical methods, including the principal component analysis, linear and quadratic discriminant analysis, clustering analysis, and regression analysis, require the knowledge of the covariance structure or some aspects thereof. Estimating a high-dimensional covariance matrix and its inverse, the precision matrix, is becoming a crucial problem in many applications including functional magnetic resonance imaging, analysis of gene expression arrays, risk management and portfolio allocation.

The standard and most natural estimator, the sample covariance matrix, performs poorly and can lead to invalid conclusions in the high-dimensional settings. For example, when $p/n \to c \in (0, \infty]$, the largest eigenvalue of the sample covariance matrix is not a consistent estimate of the largest eigenvalue of the population covariance matrix, and the eigenvectors of the sample covariance matrix can be nearly orthogonal to the truth. See Wachter (1976, 1978), Johnstone (2001), El Karoui (2003), Paul (2007), and Johnstone and Lu (2009). In particular, when $p > n$, the sample covariance matrix is not invertible, and thus cannot be applied in many applications that require estimation of the precision matrix.

To overcome the difficulty due to the high-dimensionality, structural assumptions are needed in order to estimate the covariance or precision matrix consistently. Various families of structured covariance and precision matrices have been introduced in recent years, including bandable covariance matrices, sparse covariance matrices, spiked covariance matrices, covariances with a tensor product structure, sparse precision matrices, bandable precision matrix via Cholesky decomposition, and latent graphical models. These different structural assumptions are motivated by various scientific applications, such as genomics, genetics, and financial economics. Many regularization methods have been developed accordingly to exploit the structural assumptions for estimation of covariance and precision matrices. These include the banding method in Wu and Pourahmadi (2009) and Bickel and Levina (2008a), tapering in Furrer and Bengtsson (2007) and Cai et al. (2010), thresholding in Bickel and Levina (2008b), El Karoui (2008) and Cai and Liu (2011a), penalized likelihood estimation in Huang et al. (2006), Yuan and Lin (2007), d'Aspremont et al. (2008), Banerjee et al. (2008), Rothman et al. (2008), Lam and Fan (2009), Ravikumar et al. (2011), and Chandrasekaran et al. (2012), regularizing principal components in Johnstone and Lu (2009), Zou et al. (2006), Cai, Ma, and Wu (2013), and Vu and Lei (2013), and penalized regression for precision matrix estimation in Meinshausen and Bühlmann (2006), Yuan (2010), Cai et al. (2011), Sun and Zhang (2013), and Ren et al. (2015).

Parallel to methodological advances on estimation of covariance and precision matrices there have been theoretical studies of the fundamental difficulty of the various estimation problems in terms of the minimax risks. Cai et al.

(2010) established the optimal rates of convergence for estimating a class of high-dimensional bandable covariance matrices under the spectral norm and Frobenius norm losses. Rate-sharp minimax lower bounds were obtained and a class of tapering estimators were constructed and shown to achieve the optimal rates. Cai and Zhou (2012a,b) considered the problems of optimal estimation of sparse covariance and sparse precision matrices under a range of losses, including the spectral norm and matrix ℓ_1 norm losses. Cai, Ren, and Zhou (2013) studied optimal estimation of a Toeplitz covariance matrix by using a method inspired by an asymptotic equivalence theory between the spectral density estimation and Gaussian white noise established in Golubev et al. (2010). Cai et al. (2015) solved the minimax estimation problem for a large class of sparse spiked covariance matrices under the spectral norm loss. Recently Ren et al. (2015) obtained fundamental limits on estimation of individual entries of a sparse precision matrix.

Standard techniques often fail to yield good results for many of these matrix estimation problems, and new tools are thus needed. In particular, for estimating sparse covariance matrices under the spectral norm, a new lower bound technique was developed in Cai and Zhou (2012b) that is particularly well suited to treat the "two-directional" nature of the covariance matrices, where one direction is along the rows and another along the columns. The result can be viewed as a generalization of Le Cam's method in one direction and Assouad's lemma in another. This new technical tool is useful for a range of other estimation problems. For example, it was used in Cai et al. (2012) for establishing the optimal rate of convergence for estimating sparse precision matrices and Tao et al. (2013) applied the technique to obtain the optimal rate for volatility matrix estimation.

The goal of the present paper is to provide a survey of these recent optimality results on estimation of structured high-dimensional covariance and precision matrices, and discuss some key technical tools that are used in the theoretical analyses. In addition, we will present data-driven adaptive procedures for estimation of various structured matrices. The main focus is on the bandable, Toeplitz, sparse, and sparse spiked covariance matrices as well as sparse precision matrices. Among these classes of matrices, the optimal procedures for estimating the bandable, Toeplitz, and sparse covariance matrices are obtained by "smoothing" or thresholding the sample covariance matrices based on various sparsity assumptions. In contrast, estimation of sparse spiked covariance matrices, which have sparse principal components, requires significantly different techniques to achieve optimality results. A few related problems such as sparse principal component analysis, factor models, and hypothesis testing on the covariance structure are also considered. Some open problems will be discussed at the end.

Throughout the paper, we assume that we observe a random sample $\{X^{(1)}, \ldots, X^{(n)}\}$ which consists of n independent copies of a p-dimensional random vector $X = (X_1, \ldots, X_p)'$ following some distribution with covariance matrix $\Sigma = (\sigma_{ij})$. The goal is to estimate the covariance matrix Σ and its inverse, the precision matrix $\Omega = \Sigma^{-1} = (\omega_{ij})$, based on the sample $\{X^{(1)}, \ldots, X^{(n)}\}$. Here for

ease of presentation we assume $\mathbb{E}(X) = 0$. This assumption is not essential. The non-centered mean case will be briefly discussed in Section 5.

Notations. Before we present a concise summary of the optimality results for estimating various structured covariance and precision matrices in this section, we introduce some basic notation that will be used in the rest of the paper. For any vector $x \in \mathbb{R}^p$, we use $||x||_\omega$ to denote its ℓ_ω norm with the convention that $||x|| = ||x||_2$. For any p by q matrix $M = (m_{ij}) \in \mathbb{R}^{p\times q}$, we use M' to denote its transpose. The matrix ℓ_ω operator norm is denoted by $||M||_{\ell_\omega} = \max_{||x||_\omega=1} ||Mx||_\omega$ with the convention $||M|| = ||M||_{\ell_2}$ for the spectral norm. Moreover, the entrywise ℓ_ω norm, that is, the ℓ_ω norm of M viewed as a vector, is denoted by $||M||_\omega$ and the Frobenius norm is represented by $||M||_F = ||M||_2 = (\sum_{i,j} m_{ij}^2)^{1/2}$. The submatrix with rows indexed by I and columns indexed by J is denoted by $M_{I,J}$. When the submatrix is a vector or a real number, we sometimes also use the lower case m instead of M. We use $||f||_\infty = \sup_x |f(x)|$ to denote the sup-norm of a function $f(\cdot)$, and $I\{A\}$ to denote the indicator function of an event A. We denote the covariance matrix of a random vector X by $\mathrm{Cov}(X)$ with the convention $\mathrm{Var}(X) = \mathrm{Cov}(X)$ when X is a random variable. For a symmetric matrix M, $M \succ 0$ means positive definiteness, $M \succeq 0$ means positive semi-definiteness and $\det(M)$ is its determinant. We use $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to denote its largest and smallest eigenvalues respectively. Given two sequences a_n and b_n, we write $a_n = O(b_n)$, if there is some constant $C > 0$ such that $a_n \le Cb_n$ for all n, and $a_n = o(b_n)$ implies $a_n/b_n \to 0$. The notation $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. The $n \times p$ dimensional data matrix is denoted by $\mathbf{X} = (X^{(1)}, \ldots, X^{(n)})'$ and the sample covariance matrix with known $\mathbb{E}(X) = 0$ is then defined as $\hat{\Sigma}_n = \mathbf{X}'\mathbf{X}/n = (\hat{\sigma}_{ij})$. For any index set $I \subseteq \{1, \ldots, p\}$, we denote by $\mathbf{X}_I$ the submatrix of $\mathbf{X}$ consisting of the columns of $\mathbf{X}$ indexed by I. The following definition of sub-Gaussian distributions is used throughout the paper.

Definition 1. The distribution of a random vector X is said to be sub-Gaussian with constant $\rho > 0$ if

$$\mathbb{P}\{|v'(X - \mathbb{E}X)| > t\} \le 2e^{-t^2\rho/2},$$

for all $t > 0$ and all deterministic unit vectors $\|v\| = 1$.

Several matrix norms are used for the loss functions and in the technical analysis. These include the spectral norm $\|\cdot\|$, matrix ℓ_1 operator norm $||\cdot||_{\ell_1}$, matrix ℓ_∞ operator norm $||\cdot||_{\ell_\infty}$ and Frobenius norm $||\cdot||_F$ among others. In particular, for symmetric matrices, the spectral norm is the largest singular value. The matrix ℓ_1 (ℓ_∞) operator norms are the largest column (row) sum of absolute values. Hence for symmetric matrices, these two norms coincide with each other. While the choice of the loss function varies according to specific applications and needs, the minimax behavior of the matrix estimation critically depends on the norm under which the error is measured. Matrix estimation under the Frobenius norm loss is essentially a vector estimation problem. What

really makes matrix estimation apart from it is those estimation problems under matrix operator norm losses, which are highly non-additive with respect to entries of the matrix. In particular, estimating covariance and precision matrices under the spectral norm loss brings in many new challenges as well as insights. We present the optimality results in the rest of this section under the Gaussian assumption with the focus on estimation under the spectral norm loss. More general settings will be discussed in Sections 2 and 3. A brief discussion on estimation under the Schatten q norm losses is given in Section 6.

1.1. Estimation of structured covariance matrices

We will consider in this paper optimal estimation of a range of structured covariance matrices, including bandable, Toeplitz, sparse and sparse spiked covariance matrices.

Bandable covariance matrices

The bandable covariance structure exhibits a natural "order" or "distance" among variables. This assumption is mainly motivated by time series with many scientific applications such as climatology and spectroscopy. See, for example, Friston et al. (1994) and Visser and Molenaar (1995). We consider settings where σ_{ij} is close to zero when $|i-j|$ is large. In other words, the variables X_i and X_j are nearly uncorrelated when the distance $|i-j|$ between them is large. The following parameter space was proposed in Bickel and Levina (2008a) (see also Wu and Pourahmadi (2003)),

$$\mathcal{F}_\alpha(M_0, M) = \left\{ \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le Mk^{-\alpha} \text{ for all } k, \text{ and } \lambda_{\max}(\Sigma) \le M_0 \right\}. \tag{1}$$

The parameter α specifies how fast the sequence σ_{ij} decays to zero as $j \to \infty$ for each fixed i. This can be viewed as the smoothness parameter of the class $\mathcal{F}_\alpha$, which is usually seen in nonparametric function estimation problems. In particular, for stationary processes, the parameter α is related to the smoothness of the corresponding spectral density function, which makes $\mathcal{F}_\alpha(M_0, M)$ a natural class of covariance matrices. See Grenander and Szegö (1958) for details. A larger α implies a smaller number of "effective" parameters in the model. Some other classes of bandable covariance matrices have also been considered in the literature, for example,

$$\mathcal{G}_\alpha(M_1) = \left\{ \Sigma_{p\times p} : |\sigma_{ij}| \le M_1(|i-j|+1)^{-\alpha-1} \right\}. \tag{2}$$

Note that $\mathcal{G}_\alpha(M_1) \subset \mathcal{F}_\alpha(M_0, M)$ if M_0 and M are sufficiently large. We will mainly focus on the larger class (1) in this paper. Assume that $p \le \exp(\epsilon n)$

for some constant $\epsilon > 0$, then the optimal rate of convergence for estimating the covariance matrix under the spectral norm loss over the class $\mathcal{F}_\alpha(M_0, M)$ is given as follows. See Cai et al. (2010).

Theorem 1 (Bandable Covariance Matrix)**.** *Suppose X is Gaussian. The minimax risk of estimating the covariance matrix over the bandable class given in (1) under the spectral norm loss is*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha(M_0,M)} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp \min\left\{\left(\frac{\log p}{n} + n^{-\frac{2\alpha}{2\alpha+1}}\right), \frac{p}{n}\right\}.$$

The minimax upper bound is derived by using a tapering estimator and the key rate $n^{-\frac{2\alpha}{2\alpha+1}}$ in the minimax lower bound is obtained by applying Assouad's lemma. We will discuss the important technical details in Sections 2.1 and 4.2 respectively. An adaptive block thresholding procedure, not depending on the knowledge of smoothness parameter α, is also introduced in Section 2.1.

Toeplitz covariance matrices

Toeplitz covariance matrix arises naturally in the analysis of stationary stochastic processes with a wide range of applications in many fields, including engineering, economics, and biology. See, for instance, Franaszczuk et al. (1985), Fuhrmann (1991) and Quah (2000) for specific applications. It can also be viewed as a special case of bandable covariance matrices. Similar decay or smoothness assumption like the one given in (1) is imposed, but each descending diagonal from left to right is constant for a Toeplitz matrix. Specifically, suppose the random vector $X = (X_1, \ldots, X_p)'$ consists of the first p variables of a fixed zero mean stationary process $\{X_i\}$ as $p \to \infty$. Then the Toeplitz covariance matrix Σ of X is uniquely determined by the autocovariance sequence $(\sigma_m) \equiv (\sigma_0, \sigma_1, \ldots, \sigma_{p-1}, \ldots)$ of $\{X_i\}$ with $\sigma_{ij} = \sigma_{|i-j|} = \mathbb{E}(X_i X_j)$.

It is well known that the spectrum of Toeplitz covariance matrix Σ is closely connected to the spectral density of the stationary process $\{X_i\}$ given by

$$f(x) = (2\pi)^{-1}\left[\sigma_0 + 2\sum_{m=1}^{\infty}\sigma_m \cos(mx)\right], \quad \text{for } x \in [-\pi, \pi].$$

See, e.g., Grenander and Szegö (1958). Motivated by time series applications, we consider the following class of Toeplitz covariance matrices $\mathcal{FT}_\alpha(M_0, M)$ defined in terms of the smoothness of the spectral density f. Let $\alpha = \gamma + \beta > 0$, where γ is the largest integer strictly less than α, $0 < \beta \leq 1$,

$$\mathcal{FT}_\alpha(M_0, M) = \left\{f : \|f\|_\infty \leq M_0 \text{ and } \left\|f^{(\gamma)}(\cdot + h) - f^{(\gamma)}(\cdot)\right\|_\infty \leq Mh^\beta\right\}. \tag{3}$$

In other words, the parameter space $\mathcal{FT}_\alpha(M_0, M)$ contains the Toeplitz covariance matrices whose corresponding spectral density functions are of Hölder

smoothness α. See, e.g., Parzen (1957) and Samarov (1977). Another parameter space, which directly specifies the decay rate of the autocovariance sequence (σ_m), has also been considered in the literature, see, Cai, Ren, and Zhou (2013). Under the assumption $(np/\log(np))^{\frac{1}{2\alpha+1}} < p/2$, the following theorem gives the optimal rate of convergence for estimating the Toeplitz covariance matrices over the class $\mathcal{FT}_\alpha(M_0, M)$ under the spectral norm loss.

Theorem 2 (Toeplitz Covariance Matrix)**.** *Suppose X is Gaussian. The minimax risk of estimating the Toeplitz covariance matrices over the class given in (3) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{FT}_\alpha(M_0,M)} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp \left(\frac{\log(np)}{np}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

The minimax upper bound is attained by a tapering procedure on certain estimators of autocovariance sequence, which is different from those banding estimators considered in Bickel and Levina (2008a) or Furrer and Bengtsson (2007). The minimax lower bound is established through the construction of a more informative model and an application of Fano's lemma. The essential technical details can be found in Sections 2.2 and 4.5 respectively. See Cai, Ren, and Zhou (2013) for further details.

Sparse covariance matrices

For estimating bandable and Toeplitz covariance matrices, one can take advantage of the information from the natural "order" on the variables. However, in many other applications such as genomics, there is no knowledge of distance or metric between variables, but the covariance between most pairs of the variables are often assumed to be insignificant. The class of sparse covariance matrices assumes that most of entries in each row and each column of the covariance matrix are zero or negligible. Compared to the previous two classes, there is no information on the "order" among the variables. We consider the following large class of sparse covariance matrices,

$$\mathcal{H}(c_{n,p}) = \left\{\Sigma : \max_{1\le i\le p} \sum_{j=1}^{p} \min\{(\sigma_{ii}\sigma_{jj})^{1/2}, \frac{|\sigma_{ij}|}{\sqrt{(\log p)/n}}\} \le c_{n,p}\right\}. \tag{4}$$

If an extra assumption that the variances σ_{ii} are uniformly bounded with $\max_i \sigma_{ii} \le \rho$ for some constant $\rho > 0$ is imposed, then $\mathcal{H}(c_{n,p})$ can be defined in terms of the maximal truncated ℓ_1 norm $\max_{1\le i\le p}\sum_{j=1}^{p}\min\{1, |\sigma_{ij}|(n/\log p)^{1/2}\}$, which has been considered in high-dimensional regression setting (see, for example, Zhang and Zhang (2012)).

For recovering the support of the sparse covariance matrices, it is natural to consider the following parameter space in which there are at most $c_{n,p}$ nonzero entries in each row/column of a covariance matrix,

$$\mathcal{H}_0(c_{n,p}) = \left\{\Sigma : \max_{1\le i\le p} \sum_{j=1}^{p} I\{\sigma_{ij} \ne 0\} \le c_{n,p}\right\}. \tag{5}$$

One important feature of the classes $\mathcal{H}(c_{n,p})$ and $\mathcal{H}_0(c_{n,p})$ is that they do not put any constraint on the variances σ_{ii}, $i = 1, \ldots, p$. Therefore the variances σ_{ii} can be in a very wide range and possibly $\max_i \sigma_{ii} \to \infty$. When the additional bounded variance condition $\max_i \sigma_{ii} \leq \rho$ for some constant $\rho > 0$ is imposed, it can be shown that the class $\mathcal{H}(c_{n,p})$ contains other commonly considered classes of sparse covariance matrices in the literature, including an ℓ_q ball assumption $\max_i \sum_{j=1}^p |\sigma_{ij}|^q \leq s_{n,p}$ in Bickel and Levina (2008b), and a weak ℓ_q ball assumption $\max_{1\leq j\leq p} \left\{\left|\sigma_{j[k]}\right|^q\right\} \leq s_{n,p}/k$ for each integer k in Cai and Zhou (2012a) where $\left|\sigma_{j[k]}\right|$ is the kth largest entry in magnitude of the jth row $(\sigma_{ij})_{1\leq i\leq p}$. More specifically, these two classes of sparse covariance matrices are contained in $\mathcal{H}(c_{n,p})$ with $c_{n,p} = C_q s_{n,p}(n/\log p)^{q/2}$ for some constant C_q depending on q only. The class $\mathcal{H}(c_{n,p})$ also covers the adaptive sparse covariance class $\mathcal{U}_q^*(s_{n,p})$ proposed in Cai and Liu (2011a), in which each row/column $(\sigma_{ij})_{1\leq i\leq p}$ is assumed to be in a weighted ℓ_q ball for $0 \leq q < 1$,i.e., $\max_i \sum_{j=1}^p (\sigma_{ii}\sigma_{jj})^{(1-q)/2} |\sigma_{ij}|^q \leq s_{n,p}$. Similarly $\mathcal{U}_q^*(s_{n,p}) \subset \mathcal{H}(c_{n,p})$ with $c_{n,p} = C_q s_{n,p}(n/\log p)^{q/2}$ for $0 \leq q < 1$. Although the class $\mathcal{H}(c_{n,p})$ is slightly larger than the classes defined via ℓ_q ball, weak ℓ_q ball under bounded variances condition and the class $\mathcal{U}_q^*(s_{n,p})$, the minimax risks of estimation over these different classes are the same. Indeed, the least favorable subclass of $\mathcal{H}(c_{n,p})$ chosen to establish the minimax lower bound is contained in other smaller classes. See Cai and Zhou (2012b) for further details on the minimax lower bound construction. It is worthwhile to point out that the class $\mathcal{H}(c_{n,p})$ is less general than the ones considered in El Karoui (2008) for which consistency results under the spectral norm loss were established. We advocate the larger sparse covariance class $\mathcal{H}(c_{n,p})$ in this paper not only because it contains almost all other classes considered in the literature, but also the comparison between the noise level $((\sigma_{ii}\sigma_{jj}\log p)/n)^{1/2}$ and the signal level $|\sigma_{ij}|$ captures the essence of the sparsity of the model.

Under some mild conditions $1 \leq c_{n,p} \leq C\sqrt{n/(\log p)^3}$ and $p \geq n^\phi$ for some constants ϕ and $C > 0$, the optimal rate of convergence for estimating sparse covariance matrices over the class $\mathcal{H}(c_{n,p})$ under the spectral norm is given as follows. See Cai and Zhou (2012b).

Theorem 3 (Sparse Covariance Matrix)**.** *Suppose X is Gaussian. The minimax risk of estimating a sparse covariance matrix over the class $\mathcal{H}(c_{n,p})$ given in (4) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{H}(c_{n,p})} \mathbb{E} \left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp c_{n,p}^2 \frac{\log p}{n}.$$

The distributional assumption on X can be significantly relaxed. See Section 2.3 for further details. Besides, an adaptive thresholding estimator is constructed in Section 2.3, and it is shown to be adaptive to the variability of the individual entries and attains the minimax upper bound. The lower bound argument for estimation under the spectral norm was given in Cai and Zhou (2012b) by applying Le Cam-Assouad's method, which is introduced in Section 4.1.

Sparse spiked covariance matrices

Spiked covariance matrix

$$\Sigma = \sum_{i=1}^{r} \lambda_r v_i v_i' + I, \tag{6}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and the vectors $v_1, \ldots, v_r$ are orthonormal, arises naturally in principal component analysis as well as factor models with homoscedastic noise. In particular, the first r eigenvalues of Σ are strictly larger than 1, the remaining eigenvalues. Since the spectrum of Σ has r spikes, (6) was first named spiked covariance model in Johnstone (2001).

Before defining the sparse spiked covariance structure, we introduce some notation. Let the set of p by r matrices with orthonormal columns be $O(p,r) = \{V \in \mathbb{R}^{p\times r} : V'V = I\}$. For any such $V = (v_1, v_2, \ldots, v_r) \in O(p,r)$, we denote its ith row by $V_{i,*}$. The diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. The class of sparse spiked covariance matrices imposes the sparsity on the rows of V. Specifically, we define

$$\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p}) = \left\{ \begin{array}{c} \Sigma = V\Lambda V' + I : 0 < \lambda_r \leq \cdots \leq \lambda_1 \leq \lambda_{n,p}, \\ V \in O(p,r), \sum_{j=1}^{p} I\{V_{j,*} \neq \mathbf{0}\} \leq c_{n,p} \end{array} \right\}, \tag{7}$$

where $1 \leq r_{n,p} \leq c_{n,p} \leq p$. More concretely, the group-sparse structure is imposed on the r leading eigenvectors of Σ in $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ and the number of nonzero rows of V is no more than $c_{n,p}$. As a consequence, as the sum of a sparse matrix and an identity matrix, the covariance matrix Σ itself is also sparse. In particular, there are no more than $c_{n,p}$ nonzero entries in each row and each column of Σ and hence $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p}) \subset \mathcal{H}_0(c_{n,p})$ defined in (5). However, compared with the previous three structures, sparse spiked covariance matrix has a very different structural component, the rank r matrix $\sum_{i=1}^{r} \lambda_r v_i v_i'$. This extra low-rank structure, together with the sparsity assumption yields a faster minimax rate of convergence for estimating the covariance matrices over $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ than that over sparse covariance class with sparsity $c_{n,p}$. Under the assumption $\frac{c_{n,p}}{n} \log \frac{ep}{c_{n,p}} \leq c$ for some sufficiently small constant $c > 0$, the optimal rate of convergence for estimating sparse spiked covariance matrices over the class $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ under the spectral norm is given as follows. See Cai et al. (2015).

Theorem 4 (Sparse Spiked Covariance Matrix)**.** *Suppose X is Gaussian. The minimax risk of estimating the sparse spiked covariance matrices over the class given in (7) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|^2 \asymp \min \left\{ \frac{(\lambda_{n,p} + 1)\, c_{n,p}}{n} \log \frac{ep}{c_{n,p}} + \frac{\lambda_{n,p}^2 r_{n,p}}{n}, \lambda_{n,p}^2 \right\}.$$

The minimax risk essentially has two terms $\frac{(\lambda_{n,p}+1)c_{n,p}}{n} \log \frac{ep}{c_{n,p}}$ and $\frac{\lambda_{n,p}^2 r_{n,p}}{n}$. The first term does not involve $r_{n,p}$ and is the same bound in the rank-one

$r = 1$ setting, in which estimation can be reduced into a single sparse vector estimation problem. The second term is the oracle risk when the support of V is known and can be treated as the risk in r dimensional setting. The minimax upper bound is attained by a global searching scheme for the support of V which is constructed in Section 2.4. The minimax lower bounds of the two terms are shown through applications of Fano's lemma and Le Cam's method respectively, which are introduced in Sections 4.1. See Cai et al. (2015) for further details.

1.2. Estimation of structured precision matrices

In addition to covariance matrix estimation, there is also significant interest in estimation of its inverse, the precision matrix, under the structural assumptions on the precision matrix itself. Precision matrix is closely connected to the undirected Gaussian graphical model, which is a powerful tool to model the relationships among a large number of random variables in a complex system and is used in a wide array of scientific applications. See, for instance, Wille et al. (2004) and Runge et al. (2014), for some recent applications in genomics and climate studies. It is well known that recovering the graph structure $G = (V, E)$ of an undirected Gaussian graphical model is equivalent to recovering the support of the precision matrix. In fact, if $X \sim N\left(0, \Omega^{-1}\right)$ is a graphical model with respect to G, then the entry ω_{ij} is zero if and only if the variables X_i and X_j are conditionally independent given all the remaining variables, which is equivalent to the edge $(i, j) \notin E$. (See, e.g., Lauritzen (1996).) Consequently, a sparse graph corresponds to a sparse precision matrix. We thus focus on estimation of sparse precision matrices and present the optimality results under the Gaussian assumption.

Sparse precision matrices and Gaussian graphical model

The class of sparse precision matrices assumes that most of entries in each row/column of the precision matrix are zero or negligible. The class of sparse precision matrices $\mathcal{HP}(c_{n,p}, M)$ introduced in this section is similar to the class of sparse covariance matrices defined in (4), where the sparsity is modeled by a truncated ℓ_1 type norm. We also assume the spectra of Ω are bounded from below and $\max_i \sigma_{ii}$ is bounded above for simplicity. More specifically, we define $\mathcal{HP}(c_{n,p}, M)$ by

$$\mathcal{HP}(c_{n,p}, M) = \left\{ \begin{array}{c} \Omega : \max_{1\le i\le p} \sum_{j\neq i} \min\{1, \frac{|\omega_{ij}|}{\sqrt{(\log p)/n}}\} \le c_{n,p}, \\ \frac{1}{M} \le \lambda_{\min}(\Omega), \max_i \sigma_{ii} \le M, \Omega \succ 0 \end{array} \right\}, \tag{8}$$

where M is some universal constant and the sparsity parameter $c_{n,p}$ is allowed to grow with $p, n \to \infty$.

This class of precision matrices was proposed in Ren et al. (2015) and contains similar classes proposed in Cai et al. (2011) and Cai et al. (2012), in which an extra matrix ℓ_1 norm bound $M_{n,p}$ is included in the definition. As a special case

where each $|\omega_{ij}|$ is either zero or above the level $(n/\log p)^{-1/2}$, such a matrix in $\mathcal{HP}(c_{n,p}, M)$ has at most $c_{n,p}$ nonzero entries on each row/column which is called the maximum node degree of Ω in the Gaussian graphical model. We define the following class for the support recovery purpose,

$$\mathcal{HP}_0(c_{n,p}, M) = \left\{ \begin{array}{c} \Omega : \max_{1\le i\le p} \sum_{j=1}^{p} I\{\omega_{ij} \neq 0\} \le c_{n,p}, \\ \frac{1}{M} \le \lambda_{\min}(\Omega), \max_i \sigma_{ii} \le M, \Omega \succ 0 \end{array} \right\}. \tag{9}$$

The following theorem provides the optimal rate of convergence for estimating sparse precision class under the spectral norm loss.

Theorem 5 (Sparse Precision Matrix)**.** *Suppose X is Gaussian. Assume that $1 \le c_{n,p} \le C\sqrt{n/(\log p)^3}$. The minimax risk of estimating the sparse precision matrix over the class $\mathcal{HP}(c_{n,p}, M)$ given in (8) satisfies*

$$\inf_{\hat{\Omega}} \sup_{\mathcal{HP}(c_{n,p},M)} \mathbb{E} \left\| \hat{\Omega} - \Omega \right\|^2 \asymp c_{n,p}^2 \frac{\log p}{n}.$$

In Section 3.1, we establish the minimax upper bound via a neighborhood regression approach. The estimator ANT introduced in Section 3.2 also achieves this optimal rate. The lower bound argument is provided by applying Le Cam-Assouad's method developed in Cai et al. (2012) which is discussed in Section 4.4.

For estimating sparse precision matrices, besides the minimax risk under the spectral norm, it is also important to understand the minimax risk of estimating individual entries of the precision matrix. The solution is not only helpful for the support recovery problem but also makes important advancements in the understanding of statistical inference of low-dimensional parameters in a high-dimensional setting. See Ren et al. (2015).

Theorem 6 (Entry of Sparse Precision Matrix)**.** *Suppose X is Gaussian. Assume that $(c_{n,p} \log p)/n = o(1)$. The minimax risk of estimating ω_{ij} for each i, j over the sparse precision class given in (8) is*

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{HP}(c_{n,p},M)} \mathbb{E} |\hat{\omega}_{ij} - \omega_{ij}| \asymp \max\left\{ c_{n,p} \frac{\log p}{n}, \sqrt{\frac{1}{n}} \right\}.$$

The minimax upper bound is based on a multivariate regression approach given in Section 3.2 while Le Cam's lemma is used to show the minimax lower bound in Section 4.3.

1.3. Organization of the paper

The rest of the paper is organized as follows. Section 2 presents several minimax and adaptive procedures for estimating various structured covariance matrices and establishes the corresponding minimax upper bounds under the spectral norm loss. Estimation under the factor models and sparse principal component

analysis are also discussed. Section 3 considers minimax and adaptive estimation of sparse precision matrices under the spectral norm loss. Section 3 also discusses inference on the individual entries of a sparse precision matrix and the latent graphical model. Section 4 focuses on the lower bound arguments in matrix estimation problems. It begins with a review of general lower bound techniques and then applies the tools to establish rate-sharp minimax lower bounds for various covariance and precision matrix estimation problems. The upper and lower bounds together yield immediately the optimal rates of convergence stated earlier in this section. Section 5 briefly discusses the non-centered mean case and the positive semi-definite issue in covariance and precision matrix estimation. Hypothesis testing on the covariance structure is also discussed. The paper is concluded with a discussion on some open problems on estimating covariance and precision matrices as well as their functionals in Section 6.

2. Estimation of structured covariance matrices

This section focuses on estimation of structured covariance matrices. Minimax upper bounds and adaptive procedures are introduced. Many estimators are based on "smoothing" the sample covariance matrices. These include the banding, tapering and thresholding estimators. The optimal estimator for sparse spiked covariance matrices, however, relies on a global searching scheme. Estimation of precision matrices is considered in the next section.

2.1. Bandable covariance matrices

Minimax upper bound

Bickel and Levina (2008a) introduced the class of bandable covariance matrices $\mathcal{F}_\alpha(M_0, M)$ given (1) and proposed a banding estimator

$$\hat{\Sigma}_{B,k} = (\hat{\sigma}_{ij} I\{|i-j| \le k\}) \tag{10}$$

based on the sample covariance matrix $\hat{\Sigma}_n = (\hat{\sigma}_{ij})$ for estimating a covariance matrix $\Sigma \in \mathcal{F}_\alpha(M_0, M)$. The bandwidth k was chosen to be $k_B = \left(\frac{\log p}{n}\right)^{\frac{1}{2(\alpha+1)}}$ and the rate of convergence $\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}$ for estimation under the spectral norm loss was proved under the sub-Gaussian assumption on $X = (X_1, \ldots, X_p)'$. It was unclear if this rate is optimal.

Cai et al. (2010) further studied the optimal estimation problem for the classes $\mathcal{F}_\alpha(M_0, M)$ in (1) and $\mathcal{G}_\alpha(M_1)$ in (2) under the sub-Gaussian assumption. A tapering estimator was proposed. Tapering estimators have been effectively used in the literature of climate studies. See, for instance, Gaspari and Cohn (1999), Houtekamer and Mitchell (2001) and Hamill et al. (2001). Furrer et al. (2006) and Furrer and Bengtsson (2007) previously also considered tapering estimators for estimating covariance matrices but in different settings.

Specifically, for a given even positive integer $k \leq p$, let $\omega = (\omega_m)_{0\leq m\leq p-1}$ be a weight sequence with ω_m given by

$$\omega_m = \begin{cases} 1, & \text{when } m \leq k/2 \\ 2 - \frac{2m}{k}, & \text{when } k/2 < m \leq k \\ 0, & \text{Otherwise} \end{cases}. \tag{11}$$

The tapering estimator $\hat{\Sigma}_{T,k}$ of the covariance matrix Σ is defined by

$$\hat{\Sigma}_{T,k} = (\hat{\sigma}_{ij}\omega_{|i-j|}).$$

It was shown that the tapering estimator with bandwidth $k_T = \min\{n^{\frac{1}{2\alpha+1}}, p\}$ gives the following rate of convergence under the spectral norm.

Theorem 7 (Cai et al. (2010))**.** *Suppose that X is sub-Gaussian distributed with some finite constant. Then the tapering estimator $\hat{\Sigma}_{T,k}$ with $k_T = \min\{n^{\frac{1}{2\alpha+1}}, p\}$ satisfies*

$$\sup_{\mathcal{F}_\alpha(M_0,M)} \mathbb{E}\left\|\hat{\Sigma}_{T,k_T} - \Sigma\right\|^2 \leq \min\left\{C\left(\frac{\log p}{n} + n^{-\frac{2\alpha}{2\alpha+1}}\right), C\frac{p}{n}\right\}. \tag{12}$$

Theorem 7 clearly also holds for $\mathcal{G}_\alpha(M_1)$, a subspace of $\mathcal{F}_\alpha(M_0, M)$. Note that the rate given in (12) is faster than the rate $((\log p)/n)^{\alpha/(\alpha+1)}$ obtained in Bickel and Levina (2008a) for the banding estimator $\hat{\Sigma}_{B,k}$ with the bandwidth $k_B = \left(\frac{\log p}{n}\right)^{\frac{1}{2(\alpha+1)}}$, which implies that this banding estimator is sub-optimal. A minimax lower bound is also established in Cai et al. (2010), which shows that the rate of convergence in (12) is indeed optimal. We will discuss this minimax lower bound argument in Section 4.2.

There are two key steps in the technical analysis of the tapering estimator. In the first step, it is shown that the tapering estimator $\hat{\Sigma}_{T,k}$ has a simple representation and can be written as the average of many small disjoint submatrices of size no more than k in the sample covariance matrix $\hat{\Sigma}_n$. Consequently, the distance $\left\|\hat{\Sigma}_{T,k_T} - \Sigma\right\|$ can be bounded by the maximum of distances of these submatrices from their respective means. The second key step involves the application of a large deviation result for sample covariance matrix of relatively small size under the spectral norm. This random matrix result, stated in the following lemma, is a commonly used technical tool in high-dimensional statistical problems. See Cai et al. (2010) for further details.

Lemma 1. *Suppose $Y = (Y_1, \ldots, Y_k)'$ is sub-Gaussian with constant $\rho > 0$ and with mean 0 and covariance matrix Σ. Let $Y^{(1)}, \ldots, Y^{(n)}$ be n independent copies of Y. Then there exist some universal constant $C > 0$ and some constant ρ_1 depending on ρ, such that the sample covariance matrix of $\{Y^{(1)}, \ldots, Y^{(n)}\}$, $\hat{\Sigma}_n^Y$, satisfies*

$$\mathbb{P}\left(\left\|\hat{\Sigma}_n^Y - \Sigma\right\| > t\right) \leq 2\exp(-nt^2\rho_1 + Ck),$$

for all $0 < t < \rho_1$.

See also Davidson and Szarek (2001) for more refined results under the Gaussian assumption.

For the banding estimator, Bickel and Levina (2008a) used the matrix ℓ_1 norm as the upper bound to control the spectral norm. Then bounding the risk under the spectral norm can be turned into bounding the error on each row of $\hat{\Sigma}_n$ under the vector ℓ_1 norm, which is an easier task. An analysis of the bias and variance trade-off then leads to their choice of bandwidth $k_B = (n/\log p)^{\frac{1}{2(\alpha+1)}}$. The loose control of spectral norm by the matrix ℓ_1 norm is the main reason for the sub-optimal result in Bickel and Levina (2008a) under the spectral norm loss. It is worthwhile to point out that the result in Bickel and Levina (2008a) is still not optimal under the matrix ℓ_1 norm loss due to the sub-optimal choice of the bandwidth. See the discussion on estimation under other losses at the end of this section. An interesting question is whether the banding estimator with a different bandwidth is also optimal. Indeed it can be shown that the banding estimator $\hat{\Sigma}_{B,k}$ with the bandwidth $k = \min\{n^{\frac{1}{2\alpha+1}}, p\}$ is rate-optimal. See, for example, Xiao and Bunea (2014) for a detailed calculation under the Gaussian assumption.

Adaptive estimation through block thresholding

It is evident that the construction of the optimal tapering estimator $\hat{\Sigma}_{T,k_T}$ requires the explicit knowledge of the decay rate α which is usually unknown in practice. Cai and Yuan (2012) considered the adaptive estimation problem and constructed a data-driven block thresholding estimator, not depending on α, M_0, M or M_1, that achieves the optimal rate of convergence simultaneously over the parameter spaces $\mathcal{F}_\alpha(M_0, M)$ and $\mathcal{G}_\alpha(M_1)$ for all $\alpha > 0$.

The construction of the adaptive estimator consists of two steps. In the first step, we divide the sample covariance matrix into blocks of increasing sizes as they move away from the diagonal, suggested by the decay structure of the bandable covariance matrices. The second step is to simultaneous kill or keep all the entries within each block to construct the final estimator, where the thresholding levels are chosen adaptively for different blocks. The underlying idea is to mimic the analysis for the tapering estimator $\hat{\Sigma}_{T,k_T}$ in the sense that the final estimator can be decomposed as the sum of small submatrices. However, the choice of the bandwidth is chosen adaptively here through using the block thresholding strategy, where the threshold rule is established by a novel norm compression inequality. See Theorem 3.4 of Cai and Yuan (2012) for details.

Now we briefly introduce the construction of blocks. First, we construct disjoint square blocks of size $k_{ad} \asymp \log p$ along the diagonal. Second, a new layer of blocks of size k_{ad} are created towards the top right corner along the diagonal next to the previous layer. In particular, this layer of blocks has either two or one block (of size k_{ad}) in an alternating fashion (see Figure 1). After this step, we note that the odd rows of blocks has three blocks of size k_{ad} and even rows of blocks have two blocks of size k_{ad}. This creates space to double the size of the blocks in the next step. We then repeat the first and second steps building

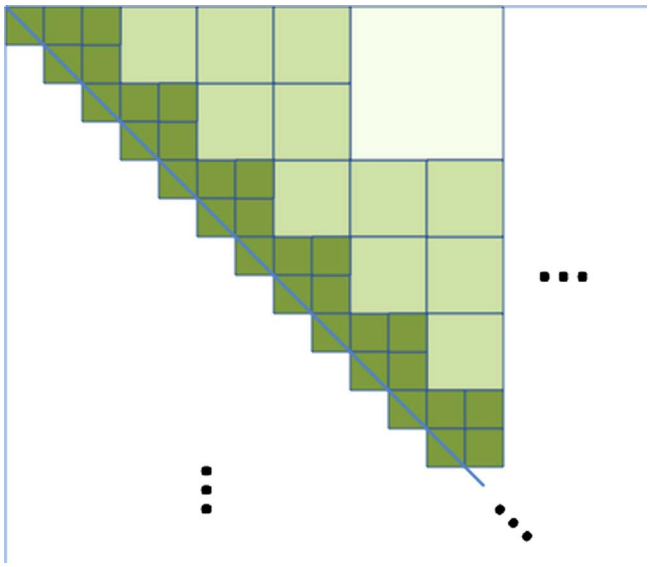

FIG 1. *Construction of blocks with increasing dimensions away from the diagonal.*

on the previous layer of blocks but double the size of the blocks until the whole upper half of the matrix is covered. In the end, the same blocking construction is done for the lower half of the matrix. It is possible that the last row and last column of the blocks are rectangular instead of being square. For the sake of brevity, we omit the discussion on these blocks. See Cai and Yuan (2012) for further details. By the construction, the blocks form a partition of $\{1, \ldots, p\}^2$. We list the indices of those blocks by $\mathcal{B} = \{B_1, \ldots, B_N\}$, assuming there are N blocks in total. The construction of the blocks is illustrated in Figure 1.

Once the blocks $\mathcal{B}$ are constructed, we define the final adaptive estimator $\hat{\Sigma}_{CY}^{Ada}$ by the following thresholding procedure on the blocks of the sample covariance matrix $\hat{\Sigma}_n$. First, we keep the diagonal blocks of size k_{ad} constructed at the very beginning, i.e., $(\hat{\Sigma}_{CY}^{Ada})_B = (\hat{\Sigma}_n)_B$ for those B in the diagonal. Second, we set all the large blocks to 0, i.e., $(\hat{\Sigma}_{CY}^{Ada})_B = 0$ for all $B \in \mathcal{B}$ with $\text{size}(B) > n/\log n$, where for $B = I \times J$, $\text{size}(B)$ is defined to be $\max\{|I|, |J|\}$. In the end, we threshold the intermediate blocks adaptively according to their spectral norm as follows. Suppose $B = I \times J$. Then $(\hat{\Sigma}_{CY}^{Ada})_B = (\hat{\Sigma}_n)_B I\{||(\hat{\Sigma}_n)_B|| > \lambda_B\}$, where

$$\lambda_B = 6(||(\hat{\Sigma}_n)_{I\times I}|| \cdot ||(\hat{\Sigma}_n)_{J\times J}||(\text{size}(B) + \log p)/n)^{1/2}.$$

Finally we obtain the adaptive estimator $\hat{\Sigma}_{CY}^{Ada}$, which is rate optimal over the classes $\mathcal{F}_\alpha(M_0, M)$ or $\mathcal{G}_\alpha(M_1)$.

Theorem 8 (Cai and Yuan (2012))**.** *Suppose that X is sub-Gaussian distributed with some finite constant. Then for all $\alpha > 0$, the adaptive estimator $\hat{\Sigma}_{CY}^{Ada}$*

constructed above satisfies

$$\sup_{\mathcal{F}_\alpha(M_0,M)} \mathbb{E}\left\|\hat{\Sigma}_{CY}^{Ada} - \Sigma\right\|^2 \le \min C\left\{\left(\frac{\log p}{n} + n^{-\frac{2\alpha}{2\alpha+1}}\right), \frac{p}{n}\right\},$$

where the constant C depends on α, M_0 and M.

In light of the optimal rate of convergence given in Theorem 1, this shows that the block thresholding estimator adaptively achieves the optimal rate over $\mathcal{F}_\alpha(M_0, M)$ for all $\alpha > 0$.

Estimation under other losses

In addition to estimating bandable covariance matrices under the spectral norm, estimation under other losses has also been considered in literature. Furrer and Bengtsson (2007) introduced a general tapering estimator, which gradually shrinks the off-diagonal entries toward zero with certain weights, to estimate covariance matrices under the Frobenius norm loss in different settings. Cai et al. (2010) studied the problem of optimal estimation under the Frobenius norm loss over the classes $\mathcal{F}_\alpha\,(M_0, M)$ and $\mathcal{G}_\alpha\,(M_1)$, and Cai and Zhou (2012a) investigated optimal estimation under the matrix ℓ_1 norm.

The optimal estimators for both the Frobenius norm and matrix ℓ_1 norm are again based on the tapering estimator (or banding estimator) with the bandwidth $k_{T,F} = n^{\frac{1}{2(\alpha+1)}}$ and $k_{T,1} = \min\left\{n^{\frac{1}{2(\alpha+1)}}, (n/\log p)^{\frac{1}{2\alpha+1}}\right\}$ respectively. The rate of convergence under the Frobenius norm of $\mathcal{F}_\alpha(M_0, M)$ is different from that of $\mathcal{G}_\alpha(M_1)$. The optimal rate of convergence under the matrix ℓ_1 norm is $\min\left\{\left((\log p/n)^{\frac{2\alpha}{2\alpha+1}} + n^{-\frac{\alpha}{\alpha+1}}\right), p^2/n\right\}$. Their corresponding minimax lower bounds are established in Cai et al. (2010) and Cai and Zhou (2012a). Comparing these estimation results, it should be noted that although the optimal estimators under different norms are all based on tapering or banding, the best bandwidth critically depends on the norm under which the estimation accuracy is measured.

2.2. Toeplitz covariance matrices

We turn to optimal estimation of Toeplitz covariance matrices. Recall that if X is a stationary process with autocovariance sequence (σ_m), then the covariance matrix $\Sigma_{p\times p}$ has the Toeplitz structure such that $\sigma_{ij} = \sigma_{|i-j|}$.

Wu and Pourahmadi (2009) introduced and studied a banding estimator based on the sample autocovariance matrix, and McMurry and Politis (2010) extended their results to tapering estimators. Xiao and Wu (2012) further improved these two results and established sharp banding estimator under the spectral norm. All these results are obtained in the framework of causal representation and physical dependence measures proposed in Wu (2005). One important assumption is that there is only one realization available, i.e. $n = 1$.

More recently, Cai, Ren, and Zhou (2013) considered the problem of optimal estimation of Toeplitz covariance matrices over the class $\mathcal{FT}_\alpha(M_0, M)$. The result is valid not only for fixed sample size but also for the general case when $n, p \to \infty$.

An optimal tapering estimator is constructed by tapering the sample autocovariance sequence. More specifically, recall $\hat{\Sigma}_n = (\hat{\sigma}_{ij})$ is the sample covariance matrix. For $0 \le m \le p-1$, set

$$\tilde{\sigma}_m = (p-m)^{-1} \sum_{s-t=m} \hat{\sigma}_{st},$$

which is an unbiased estimator of σ_m. The estimator $\tilde{\sigma}_m$ was also proposed in Bickel and Levina (2004). Then the tapering Toeplitz estimator $\hat{\Sigma}_{T,k}^{Toep} = (\breve{\sigma}_{st})$ of Σ with bandwidth k is defined as $\breve{\sigma}_{st} = \omega_{|s-t|}\tilde{\sigma}_{|s-t|}$, where the weight ω_m is defined in (11). By picking the best choice of the bandwidth $k_{To} = (np/\log(np))^{\frac{1}{2\alpha+1}}$, the optimal rate of convergence is given as follows.

Theorem 9 (Cai, Ren, and Zhou (2013))**.** *Suppose that X is Gaussian distributed with a Toeplitz covariance matrix $\Sigma \in \mathcal{FT}_\alpha(M_0, M)$ and suppose $(np/\log(np))^{\frac{1}{2\alpha+1}} \le p/2$. Then the tapering estimator $\hat{\Sigma}_{T,k}^{Toep}$ with $k_{To} = (np/\log(np))^{\frac{1}{2\alpha+1}}$ satisfies*

$$\sup_{\Sigma \in \mathcal{FT}_\alpha(M_0,M)} \mathbb{E} \left\| \hat{\Sigma}_{T,k_{To}}^{Toep} - \Sigma \right\|^2 \le C \left(\frac{\log(np)}{np} \right)^{\frac{2\alpha}{2\alpha+1}}.$$

The performance of the banding estimator was also considered in Cai, Ren, and Zhou (2013). Surprisingly, the best banding estimator is inferior to the optimal tapering estimator over the class $\mathcal{FT}_\alpha(M_0, M)$, which is due to a larger bias term caused by the banding estimator. See Cai, Ren, and Zhou (2013) for further details. At a high level, the upper bound proof is based on the fact that the spectral norm of a Toeplitz matrix can be bounded above by the sup-norm of the corresponding spectral density function. which leads to the appearance of the logarithmic term in the upper bound. A lower bound argument will be introduced in Section 4.5. The appearance of the logarithmic term indicates the significant difference in the technical analyses between estimating bandable and Toeplitz covariance matrices.

2.3. Sparse covariance matrices

We now consider optimal estimation of another important class of structured covariance matrices – sparse covariance matrices. This problem has been considered by Bickel and Levina (2008b), Rothman et al. (2009) and Cai and Zhou (2012a,b). These works assume that the variances are uniformly bounded, i.e., $\max_i \sigma_{ii} \le \rho$ for some constant $\rho > 0$. Under such a condition, a universal thresholding estimator $\hat{\Sigma}^{U,Th} = \left(\hat{\sigma}_{ij}^{U,Th}\right)$ is proposed with $\hat{\sigma}_{ij}^{U,Th} =$

$\hat{\sigma}_{ij} \cdot I\{|\hat{\sigma}_{ij}| \geq \gamma (\log p/n)^{1/2}\}$ for some sufficiently large constant $\gamma = \gamma(\rho)$. This estimator is not adaptive as it depends on the unknown parameter ρ. The rates of convergence $s_{n,p} (\log p/n)^{(1-q)/2}$ under the spectral norm and matrix ℓ_1 norm in probability are obtained over the classes of sparse covariance matrices, in which each row/column is in an ℓ_q ball or a weak ℓ_q ball, $0 \leq q < 1$. i.e. $\max_i \sum_{j=1}^p |\sigma_{ij}|^q \leq s_{n,p}$ or $\max_{1\leq j\leq p} \left\{\left|\sigma_{j[k]}\right|^q\right\} \leq s_{n,p}/k$ respectively.

Sparse covariance matrices: Adaptive minimax upper bound

Estimation of a sparse covariance matrix is intrinsically a heteroscedastic problem in the sense that the variances of the entries of the sample covariance matrix $\hat{\sigma}_{ij}$ can vary over a wide range. A universal thresholding estimator essentially treats the problem as a homoscedastic problem and does not perform well in general. Cai and Liu (2011a) proposed a data-driven estimator $\hat{\Sigma}^{Ad,Th}$ which adaptively thresholds the entries according to their individual variabilities,

$$\begin{aligned} \hat{\Sigma}^{Ad,Th} &= (\hat{\sigma}_{ij}^{Ad,Th}),\ \hat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_{ki}X_{kj} - \hat{\sigma}_{ij})^2 \\ \hat{\sigma}_{ij}^{Ad,Th} &= \hat{\sigma}_{ij} \cdot I\{|\hat{\sigma}_{ij}| \geq \delta(\hat{\theta}_{ij} \log p/n)^{1/2}\}. \end{aligned} \tag{13}$$

The advantages of this adaptive procedure are that it is fully data-driven and no longer requires the variances σ_{ii} to be uniformly bounded. The estimator $\hat{\Sigma}^{Ad,Th}$ attains the following rate of convergence adaptively over sparse covariance classes $\mathcal{H}(c_{n,p})$ defined in (4) under the matrix ℓ_ω norms for all $\omega \in [1, \infty]$.

Theorem 10 (Sparse Covariance Matrices)**.** *Suppose each standardized component $Y_i = X_i/\sigma_{ii}^{1/2}$ is sub-Gaussian distributed with some finite constant and a mild condition $\min_{ij} \mathrm{Var}(Y_iY_j) \geq c_0$ for some positive constant c_0. Then the adaptive estimator $\hat{\Sigma}^{Ad,Th}$ constructed above in (13) satisfies*

$$\inf_{\mathcal{H}(c_{n,p})} \mathbb{P}\left(\|\hat{\Sigma}^{Ad,Th} - \Sigma\|_{\ell_\omega} \leq C c_{n,p}((\log p)/n)^{1/2}\right) \geq 1 - O((\log p)^{-1/2} p^{-\delta+2}), \tag{14}$$

where $\omega \in [1, \infty]$.

A similar idea was applied in the practical implementation without theoretical justification in El Karoui (2008). The proof of Theorem 10 essentially follows from the analysis in Cai and Liu (2011a) for $\hat{\Sigma}^{Ad,Th}$ under the matrix ℓ_1 norm over a smaller class of sparse covariance matrices $\mathcal{U}_q^*(s_{n,p})$, where each row/column $(\sigma_{ij})_{1\leq i\leq p}$ is assumed to be in a weighted ℓ_q ball, $\max_i(\sigma_{ii}\sigma_{jj})^{(1-q)/2} |\sigma_{ij}|^q \leq s_{n,p}$. That result is automatically valid for all matrix ℓ_ω norms, following the claim in Section 6 of Cai and Zhou (2012b), where the Riesz-Thorin interpolation theorem is applied. The key technical tool in the analysis is the following large deviation result for self-normalized entries of the sample covariance matrix.

Lemma 2. *Under the assumptions of Theorem 10, for any small $\varepsilon > 0$,*

$$\mathbb{P}(\,|\hat{\sigma}_{ij} - \sigma_{ij}|\,/\hat{\theta}_{ij}^{1/2} \geq \sqrt{\alpha(\log p)/n}) = O((\log p)^{-1/2} p^{-(\alpha/2-\epsilon)}),$$

Lemma 2 follows from a moderate deviation result in Shao (1999). See Cai and Liu (2011a) for further details.

Theorem 10 states the rate of convergence in probability. The same rate of convergence holds in expectation with a mild sample size condition $p \geq n^{\phi}$ for some $\phi > 0$. A minimax lower bound argument under the spectral norm is provided in Cai and Zhou (2012b) (see also Section 4.4), which shows that $\hat{\Sigma}^{Ad,Th}$ is indeed rate optimal under all matrix ℓ_ω norms. In contrast, the universal thresholding estimator $\hat{\Sigma}^{U,Th}$ is sub-optimal over $\mathcal{H}(c_{n,p})$ due to the possibility of $\max_i(\sigma_{ii}) \to \infty$.

Besides the matrix ℓ_ω norms, Cai and Zhou (2012b) considered a unified result on estimating sparse covariance matrices under a class of Bregman divergence losses which include the commonly used Frobenius norm as a special case. Following a similar proof there, it can be shown that the estimator $\hat{\Sigma}^{Ad,Th}$ also attains the optimal rate of convergence under the Bregman divergence losses over the large parameter class $\mathcal{H}(c_{n,p})$. In addition to the hard thresholding estimator introduced above, Rothman et al. (2009) considered a class of thresholding rules with more general thresholding functions including soft thresholding, SCAD and adaptive Lasso. It is straightforward to extend all results above to this setting. Therefore, the choice of the thresholding function is not important as far as the rate optimality is concerned.

It is worthwhile to point out that all related results hold under assumptions on the marginals of X only. This is significantly different from other covariance matrices estimation problems, where joint distributional assumptions are typically imposed. Among these results, distributions with polynomial-type tails have been considered by Bickel and Levina (2008b), El Karoui (2008) and Cai and Liu (2011a). In particular, Cai and Liu (2011a) showed that the adaptive thresholding estimator $\hat{\Sigma}^{Ad,Th}$ attains the same rate of convergence $c_{n,p}((\log p)/n)^{1/2}$ as the one for the sub-Gaussian case (14) in probability over the class $\mathcal{U}_q^*(s_{n,p})$, assuming that for some $\gamma, C > 0$, $p \leq Cn^{\gamma}$ and for some $\epsilon > 0$, such that $\mathbb{E}\,|X_j|^{4\gamma+4+\epsilon} \leq K$ for all j. The superiority of the adaptive thresholding estimator $\hat{\Sigma}^{Ad,Th}$ for heavy-tailed distributions is also due to the moderate deviation for the self-normalized statistic $\hat{\sigma}_{ij}/\hat{\theta}_{ij}$ (see Shao (1999)). It is easy to see that this still holds over the class $\mathcal{H}(c_{n,p})$. Recently, Chen et al. (2013) and Basu and Michailidis (2015) considered sparse covariance matrix estimation for time series data based on certain dependence measures, which play an important role in the rates of convergence. In particular, the results in Basu and Michailidis (2015) rely on a new measure of stability for stationary processes without using the commonly imposed functional dependence measure in Wu (2005).

Support recovery

A closely related problem to estimating a sparse covariance matrix is the recovery of the support of the covariance matrix. This problem has been considered by, for example, Rothman et al. (2009) and Cai and Liu (2011a). For support recovery, it is natural to consider the parameter space $\mathcal{H}_0(c_{n,p})$. Define the

support of $\Sigma = (\sigma_{ij})$ by $\text{supp}(\Sigma) = \{(i,j) : \sigma_{ij} \neq 0\}$. Rothman et al. (2009) applied the universal thresholding estimator $\hat{\Sigma}^{U,Th}$ to estimate the support of true sparse covariance in $\mathcal{H}_0(c_{n,p})$, assuming bounded variances $\max_i(\sigma_{ii}) \leq \rho$. In particular, it successfully recovers the support of Σ in probability, provided that the magnitudes of the nonzero entries are above a certain threshold, i.e., $\min_{(i,j)\in\text{supp}(\Sigma)} |\sigma_{ij}| > \gamma_0 (\log p/n)^{1/2}$ for a sufficiently large $\gamma_0 > 0$.

Cai and Liu (2011a) extended this result under a weaker assumption on entries in the support using the adaptive thresholding estimator $\hat{\Sigma}^{Ad,Th}$. We state the result below.

Theorem 11 (Cai and Liu (2011a))**.** *Let $\delta \geq 2$. Suppose the assumptions in Theorem 10 hold and for all $(i,j) \in \text{supp}(\Sigma)$*

$$|\sigma_{ij}| > (2 + \delta + \gamma)(\theta_{ij} \log p/n)^{1/2},$$

for some constant $\gamma > 0$, where $\theta_{ij} = \text{Var}(X_i X_j)$. Then we have

$$\inf_{\mathcal{H}_0(c_{n,p})} \mathbb{P}(\text{supp}(\hat{\Sigma}^{Ad,Th}) = \text{supp}(\Sigma)) \to 1.$$

2.4. Spiked sparse covariance matrices

Minimax upper bound

We now turn to the optimal estimation of sparse spiked covariance matrices over $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ in (7) under the squared spectral norm loss. Recall the spiked covariance structure $\Sigma = V\Lambda V' + I$ is defined in (6). This structure was originally considered in Johnstone (2001) and have been studied by several papers under the sparse principal component analysis setting, including Paul (2007), Nadler (2008), Johnstone and Lu (2009), Amini and Wainwright (2009), Birnbaum et al. (2013), Ma (2013), Vu and Lei (2013), Cai et al. (2013), Berthet and Rigollet (2013) and Cai et al. (2015). However, most of the works considered estimating either individual leading eigenvector v_i or the principal subspace VV' spanned by $r_{n,p}$ leading eigenvectors under the Frobenius norm loss. A brief discussion along this line is presented later in this section. Despite its close relationship with estimation of the covariance matrix Σ via the well-known sin-theta theorem (Davis and Kahan (1970)), the optimality estimation of sparse spiked covariance matrices cannot be obtained immediately, especially under the setting where $r_{n,p}$ and $\lambda_{n,p}$ go to infinity as $n, p \to \infty$.

In a recent work, Cai et al. (2015) considered minimax estimation of sparse spiked covariance matrices under the squared spectral norm. The key step of constructing optimal estimators is a global searching scheme to find the support of V. More specifically, let the support of V be $\text{supp}(V) = \{i : V_{i,*} \neq \mathbf{0}\} \subset \{1, \ldots, p\}$ and the cardinality of any set S be D_S. Recall $\hat{\Sigma} = (\hat{\sigma}_{ij})$ is the sample covariance matrix. Then the estimator of $\text{supp}(V)$ is defined by an arbitrary

element in the following set

$$\mathcal{SV}(c_{n,p}) = \left\{ \begin{array}{l} S \subset \{1,\ldots,p\} : D_S = c_{n,p}, \text{ and for all } A \subset S^c \text{ with } D_A \le c_{n,p}, \\ \left\|\hat{\Sigma}_{A,A} - I\right\| \le \eta(D_A, n, p, \gamma), \left\|\hat{\Sigma}_{A,S}\right\| \le 2\sqrt{\left\|\hat{\Sigma}_{S,S}\right\|\eta(c_{n,p}, n, p, \gamma)} \end{array} \right\},$$

where the thresholding quantity

$$\eta(a, n, p, \gamma) = 2\left(\sqrt{\frac{a}{n}} + \sqrt{\frac{\gamma}{n} a \log(ep/a)}\right) + \left(\sqrt{\frac{a}{n}} + \sqrt{\frac{\gamma}{n} a \log(ep/a)}\right)^2,$$

with $\gamma \ge 10$ under the Gaussian assumption. Intuitively, the two deviation criterions in $\mathcal{SV}(c_{n,p})$ admit $\mathrm{supp}(V)$ as a feasible element and at the same time rule out the possibility that S^c overlaps with $\mathrm{supp}(V)$ for any $S \in \mathcal{SV}(c_{n,p})$. The quantity $\eta(a, n, p, \gamma)$ is carefully picked under the Gaussian assumption based on the Davidson-Szarek bound (Davidson and Szarek (2001)). See Cai et al. (2015) (also Proposition D.1 in Ma (2013)) for further details. Given an estimator $\hat{S} \in \mathcal{SV}(c_{n,p})$ of $\mathrm{supp}(V)$, the final sparse spiked covariance matrix estimator can be defined as

$$\hat{\Sigma}^{CMW} = \left(\hat{\sigma}_{ij}^{CMW}\right), \text{ where } \hat{\sigma}_{ij}^{CMW} = \hat{\sigma}_{ij} I\{i \in \hat{S}, j \in \hat{S}\} + 1I\{i = j \notin \hat{S}\}. \quad (15)$$

We set $\hat{\Sigma}^{CMW} = I$ if $\mathcal{SV}(c_{n,p})$ is the empty set. It was shown that the global search estimator $\hat{\Sigma}^{CMW}$ attains the following rate of convergence under the spectral norm.

Theorem 12 (Cai et al. (2015))**.** *Suppose that X is Gaussian distributed. If $\frac{c_{n,p}}{n} \log \frac{ep}{c_{n,p}} \le c$ for a sufficiently small constant $c > 0$, then*

$$\sup_{\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})} \mathbb{E}\left\|\hat{\Sigma}^{CMW} - \Sigma\right\|^2 \le C\left(\frac{(\lambda_{n,p} + 1)\, c_{n,p}}{n} \log \frac{ep}{c_{n,p}} + \frac{\lambda_{n,p}^2 r_{n,p}}{n}\right).$$

At a high level, the first term in the upper bound above does not involve $r_{n,p}$ and is due to estimation of the leading eigenvector. The second term essentially follows from the estimation error of eigenvalue for a $c_{n,p}$ by $c_{n,p}$ matrix with rank $r_{n,p}$. It is clear that under the setting in which $\lambda_{n,p}$ is bounded by a universal constant, the rate of convergence reduces to $\frac{c_{n,p}}{n} \log \frac{ep}{c_{n,p}}$ since $r_{n,p}$ cannot be larger than $c_{n,p}$. Under such a setting, it is interesting to compare the rates of convergence over $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ and a larger class $\mathcal{H}_0(c_{n,p})$ considered in Section 2.3. Theorems 10 and 12 imply that the rate is reduced from $\frac{c_{n,p}^2 \log p}{n}$ to $\frac{c_{n,p}}{n} \log \frac{ep}{c_{n,p}}$. This much faster rate of convergence can be achieved because of the extra spike structure.

Unlike other estimators considered in the previous sections which are obtained from the sample covariance via a direct "smoothing" step, the estimator $\hat{\Sigma}^{CMW}$ is obtained by a global search for the support of V, which is computationally intensive. It is of interest to ask whether a computationally efficient while still minimax optimal estimator exists. Consider a special case of rank one

$r_{n,p} = 1$. Some recent works imply that without a strong sample size condition, namely $n > Cc_{n,p}^2 \frac{\log p}{\lambda_{n,p}^2}$ for a sufficiently large constant $C > 0$, minimax estimation cannot be achieved by any randomised polynomial-time algorithm. Indeed, Berthet and Rigollet (2013) first showed that there is no computationally efficient algorithm for a sparse principal component detection problem when the strong sample size condition is violated, assuming the widely-believed hardness of Planted Clique detection problem. See Wang et al. (2014) for the extension to the problem of estimating leading eigenvector. Both works established this computational lower bound by reducing the PCA problems into the Planted Clique detection problem but required a larger parameter space which includes many general distributions besides Gaussian distribution. In a recent work, Gao et al. (2014) bridged the gap and established this result for estimating leading eigenvector, faithful to the Gaussian spiked covariance model. It is immediate to extend it into the estimation of covariance matrix Σ because of $r_{n,p} = 1$. See Gao et al. (2014) for further details.

When the upper bound in Theorem 12 is larger than $\lambda_{n,p}^2$, it is trivial to use identity matrix as the estimation with an upper bound rate $\lambda_{n,p}^2$. A minimax lower bound is also constructed in Cai et al. (2015) to show that the minimum of $\lambda_{n,p}^2$ and the rate of convergence in Theorem 12 is indeed optimal.

Sparse principal component analysis

Principal component analysis (PCA) is one of the most commonly used techniques for dimension reduction and feature extraction with many applications, including image recognition, data compression, and clustering. PCA is closely related to, but different from, covariance matrix estimation under the matrix norm losses. In the classical fixed p setting, the leading eigenvector or eigenspace of Σ can be consistently estimated by the counterparts operated on the sample covariance matrix $\hat{\Sigma}_n$ as $n \to \infty$. However this standard PCA approach yields inconsistent estimators in the high-dimensional settings when the spectra of the population covariance matrix remain bounded. See, for example, Paul (2007), d'Aspremont et al. (2007) and Johnstone and Lu (2009). It is worthwhile to point out that this is different from the high-dimensional factor model considered in Fan et al. (2013), where leading eigenvalues increase at least in order of p and as a result the standard PCA still performs well.

Various regularized approaches have been introduced in the literature for PCA, assuming certain sparsity structures on the leading eigenvectors. Zou et al. (2006) imposed Lasso type sparsity constraints on the eigenvector after transforming the PCA problem into a regression problem. d'Aspremont et al. (2007) proposed a semi-definite program as a relaxation to the ℓ_0 penalty. Shen and Huang (2008) applied a regularized low-rank approach with its consistency established in Shen et al. (2013). See also Jolliffe et al. (2003) and Witten et al. (2009) for other methodologies.

Theoretical analysis for PCA has so far mainly focused on the spiked covariance matrix model $\Sigma = \sum_{i=1}^r \lambda_r v_i v_i' + I$, defined in (6). It is commonly assumed

that the λ_i's are bounded from below and above by some universal constants and each eigenvector v_i is either in a weak ℓ_q ball with radius $c_{n,p}^{1/q}$ for $0 < q < 2$ or exactly sparse with $c_{n,p}$ nonzero entries for $q = 0$.

Johnstone and Lu (2009) considered the single spike case $r = 1$ and proposed a diagonal thresholding (DT) procedure. In particular, a thresholding procedure is applied to each $\hat{\sigma}_{ii}$ to pick out those coordinates of strong signals with magnitude at least at the level of $((\log p)/n)^{1/4}$ and then the standard PCA is performed on this subset of coordinates. The final estimator of the leading eigenvector is obtained by padding zeros to the remaining coordinates. Consistency is shown in the paper for estimating the leading eigenvector v_1 under the squared ℓ_2 loss. Amini and Wainwright (2009) studied the theoretical properties of the leading eigenvectors obtained in Johnstone and Lu (2009) and d'Aspremont et al. (2007) with a focus on the model selection setting, in which the leading eigenvector v_1 is exactly sparse, i.e. $q = 0$.

Birnbaum et al. (2013) established the minimax rates of convergence $c_{n,p}(\log p/n)^{1-q/2}$ of the individual leading eigenvectors for finite r with distinct leading eigenvalues $\lambda_i \neq \lambda_j$ for $i \neq j$. The DT method is shown to be sub-optimal but can be used as the first step of a two-stage optimal coordinate selection approach called ASPCA. The purpose of the second stage is to further pick out those coordinates with magnitude larger than the optimal threshold level $((\log p)/n)^{1/2}$. Ma (2013) proposed an iterative thresholding procedure (ITSPCA) based on DT, which also attains the same optimal rate for estimating each leading eigenvector. Estimation of the principal subspace spanned by the leading r eigenvectors is more appropriate when some of the leading eigenvalues have multiplicity great than one. Ma (2013) further established the rates of convergence of ITSPCA for estimating leading principal subspace under a loss function defined by the squared Frobenius distance between the projection matrices of the leading principal subspace $V = (v_1, v_2, \ldots, v_r)$ and its estimator $\hat{V}$, i.e., $\|VV' - \hat{V}\hat{V}'\|_F^2$.

All results discussed above assumed finite rank r. Cai, Ma, and Wu (2013) further considered the optimality problems of estimating the leading principal subspace under a group sparsity assumption, allowing the number of spikes r to diverge to infinity. Specifically, it is assumed that the vector obtained from the ℓ_2 norm of each row of V is in a weak ℓ_q ball with radius $c_{n,p}^{1/q}$ for $0 < q < 2$ or exactly sparse with $c_{n,p}$ nonzero entries for $q = 0$. The minimax rate of convergence is established and an aggregation procedure is constructed and shown to attain the optimal rate $c_{n,p}((r + \log \frac{ep}{c_{n,p}})/(nh(\lambda)))^{1-q/2}$, where $h(\lambda) = \lambda^2/(1+\lambda)$ and leading eigenvalues $\lambda_i \asymp \lambda$ for all i. In particular, the rate is optimal with respect to all the parameters $n, p, r, \lambda, c_{n,p}$. However, this aggregation procedure is computationally infeasible and an optimal adaptive procedure is then proposed in Cai, Ma, and Wu (2013). Vu and Lei (2013) extended the spiked covariance model (6) into a more general setting and considered the problem of optimally estimating the leading eigenspace under group sparsity or column sparsity in a slightly difference setting. Compared to the optimal rates in Cai, Ma, and Wu (2013), the dependency on λ is not optimal for the method proposed in Vu and Lei (2013). In a related paper, Cai et al. (2015) further studied the mini-

max estimation the principal subspace V under the squared spectral norm loss $\|VV' - \hat{V}\hat{V}'\|^2$ based on the estimator $\hat{\Sigma}^{CMW}$ in (15) proposed for the minimax estimation of spiked covariance matrix we discussed at the beginning of this section.

Factor model

Factor models in the high-dimensional setting have been used in a range of applications in finance and economics such as modeling wage rates and optimal portfolio allocations. See, for instance, Engle and Watson (1981) and Goldfarb and Iyengar (2003). Despite closely related to the spiked covariance models where the covariance also can be written as the sum of a low-rank and a sparse matrix, estimation of factor models is different from that of sparse spiked covariance matrices. We consider the following multi-factor model (See Ross (1976, 1977) and Chamberlain and Rothschild (1983).) $X_j = \mathbf{b}_j' F + U_j$, where $\mathbf{b}_j$ is a deterministic k by 1 vector of factor loadings, F is the random vector of common factors and U_j is the error component of X_j. Set $U = (U_1, \ldots, U_p)'$ be the error vector of X and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$. Assume that U and factors F are independent, then it is easy to see the covariance of X can be written as $\Sigma = \mathbf{B}\mathrm{Cov}(F)\mathbf{B}' + \Sigma_U$, where $\Sigma_U = \mathrm{Cov}(U) = \left(\sigma_{ij}^U\right)$ is the error covariance matrix. Usually a small value of k is assumed and a sparse structure, such as the diagonal structure, is imposed on Σ_U. Therefore in factor models, the covariance matrix Σ can be represented as the sum of a low-rank matrix and a sparse matrix.

Fan et al. (2008) considered a factor model assuming that the error components are independent, which results Σ_U to be a diagonal matrix. This result was extended and improved in Fan et al. (2011) by further assuming general sparse structure on Σ_U. More specifically, there are no more than $c_{n,p}$ nonzero entries in each row/column of Σ_U, i.e. $\Sigma_U \in \mathcal{H}_0(c_{n,p})$. Let the ith observation $X^{(i)} = \mathbf{B}F^{(i)} + U^{(i)}$ for $i = 1, \ldots, n$, where $U^{(i)} = \left(U_1^{(i)}, \ldots, U_p^{(i)}\right)'$ is its error vector. Both works assume that the factors $F^{(i)}$, $i = 1, \ldots, n$ are observable and hence the number of factors k is known as well. In other words, the models are more about the regression models. Indeed, this allows using ordinary least squares estimator $\hat{\mathbf{b}}_j$ to estimate loadings $\mathbf{b}_j$ accurately first and then estimating the errors $\hat{U}_j^{(i)} = X_j^{(i)} - \hat{\mathbf{b}}_j' F^{(i)}$ by the residuals. Additional adaptive thresholding procedure is then applied to the error covariance matrix estimator $\hat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^n \hat{U}^{(i)}(\hat{U}^{(i)})'$ to estimate Σ_U, motivated by the procedure in (13). Under certain conditions, including bounded variances $\max_i \mathrm{Var}(X_i)$, bounded $\lambda_{\min}(\Sigma_U)$, $\lambda_{\min}(\mathrm{Cov}(F))$ and exponential tails of F and U, the rates of convergence $c_{n,p}k\left(\log p/n\right)^{1/2}$ under the spectral norm are obtained for estimating Σ_U^{-1} and Σ^{-1}. Note the number of factors k plays a role on the rates, compared to the one in (14).

Recently, Fan et al. (2013) considered the setting in which the factors are unobservable and must be estimated from the data as well. This imposes new challenges since there are two matrices to estimate while only a noisy version

of their sum is observed. To overcome this difficulty, Fan et al. (2013) assumed the factors are pervasive in the sense that a non–negligible fraction of factor loadings is bounded away from zero by a universal constant. As a result, the k eigenvalues of the low-rank matrix $\mathbf{B}\mathrm{Cov}(F)\mathbf{B}'$ diverge at the rate $O(p)$ while the spectra of the sparse matrix Σ_U is assumed to be bounded from below and above. Under this assumption, after simply running the SVD on the sample covariance matrix $\hat{\Sigma}_n$, the matrix $\mathbf{B}\mathrm{Cov}(F)\mathbf{B}'$ can be accurately estimated by the matrix formed by the first k principal components of $\hat{\Sigma}_n$ and the sparse matrix Σ_U can then be estimated by adaptively thresholding the remaining principal components. k is assumed to be finite rather than diverging in Fan et al. (2013). Under other similar assumptions as those in Fan et al. (2011), the rates of convergence $c_{n,p}(\log p/n)^{1/2}$ for estimating Σ_U^{-1} and Σ^{-1} under the spectral norm are derived, assuming $\Sigma_U \in \mathcal{H}_0(c_{n,p})$.

We would like to point out that there also is a growing literature on the study of decomposition from the sum of a low-rank matrix and a sparse matrix. However, the focus is mainly on the data matrix instead of the covariance structure with the goal of identification. The methods are mainly based on certain "incoherence condition" between the two matrices to ensure identifiability while the spectra of the two matrices are on the same order. Let $L = U\Lambda V'$ be the SVD of the low-rank matrix L with rank r. For example, the incoherence condition with parameter μ defined in Candès et al. (2011) states that $\max_i \|U_{i,*}\|^2 \le \mu r/n$, $\max_i \|V_{i,*}\|^2 \le \mu r/n$ and $\|UV'\|_\infty \le n^{-1}\sqrt{\mu r}$. Hence under this incoherence condition, the low-rank matrix cannot be a sparse matrix. It is worthwhile to point out that such an incoherence condition is not required for the factor models discussed in this section and is not satisfied by the sparse spiked covariance matrices in Section 2.4, where the low-rank matrix is also sparse. See, e.g., Candès et al. (2011) and Agarwal et al. (2012).

3. Minimax upper bounds of estimating sparse precision structure

We turn in this section to optimal estimation of sparse precision matrices and recovering its support which have close connections to Gaussian graphical models. The problem has drawn considerable recent attentions. We have seen in the last section that optimal estimators of structured covariance matrices, with the exception of sparse spiked covariance matrices, are usually obtained from the sample covariance matrix through certain direct "smoothing" operations such as banding, tapering, or thresholding. Compared to those methods, estimation of the structured precision matrices is more involved due to the lack of a natural pivotal estimator and is usually obtained through some regression or optimization procedures.

There are two major approaches to estimation of sparse precision matrices: neighborhood-based and penalized likelihood approaches. Neighborhood-based approach runs a Lasso regression or Dantzig selector of each variable on all other variables to estimate the precision matrix column by column. This approach requires running p Lasso regressions. We focus on it in Section 3.1 with

an emphasis on the adaptive rate optimal procedure proposed in Sun and Zhang (2012). An extension of this approach to regressing two variables against others lead to a statistical inference result on each entry ω_{ij}. In Section 3.2, we introduce such a method proposed in Ren et al. (2015) and consider support recovery as well. Penalized likelihood approach is surveyed in Section 3.3, together with latent graphical model structure estimation problems.

3.1. Sparse precision matrix: Adaptive minimax upper bound under spectral norm

Under the Gaussian assumption, the motivation of neighborhood-based approach is the following conditional distribution of X_j given all other variables X_{j^c},

$$X_j | X_{j^c} \sim N\left(-\omega_{jj}^{-1}\omega_{j^c j} X_{j^c}, \omega_{jj}^{-1}\right), \tag{16}$$

where $\omega_{j^c j}$ is the jth column of Ω with the jth coordinate removed. See, for example, Anderson (2003).

Meinshausen and Bühlmann (2006) first proposed the neighborhood selection approach and applied the standard Lasso regression on $\mathbf{X}_j$ against $\mathbf{X}_{j^c}$ to estimate nonzero entries in each row. The goal of this paper however is to identify the support of Ω. In the same spirit, Yuan (2010) applied the Dantzig selector version of this regression to estimate Ω column by column. i.e. $\min \|\beta\|_1$ s.t. $\left\|\left(\mathbf{X}'_{j^c}\mathbf{X}_j - \mathbf{X}'_{j^c}\mathbf{X}_{j^c}\beta\right)/n\right\|_\infty \le \tau$. Cai et al. (2011) further proposed an estimator called CLIME by solving a related optimization problem

$$\arg\min\left\{\|\Omega\|_1 : \ \|\hat{\Sigma}_n\Omega - I\|_\infty \le \tau\right\}.$$

In practice, the tuning parameter τ is chosen via cross-validation. However the theoretical choice of $\tau = CM_{n,p}\sqrt{\log p/n}$ requires the knowledge of the matrix ℓ_1 norm $M_{n,p} = ||\Omega||_{\ell_1}$, which is unknown. Cai et al. (2012) introduced an adaptive version of CLIME $\hat{\Omega}_{ACLIME}$, which is data-driven and adaptive to the variability of individual entries of $\hat{\Sigma}_n\Omega - I$. Over the class $\mathcal{HP}(c_{n,p}, M)$, the estimators proposed in Yuan (2010), Cai et al. (2011) and Cai et al. (2012) can be shown to attain the optimal rate under the spectral norm if the matrix ℓ_1 norm $M_{n,p}$ is bounded. Besides the Gaussian case, both Cai et al. (2011) and Cai et al. (2012) also considered sub-Gaussian and polynomial tail distributions. It turns out that if each X_i has finite $4+\varepsilon$ moments, under some mild condition on the relationship between p and n, the rate of convergence is the same as those in the Gaussian case.

Now we introduce the minimax upper bound for estimating Ω under the spectral norm over the class $\mathcal{HP}(c_{n,p}, M)$. Sun and Zhang (2012) constructed an estimator for each column with the scaled Lasso, a joint estimator for the regression coefficients and noise level. For simplicity, we assume X is Gaussian. For each $j = 1, \ldots, p$, the scaled Lasso is applied to the linear regression of the

jth column $\mathbf{X}_j$ of the data matrix against all other columns $\mathbf{X}_{j^c}$ as follows:

$$\left\{\hat{\beta}_j, \hat{\eta}_j\right\} = \arg\min_{b,\eta} \left\{ \frac{\|\mathbf{X}_j - \mathbf{X}_{j^c} b\|^2}{2n\eta} + \frac{\eta}{2} + \lambda \sum_{k \neq j} \sqrt{\hat{\sigma}_{kk}}\, |b_k| \right\},$$
$$b \in \mathbb{R}^{p-1} \text{ and indexed by } j^c, \tag{17}$$

where $\lambda = A\sqrt{(\log p)/n}$ for some constant $A > 2$ and $\hat{\sigma}_{kk}$ is the sample variance of X_k. The optimization (17) is jointly convex in (b, η), hence an iterative algorithm, which in each iteration first estimates b given η, then estimates η given the b just estimated, is guaranteed to converge to the global solution. This iterative algorithm is run until convergence of $\{\hat{\beta}_j, \hat{\eta}_j\}$ for each scaled Lasso regression (17). After computing the solution $\{\hat{\beta}_j, \hat{\eta}_j\}$, the estimate of jth column $\hat{\omega}_j^{SL}$ of Ω is given by $\hat{\omega}_{jj}^{SL} = \hat{\eta}_j^{-2}$ and $\hat{\omega}_{j^c j}^{SL} = -\hat{\beta}_j \hat{\eta}_j^{-2}$. The final estimator $\hat{\Omega}^{SL}$ is obtained by putting the columns $\hat{\omega}_j^{SL}$ together and applying an additional symmetrization step, i.e.,

$$\tilde{\Omega}^{SL} = \arg\min_{M: M' = M} ||\hat{\Omega}^{SL} - M||_{\ell_1},\ \hat{\Omega}^{SL} = (\hat{\omega}_{ij}^{SL}).$$

Without assuming bounded matrix ℓ_1 norm on Ω, the rate of convergence of $\tilde{\Omega}^{SL}$ is given under the spectral norm as follows

Theorem 13 (Sun and Zhang (2012)). *Suppose that X is Gaussian and $c_{n,p}(\frac{\log p}{n})^{1/2} = o(1)$. The parameter class $\mathcal{HP}(c_{n,p}, M)$ is defined in (8). Then the estimator $\tilde{\Omega}^{SL}$ with $\lambda = A\sqrt{(\log p)/n}$ for some constant $A > 2$ satisfies*

$$\sup_{\mathcal{HP}(c_{n,p}, M)} \mathbb{E} \left\| \tilde{\Omega}^{SL} - \Omega \right\| \leq C c_{n,p} \left(\frac{\log p}{n} \right)^{1/2}. \tag{18}$$

The key technical tool in the analysis is the oracle inequality for the prediction error as well as the bound on the error under the ℓ_1 norm in the high-dimensional sparse linear regression setting for the Lasso estimator and Dantzig selector. We state the results for Lasso as a lemma in the following simple case. Assume the observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ have the following form

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{W},$$

where $\mathbf{X}$ is an n by p design matrix and $\mathbf{W} = (W_1, W_2, \ldots, W_n)'$ is the vector of i.i.d. independent sub-Gaussian noise with variance σ^2. Suppose the coefficient β is sparse with no more than s nonzero coordinates, i.e. $\|\beta\|_0 \leq s$. The Lasso estimator of β is defined by

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - \mathbf{X}b\|^2}{2n} + \lambda \|b\|_1 \right\}. \tag{19}$$

Moreover, we assume that the rows of $\mathbf{X}$ are i.i.d. copies of some sub-Gaussian distribution X with mean 0 and covariance matrix $\mathrm{Cov}(X)$ whose spectra are bounded from below and above by constants.

Lemma 3. *Assume that* $(s\log p)/n = o(1)$. *For any given* $M > 0$, *there exists a sufficiently large constant* $A > 0$ *depending on* M *and the spectra of* $\mathrm{Cov}(X)$ *such that the following results hold with probability* $1 - O(p^{-M})$ *for the Lasso estimator with the tuning parameter* $\lambda > A\sigma\sqrt{(\log p)/n}$ *in (19),*

$$\left\|\hat{\beta}_L - \beta\right\|_1 \leq Cs\sigma\sqrt{(\log p)/n},$$
$$\left\|\mathbf{X}\left(\hat{\beta}_L - \beta\right)\right\|^2 \leq Cs\sigma^2(\log p)/n,$$

where the constant C *depends on* M *and the spectra of* $\mathrm{Cov}(X)$.

Lemma 3 follows from the standard Lasso regression results. See, for example, Bickel et al. (2009) for further details. Note that under the sub-Gaussian assumption on the random design matrix $\mathbf{X}$ and the assumption $(s\log p)/n = o(1)$, the required properties on the Gram matrix $\mathbf{X}'\mathbf{X}/n$ such as the compatibility factor condition (van de Geer and Bühlmann (2009)), the restricted eigenvalue condition (Bickel et al. (2009)) or the cone invertibility factors condition (Ye and Zhang (2010)) are automatically satisfied with probability $1 - c_1\exp(-c_2 n)$, where constants c_1 and c_2 depend on the spectra of $\mathrm{Cov}(X)$ and σ. See, for example, Rudelson and Zhou (2013) for details.

Another contribution of $\tilde{\Omega}^{SL}$ is that the procedure is tuning-free in the sense that λ is well specified. Finally, the assumptions in Theorem 13 can be further weakened and a smaller λ is also valid. See Sun and Zhang (2012) for further details.

3.2. Individual entries of sparse precision matrix: Asymptotic normality

Given the connection between the entry ω_{ij} and the corresponding edge (i, j) in a Gaussian graph, it is of significant interest to make inference on and provide a confidence interval for ω_{ij}. Furthermore, the analysis would lead to results on support recovery. Along this line, to estimate a given entry ω_{ij}, Ren et al. (2015) extended the neighborhood-based approach to regress two variables against the remaining ones, based on the following conditional distribution,

$$X_A|X_{A^c} \sim N\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}X_{A^c}, \Omega_{A,A}^{-1}\right), \quad \text{with} \quad \Theta_{A,A} = \Omega_{A,A}^{-1} = \begin{pmatrix} \theta_{ii} & \theta_{ij} \\ \theta_{ji} & \theta_{jj} \end{pmatrix} \tag{20}$$

where $A = \{i, j\}$ is the index set of the two variables. Recall that ω_{ij} sits in the coefficients $-\omega_{jj}^{-1}\omega_{j^c j}$ of neighborhood selection model in (16), our goal thus was to estimate the coefficients as a whole under some vector norm loss. In comparison, here we only need to estimate the noise level in the regression model (20) since ω_{ij} is one of the three parameters in $\Theta_{A,A}$. This leads to the multivariate regression with two response variables X_i and X_j. Scaled Lasso

regression is applied in Ren et al. (2015) as follows. For each $m \in A = \{i, j\}$,

$$\left\{\hat{\beta}_m, \hat{\theta}_{mm}^{1/2}\right\} = \arg\min_{b\in\mathbb{R}^{p-2},\sigma\in\mathbb{R}} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c} b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k\in A^c} \sqrt{\hat{\sigma}_{kk}}\,|b_k| \right\}, \tag{21}$$

where the vector b is indexed by A^c and $\hat{\sigma}_{kk}$ is the sample variance of X_k. Define the $(p-2)\times 2$ dimensional coefficients $\hat{\beta} = (\hat{\beta}_{\mathbf{i}}, \hat{\beta}_{\mathbf{j}})$ and the residuals of the scaled Lasso regression by $\hat{\epsilon}_A = \mathbf{X}_A - \mathbf{X}_{A^c}\hat{\beta}$. The estimator of $\Theta_{A,A}$ can be given by $\hat{\Theta}_{A,A} = \hat{\epsilon}_A'\hat{\epsilon}_A/n$. Finally, the estimator $\hat{\Omega}_{A,A} = (\hat{\omega}_{kl})_{k,l\in A}$ is obtained by simply inverting $\hat{\Theta}_{A,A}$, i.e. $\hat{\Omega}_{A,A} = \hat{\Theta}_{A,A}^{-1}$. In particular,

$$\hat{\omega}_{ij} = -\hat{\theta}_{ij}/(\hat{\theta}_{ii}\hat{\theta}_{jj} - \hat{\theta}_{ij}^2). \tag{22}$$

The rate of convergence of estimating each ω_{ij} is then provided under a certain sparsity assumption over $\mathcal{HP}(c_{n,p}, M)$. In particular when $c_{n,p} = o\left(\sqrt{n}/\log p\right)$, an asymptotic efficiency result and the corresponding confidence interval are obtained.

Theorem 14 (Ren et al. (2015))**.** *Let* $\lambda = \sqrt{\frac{2\delta\log p}{n}}$ *for any* $\delta > 1$ *in Equation (21). Assume* $(c_{n,p}\log p)/n = o(1)$*, then for any small* $\epsilon > 0$*, there exists a constant* $C_1 = C_1(\epsilon) > 0$ *such that*

$$\sup_{\mathcal{HP}(c_{n,p},M)} \sup_{i,j} \mathbb{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| > C_1 \max\left\{ c_{n,p}\frac{\log p}{n}, \sqrt{\frac{1}{n}} \right\} \right\} \le \epsilon. \tag{23}$$

Furthermore, $\hat{\omega}_{ij}$ *is asymptotically efficient*

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} N\left(0, 1\right), \tag{24}$$

when $c_{n,p} = o(\frac{\sqrt{n}}{\log p})$*, where* $F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2$.

The key technical tool in the analysis is also related to Lemma 3 but focuses on the prediction error rather than estimation under the ℓ_1 norm. The advantage of estimator $\hat{\omega}_{ij}$ over $\hat{\omega}_{ij}^{SL}$ defined in Section 3.1 is that by estimating the noise level rather than each coefficient, the estimation accuracy can be significantly improved. However, we have to pay for the accuracy with the computational costs. If our goal is to estimate all those entries above the threshold level $(\log p/n)^{1/2}$ individually, we can first apply the method proposed in Liu (2013) with p regressions to pick those order $pc_{n,p}$ entries above the threshold level, then order $pc_{n,p}$ regressions have to be done to estimate them individually. In contrast, methods in Section 3.1 only require p regressions. Minimax lower bounds are also provided in Ren et al. (2015) to show the estimator $\hat{\omega}_{ij}$ is indeed rate optimal. See Section 4.3 for details. This methodology can be routinely extended into a more general form with A replaced by some subset $B \subset \{1, 2, \ldots, p\}$ with bounded size. Then the inference result can be obtained to estimate a smooth functional of $\Omega_{B,B}^{-1}$.

Other related works in high-dimensional regression also can be applied to the current setting. Zhang and Zhang (2014) proposed a relaxed projection approach for making inference of each coefficient in a regression setting. See also van de Geer et al. (2014) and Javanmard and Montanari (2014). Although the procedures in those works seem different from $\hat{\omega}_{ij}$ in (22), essentially all methods try to estimate the partial correlation of X_i and X_j and hence they are asymptotically equivalent. Liu (2013) recently developed a multiple testing procedure with the false discovery rate (FDR) control for testing the entries of Ω, $H_{0ij}\colon \omega_{ij} = 0$. Surprisingly, to test all ω_{ij}, it only requires running p regressions as (17).

The support recovery problem is closely related due to the graphical interpretation as well. Based on the estimator $\hat{\omega}_{ij}$ in (22), Ren et al. (2015) applied an additional thresholding procedure, adaptive to the Fisher information F_{ij} to recover the sign of Ω. More specifically, define $\mathcal{S}(\mathbf{\Omega}) = \{\text{sgn}(\omega_{ij}),\ 1 \le i, j \le p\}$ and

$$\begin{aligned} \hat{\Omega}^{ANT} &= (\hat{\omega}_{ij}^{ANT})_{p\times p}, \quad \text{where } \hat{\omega}_{ii}^{ANT} = \hat{\omega}_{ii}, \text{ and } \hat{\omega}_{ij}^{ANT} = \hat{\omega}_{ij} I\{|\hat{\omega}_{ij}| \ge \tau_{ij}\} \\ \text{with } \tau_{ij} &= \sqrt{(2\xi_0(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)\log p)/n} \text{ for } i \ne j. \end{aligned} \tag{25}$$

Theorem 15 (Ren et al. (2015))**.** *Let* $\lambda = \sqrt{\frac{2\delta \log p}{n}}$ *for any* $\delta > 3$ *and* $\xi_0 > 2$ *in the thresholding level (25). Assume* $c_{n,p} = o(\sqrt{n/\log p})$ *and* $|\omega_{ij}| \ge \sqrt{(8\xi_0(\omega_{ii}\omega_{jj} + \omega_{ij}^2)\log p)/n}$ *for any* $\omega_{ij} \ne 0$*. Then we have*

$$\inf_{\mathcal{HP}_0(c_{n,p},M)} \mathbb{P}\left(\mathcal{S}(\hat{\Omega}^{ANT}) = \mathcal{S}(\Omega)\right) \to 1. \tag{26}$$

The sufficient condition on each nonzero entry in the Theorem 15 is much weaker compared with other results in the literature, where the smallest magnitude of the nonzero entries is required to be above the threshold level $\|\Omega\|_{\ell_1}\sqrt{(\log p)/n}$. It is worthwhile to point out that based on Theorem 14, it can be easily shown that $\hat{\Omega}^{ANT}$ also attains the optimal rates of convergence under the spectral norm over $\mathcal{HP}(c_{n,p}, M)$ as that in Theorem 13.

3.3. Related results

Penalized likelihood approaches Penalized likelihood methods have also been introduced for estimating sparse precision matrices. It is easy to see that under the Gaussian assumption the negative log-likelihood up to a constant, can be written as $l\left(X^{(1)}, \ldots, X^{(n)}; \Omega\right) = \text{tr}(\hat{\Sigma}_n\Omega) - \log\det(\Omega)$, where $\det(\Omega)$ is the determinant of Ω. To incorporate the sparsity of Ω, we consider the following penalized log-likelihood estimator with Lasso-type penalty

$$\hat{\Omega}_\lambda = \arg\min_{\Omega\succ 0} \text{tr}(\hat{\Sigma}_n\Omega) - \log|\Omega| + \lambda\sum_{i,j}|\omega_{ij}|, \tag{27}$$

where $\Omega \succ 0$ means symmetric positive definite. Some results are derived by using ℓ_1 penalty on the off-diagonal entries $\sum_{i\neq j}|\omega_{ij}|$ rather than all entries. We will review some theoretical properties and computational issue respectively below.

Yuan and Lin (2007) first proposed using $\hat{\Omega}_\lambda$ and studied its asymptotic properties for fixed p as $n \to \infty$. Rothman et al. (2008) analyzed the high-dimensional behavior of this estimator. Assuming that spectra of Ω are bounded from below and above, the rates of convergence $((p+s)\log p/n)^{1/2}$ and $((1+s)\log p/n)^{1/2}$ under the Frobenius norm and spectral norm are obtained respectively with $s = \sum_{i\neq j} I\{\omega_{ij} \neq 0\}$ being the number of nonzero off-diagonal entries. The sparsity s is defined globally, which is different from the local sparsity $c_{n,p}$ defined in (9) and can be as large as $pc_{n,p}$ over the class $\mathcal{HP}_0(c_{n,p}, M)$. Hence in the worst case scenario, the rate of convergence $((1+s)\log p/n)^{1/2}$ under the spectral norm is not as good as that $(c_{n,p}\log p/n)^{1/2}$ derived by neighborhood-based approach in (18). Lam and Fan (2009) studied a generalization of (27) and replace the Lasso penalty by general non-convex penalties such as SCAD to overcome the bias issue. Ravikumar et al. (2011) applied the primal-dual witness construction to derive the rate of convergence $(\log p/n)^{1/2}$ under the sup-norm which in turn leads to convergence rates in the Frobenius and spectral norms as well as support recovery under certain regularity conditions. The results heavily depend on a strong irrepresentability condition imposed on the Hessian matrix $\Gamma = \Sigma \otimes \Sigma$, where $\otimes$ is the tensor (or Kronecker) product. Both sub-Gaussian and polynomial tail cases are considered. However this method cannot be extended to allowing many small nonzero entries such as the class $\mathcal{HP}(c_{n,p}, M)$.

Latent Gaussian graphical model We have seen the connections between the precision matrix and the corresponding Gaussian graph, in which we assume all variables are fully observed. In some applications, one may not have access to all the relevant variables. Suppose we only observe p coordinates $X = (X_1, \ldots, X_p)'$ of a $(p+r)$-dimensional Gaussian vector $(X', Y')'$, where $Y = (Y_1, \ldots, Y_r)'$ represent the latent coordinates. Denote the covariance matrix of all variables by $\Sigma_{(X,Y)}$. It is natural to assume that the fully observed $(p+r)$-dimensional Gaussian graphical model has a sparse dependence graph. In other words, the precision matrix $\Omega_{(X,Y)} = \Sigma^{-1}_{(X,Y)}$ is sparse. Represent the precision matrix $\Omega_{(X,Y)}$ in the following block form

$$\Omega_{(X,Y)} = \begin{pmatrix} \Omega_{XX} & \Omega_{XY} \\ \Omega_{YX} & \Omega_{YY} \end{pmatrix}.$$

In such a case, the $p \times p$ precision matrix Ω of the observed coordinates X can be written as the difference by the Schur complement formula,

$$\Omega = \Omega_{XX} - \Omega_{XY}\Omega_{YY}^{-1}\Omega_{YX} = S^* - L^*.$$

Here $S^* = \Omega_{XX}$ is a sparse matrix corresponding to the structure of the subgraph induced by those observed p variables and $L^* = \Omega_{XY}\Omega_{YY}^{-1}\Omega_{YX}$ is a low-

rank matrix with rank at most r, which is the number of the unobserved latent variables.

Chandrasekaran et al. (2012) proposed a penalized likelihood approach to estimate both the sparse structure S^* and the low-rank part L^* as the solutions to

$$\min_{(S,L):S-L\succ 0,\ L\succeq 0} \operatorname{tr}\left((S-L)\,\hat{\Sigma}_n\right) - \log\det(S-L) + \chi_n\left(\gamma\sum_{i,j}|s_{ij}| + \operatorname{tr}(L)\right). \tag{28}$$

Here $\hat{\Sigma}_n$ is the sample covariance matrix of the observed coordinates X and $\operatorname{tr}(L)$ is the trace of L, which is used to induce the low-rank structure on L. Consistency results were established under a strong irrepresentability condition and assumptions on the minimum magnitude of the nonzero entries of S^* and the minimum nonzero eigenvalue of L^*. Ren and Zhou (2012) relaxed the assumptions by considering the parameter space $\mathcal{HP}_0(c_{n,p}, M)$ (9) for S^* with bounded matrix ℓ_1 norm and a "spread-out" parameter space for L^*. Ren et al. (2015) further removed the matrix ℓ_1 norm condition based on the estimator $\hat{\Omega}^{ANT}$ obtained in (25).

3.4. Computational issues

Estimating high-dimensional precision matrices is a computationally challenging problem. Pang et al. (2014) proposed an efficient parametric simplex algorithm to implement the CLIME estimator. In particular, their algorithm efficiently calculates the full piecewise-linear regularization path and provides an accurate dual certificate as stopping criterion. An R package 'fastclime' coded in C has been developed. See Pang et al. (2014) for more details.

For the semi-definite program (27), Yuan and Lin (2007) solved the problem using interior-point method for the general max-det problem which is proposed by Vanderberghe et al. (1998). Rothman et al. (2008) derived their algorithm based on the Cholesky decomposition and the local quadratic approximation such that cyclical coordinate descent approach can be applied. Define $W = \hat{\Omega}_\lambda^{-1}$. Banerjee et al. (2008) showed that one can solve the dual of (27) through optimizing over each row/column of W in a block coordinate descent form (see also d'Aspremont et al. (2008)). In fact, solving this dual form is equivalent to solving p coupled Lasso regression problems, which are related to the neighborhood-based approach considered in Section 3.1. Friedman et al. (2008) further proposed the GLasso algorithm which takes advantage of the fast coordinate descent algorithms (Friedman et al. (2007)) to solve it efficiently. The computational complexity of GLasso is $O(p^3)$. In comparison, the algorithms of Banerjee et al. (2008) and Yuan and Lin (2007) have higher computational costs. A modified GLasso algorithm is proposed by Witten et al. (2011) to improve the speed from $O(p^3)$ to $O(p^2 + \sum_{i=1}^m |C_i|^3)$ when the solution $\hat{\Omega}_\lambda$ is block diagonal with blocks $C_1, \ldots, C_m$, where $|C_i|$ is the size of the ith block.

Hsieh et al. (2011) apply a second-order algorithm to solve (27) and achieve superlinear rate of convergence. Their algorithm is based on a modified Newton's

method which leverages the sparse structure of the solution. Recently, Hsieh et al. (2013) further improved this result and claimed that the optimization problem (27) can be solved even for a million variables. A block coordinate descent method with the blocks chosen via a clustering approach is used to avoid the memory bottleneck of storing the gradient W when dimension p is very large.

4. Lower bounds

A major step in establishing a minimax theory is the derivation of rate sharp minimax lower bounds. In this section, we first review a few effective lower bound arguments based on hypothesis testing. These include Le Cam's method, Assouad's Lemma and Fano's Lemma which have been commonly used in the more conventional nonparametric estimation problems. See Yu (1997) and Tsybakov (2009) for further discussions. We will also discuss a new lower bound technique developed in Cai and Zhou (2012b) that is particularly well suited for treating "two-directional" problems such as matrix estimation, where one direction is along the rows and another along the columns. The technique can be viewed as a generalization of both Le Cam's method and Assouad's Lemma. We will then apply these lower bound arguments to the various covariance and precision matrix estimation problems discussed in the previous sections to obtain minimax lower bounds, which match the corresponding upper bound results in the last two sections. The upper and lower bounds together yield the optimal rates of convergence given in Section 1.

4.1. General minimax lower bound techniques

Le Cam's method

Le Cam's method is based on a two-point testing argument. See Le Cam (1973) and Donoho and Liu (1991). In nonparametric estimation problems, Le Cam's method often provides the minimax lower bound for estimating a real-valued functional. See, for instance, Bickel and Ritov (1988) and Fan (1991) for the quadratic functional estimation problems.

Let X be an observation from a distribution $\mathbb{P}_\theta$ where θ belongs to a parameter set Θ. For two distributions $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ with densities p_{θ_1} and p_{θ_2} with respect to a common dominating measure μ, the total variation affinity is given by $\|\mathbb{P}_{\theta_1} \wedge \mathbb{P}_{\theta_2}\| = \int p_{\theta_1} \wedge p_{\theta_2} d\mu$. Le Cam's method relates the testing problem with the total variation affinity. In other words, when the total variation affinity between the two distributions is bounded away from zero, it is impossible to test between those two distributions perfectly. As a consequence, a lower bound can be measured by the distance of the two parameters θ_1 and θ_2, which index those two distributions.

In the current paper, we introduce a version of Le Cam's method which tests the simple hypothesis $H_0 : \theta = \theta_0$ against a composite alternative $H_1 : \theta \in \Theta_1$

with a finite parameter set $\Theta = \{\theta_0, \theta_1, \ldots, \theta_D\}$. Let L be a loss function on Θ_*, the domain of θ. Define

$$\ell_{\min} = \min_{1 \le i \le D} \inf_{t \in \Theta_*} \left[L\left(t, \theta_0\right) + L\left(t, \theta_i\right) \right]$$

and denote $\bar{\mathbb{P}} = \frac{1}{D} \sum_{i=1}^{D} \mathbb{P}_{\theta_i}$. Le Cam's method gives a lower bound for the maximum estimation risk over the parameter set Θ.

Lemma 4 (Le Cam)**.** *Let T be any estimator of θ based on an observation X from a distribution $\mathbb{P}_\theta$ with $\theta \in \Theta = \{\theta_0, \theta_1, \ldots, \theta_D\}$, then*

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L\left(T, \theta\right) \ge \frac{1}{2} l_{\min} \left\| \mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}} \right\|. \tag{29}$$

Assouad's lemma

Assouad's lemma works with a hypercube $\Theta = \{0,1\}^r$. It is based on testing a number of pairs of simple hypotheses and is connected to multiple comparisons. See Assouad (1983). In nonparametric estimation problems, Assouad's lemma is often successful in obtaining the minimax lower bound for many global estimation problems such as estimating the whole density or regression functions in certain smoothness classes.

For a parameter $\theta = (\theta_1, \ldots, \theta_r)$ where $\theta_i \in \{0,1\}$, one tests whether $\theta_i = 0$ or 1 for each $1 \le i \le r$ based on the observation X. In other words, we decompose the global estimation problem into r sub-problems. For each sub-problem or each pair of simple hypotheses, there is a certain loss for making an error in the comparison. The lower bound given by Assouad's lemma is a combination of losses from testing all pairs of simple hypotheses. To make connection to Le Cam's method, we can view the loss for making an error due to each sub-problem is obtained by applying Le Cam's method. In particular, when $r = 1$, Assouad's lemma becomes Le Cam's method with $D = 1$ in Lemma 4, which tests a simple null hypothesis against a simple alternative.

Let

$$H\left(\theta, \tilde{\theta}\right) = \sum_{i=1}^{r} \left| \theta_i - \tilde{\theta}_i \right| \tag{30}$$

be the Hamming distance on Θ. Assouad's lemma gives a lower bound for the maximum risk over the hypercube Θ of estimating an arbitrary quantity $\psi\left(\theta\right)$ belonging to a metric space with metric d. It works especially well when the metric d is decomposable with respect to the Hamming distance.

Lemma 5 (Assouad)**.** *Let $X \sim \mathbb{P}_\theta$ with $\theta \in \Theta = \{0,1\}^r$ and let $T = T(X)$ be an estimator of $\psi(\theta)$ based on X. Then for all $s > 0$*

$$\max_{\theta \in \Theta} 2^s \mathbb{E}_\theta d^s\left(T, \psi\left(\theta\right)\right) \ge \min_{\theta \ne \tilde{\theta}} \frac{d^s\left(\psi\left(\theta\right), \psi\left(\tilde{\theta}\right)\right)}{H\left(\theta, \tilde{\theta}\right)} \cdot \frac{r}{2} \cdot \min_{H\left(\theta, \tilde{\theta}\right) = 1} \left\| \mathbb{P}_\theta \wedge \mathbb{P}_{\tilde{\theta}} \right\|. \tag{31}$$

The lower bound in (31) has three factors. The first factor is the minimum cost of making a mistake per comparison and the second one is the expected number of mistakes one would make when each pair of simple hypotheses is indistinguishable. The last factor, which is usually bounded below by some positive constant in applications, is the total variation affinity for each sub-problem.

Le Cam-Assouad's method

The Le Cam-Assouad's method, which was first introduced in Cai and Zhou (2012b), is designed to treat problems such as estimation of sparse matrices with constraints on both rows and columns. Again, let $X \sim \mathbb{P}_\theta$ where $\theta \in \Theta$. The parameter space Θ of interest has a special structure which can be viewed as the Cartesian product of two components Γ and Λ. For a given positive integer r and a finite set $B \subset \mathbb{R}^p \setminus \{\mathbf{0}_{1\times p}\}$, define $\Gamma = \{0,1\}^r$ and $\Lambda \subseteq B^r$. Define

$$\Theta = \Gamma \otimes \Lambda = \{\theta = (\gamma, \lambda) : \gamma \in \Gamma \text{ and } \lambda \in \Lambda\}. \tag{32}$$

The Le Cam-Assouad's method reduces to the classical Assouad's lemma when Λ contains only one element, and becomes Le Cam's method when $r = 1$. The advantage of this method is that it breaks down the lower bound calculations for the whole matrix estimation problem into calculations for individual rows so that the overall analysis is tractable.

For $\theta = (\gamma, \lambda) \in \Theta$, denote the projection of θ to Γ by $\gamma(\theta) = (\gamma_i(\theta))_{1\le i\le r}$ and to Λ by $\lambda(\theta) = (\lambda_i(\theta))_{1\le i\le r}$. Let D_Λ be the cardinality of Λ. For a given $a \in \{0,1\}$ and $1 \le i \le r$, we define the mixture distribution $\bar{\mathbb{P}}_{a,i}$ by

$$\bar{\mathbb{P}}_{a,i} = \frac{1}{2^{r-1} D_\Lambda} \sum_\theta \{\mathbb{P}_\theta : \ \gamma_i(\theta) = a\}. \tag{33}$$

So $\bar{\mathbb{P}}_{a,i}$ is the mixture distribution over all P_θ with $\gamma_i(\theta)$ fixed to be a while all other components of θ vary over all possible values. In the construction of the parameter set for establishing the minimax lower bound of matrix estimation problems, each $\theta = (\gamma, \lambda)$ is uniquely associated with some symmetric matrix. Usually r is the number of possibly non-zero rows in the upper triangle of the matrix, and each element λ of Λ is associated with a matrix by making r non-zero rows of the matrix equal to the r coordinates of λ in order.

Lemma 6 (Le Cam-Assouad)**.** *For any estimator T of $\psi(\theta)$ based on an observation from a probability distribution in $\{\mathbb{P}_\theta, \theta \in \Theta\}$, and any $s > 0$*

$$\max_\Theta 2^s \mathbb{E}_\theta d^s(T, \psi(\theta)) \ge \alpha \frac{r}{2} \min_{1\le i\le r} \left\| \bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i} \right\| \tag{34}$$

where $\bar{\mathbb{P}}_{a,i}$ is defined in Equation (33) and α is given by

$$\alpha = \min_{\{(\theta,\tilde\theta): H(\gamma(\theta),\gamma(\tilde\theta))\ge 1\}} \frac{d^s(\psi(\theta), \psi(\tilde\theta))}{H(\gamma(\theta), \gamma(\tilde\theta))}. \tag{35}$$

Fano's lemma

Fano's lemma, like Assouad's lemma, is also based on multiple hypotheses testing argument and has been widely used for global estimation problems in nonparametric settings. Fano's lemma applies to a more general setting and hence is stronger than Assouad's lemma, although the latter one seems easier to use in many applications. For their relationship, we refer to Yu (1997) for more details.

For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density p and q with respect to a common dominating measure μ, write the Kullback-Leibler divergence as $K(\mathbb{P},\mathbb{Q}) = \int p \log \frac{p}{q} d\mu$. The following lemma, which can be viewed as a version of Fano's lemma, gives a lower bound for the minimax risk over the parameter set $\Theta = \{\theta_0, \theta_1, \ldots, \theta_{m_*}\}$. See (Tsybakov, 2009, Section 2.6) for more detailed discussions.

Lemma 7 (Fano). *Let $\Theta = \{\theta_i : i = 0, \ldots, m_*\}$ be a parameter set and d be a distance over Θ. Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability distributions satisfying*

$$\frac{1}{m_*} \sum_{1 \le i \le m_*} K\left(\mathbb{P}_{\theta_i}, \mathbb{P}_{\theta_0}\right) \le c \log m_* \tag{36}$$

with $0 < c < 1/8$. Let $\hat{\theta}$ be any estimator based on an observation from a distribution in $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Then

$$\sup_{\theta \in \Theta} \mathbb{E} d^2\left(\hat{\theta}, \theta\right) \ge \min_{i \ne j} \frac{d^2\left(\theta_i, \theta_j\right)}{4} \frac{\sqrt{m_*}}{1+\sqrt{m_*}} \left(1 - 2c - \sqrt{\frac{2c}{\log m_*}}\right).$$

Another advantage of Fano's lemma is that it sometimes only relies on the analytical behavior of the packing number of the parameter space with respect to the loss function d and avoids constructing an explicit subsets of parameter spaces, which can be a challenging task in many situations. See Yang and Barron (1999) for details. Hence the accurate rate of packing number at the logarithm level is the key to applying Fano's lemma and obtaining the minimax lower bounds. Rigollet and Tsybakov (2012) applied this method to improve the assumptions in the minimax lower bound argument for estimating sparse covariance matrices under the matrix ℓ_1 norm in Cai and Zhou (2012a).

4.2. Application of Assouad's lemma to estimating bandable covariance matrices

In Section 2.1, we constructed the tapering estimator $\hat{\Sigma}_{T,k_T}$ and claimed in Theorem 7 that it attains the rate of convergence $\min\left\{\left(\frac{\log p}{n} + n^{-\frac{2\alpha}{2\alpha+1}}\right), \frac{p}{n}\right\}$ over the bandable class $\mathcal{F}_\alpha\left(M_0, M\right)$ defined in (1) under the spectral norm. We now apply Assouad's lemma to give a lower bound $n^{-\frac{2\alpha}{2\alpha+1}}$.

The basic strategy is to carefully construct a finite least favorable subset of the corresponding parameter space in the sense that the difficulty of estimation over the subset is essentially the same as that of estimation over the whole

parameter space. The finite collection that is appropriate for this lower bound argument is defined as follows. For given positive integers k and m with $k \leq p/2$ and $1 \leq m \leq k$, define the $p \times p$ matrix $D(m,k) = (d_{ij})_{p\times p}$ with

$$d_{ij} = I\left\{i = m \text{ and } m+1 \leq j \leq 2k, \text{ or } j = m \text{ and } m+1 \leq i \leq 2k\right\}.$$

Set $k = n^{\frac{1}{2\alpha+1}}$ and $a = k^{-(\alpha+1)}$. We then define the collection of 2^k covariance matrices as

$$\mathcal{F}_{\text{sub}} = \left\{\Sigma(\theta) : \Sigma(\theta) = I + \tau a \sum_{m=1}^{k} \theta_m D(m,k), \quad \theta = (\theta_m) \in \{0,1\}^k\right\} \tag{37}$$

where I is the $p \times p$ identity matrix and $0 < \tau < 2^{-\alpha-1}M$. Without loss of generality we assume that $M_0 > 1$.

Consider the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{\text{sub}}$ and the joint distribution P_θ. For $0 < \tau < 2^{-\alpha-1}M$ it is easy to check that $\mathcal{F}_{\text{sub}} \subset \mathcal{F}_\alpha(M_0, M)$ for sufficiently large n. Hence applying Lemma 5 to the parameter space $\mathcal{F}_{\text{sub}}$, we have

$$\begin{aligned}\inf_{\hat{\Sigma}} \max_{\Sigma\in\mathcal{F}_\alpha} 2^2 E_\Sigma \left\|\hat{\Sigma} - \Sigma\right\|^2 &\geq \inf_{\hat{\Sigma}} \max_{\theta\in\{0,1\}^k} 2^2 E_\theta \left\|\hat{\Sigma} - \Sigma(\theta)\right\|^2 \\ &\geq \min_{H\left(\theta,\tilde{\theta}\right)\geq 1} \frac{\left\|\Sigma(\theta) - \Sigma\left(\tilde{\theta}\right)\right\|^2}{H\left(\theta,\tilde{\theta}\right)} \frac{k}{2} \min_{H\left(\theta,\tilde{\theta}\right)=1} \left\|P_\theta \wedge P_{\tilde{\theta}}\right\|.\end{aligned}$$

It is easy to check by our construction that there exists some constant $c_1 > 0$, such that the first factor above is lower bounded by $c_1 k a^2$, i.e.,

$$\min_{H\left(\theta,\tilde{\theta}\right)\geq 1} \frac{\left\|\Sigma(\theta) - \Sigma\left(\tilde{\theta}\right)\right\|^2}{H\left(\theta,\tilde{\theta}\right)} \geq c_1 k a^2.$$

In addition, it can be shown that the total variation affinities between the pairs of distributions satisfy $\min_{H\left(\theta,\tilde{\theta}\right)=1} \left\|P_\theta \wedge P_{\tilde{\theta}}\right\| \geq c_2$ for some positive constant c_2. Putting all together, the minimax lower bound $n^{-\frac{2\alpha}{2\alpha+1}}$ follows from the above results with the choice of $k = n^{\frac{1}{2\alpha+1}}$.

Other lower bound arguments based on Assouad's lemma have also been established in the literature. For instance, The lower bound rates $n^{-\frac{2\alpha+1}{2\alpha+2}}$ and $n^{-\frac{\alpha}{\alpha+1}}$ of estimating bandable covariance matrix under the Frobenius norm and matrix ℓ_1 norm over class $\mathcal{G}_\alpha(M_1)$ defined in (2) respectively. See Cai et al. (2010) and Cai and Zhou (2012a). We omit the details here.

4.3. Application of Le Cam's method to estimating entries of precision matrices

In Section 3.2, we discussed that the estimator $\hat{\omega}_{ij}$ for each pair of i, j attains the rate of convergence $\max\{C_1(c_{n,p}\log p)/n, C_2 n^{-1/2}\}$. Since the proof

of parametric lower bound $n^{-1/2}$ is trivial, we focus on the novel lower bound $(c_{n,p}\log p)/n$ only and apply Le Cam's method to show that it is indeed a lower bound for estimating ω_{ij}. Assume $p \geq c_{n,p}^{\upsilon}$ with $\nu > 2$. In particular, a finite collection of distributions $\mathcal{F}_{\text{sub}} \subset \mathcal{HP}(c_{n,p}, M)$ is carefully constructed in the next paragraph.

Without loss of generality, we consider estimating ω_{11} and ω_{12}. Define $\Omega_0 = (\omega_{kl}^{(0)})_{p\times p} = \Sigma_0^{-1}$, where $\Sigma_0 = (\sigma_{kl}^{(0)})_{p\times p}$ is a matrix with all diagonal entries equal to 1, $\sigma_{12}^{(0)} = \sigma_{21}^{(0)} = b$ and the rest all zeros. The constant b will be chosen later. It is easy to see that Ω_0 also has a very simple form with $\omega_{11}^{(0)} = \omega_{22}^{(0)} = (1-b^2)^{-1}$, $\omega_{12}^{(0)} = -b(1-b^2)^{-1}$, $\omega_{ii}^{(0)} = 1$ for $i \geq 3$ and all other entries being zeros. Besides, denote by $\mathcal{A}$ the collection of all $p \times p$ symmetric matrices with exactly $(c_{n,p}-1)$ entries equal to 1 between the third and the last entries on the first row/column and the rest all zeros. Now based on Ω_0 and $\mathcal{A}$, the finite collection $\mathcal{F}_{\text{sub}} = \{\Omega_0, \Omega_1, \ldots, \Omega_{m_*}\}$ can be defined formally as follows

$$\mathcal{F}_{\text{sub}} = \left\{\Omega : \Omega = \Omega_0 \text{ or } \Omega = (\Sigma_0 + aA)^{-1}, \text{ for some } A \in \mathcal{A}\right\},$$

where $a = (\tau_1 \log p/n)^{1/2}$ and $b = (1-1/M)/2$ for some sufficiently small positive τ_1, depending on M, b and ν. In other words, we use subscripts $1, \ldots, m_*$ to label those $\Omega = (\Sigma_0 + aA)^{-1}$ over $A \in \mathcal{A}$. The cardinality of $\mathcal{F}_{\text{sub}}$ is $1 + m_*$, where $m_* = \binom{p-2}{c_{n,p}-1}$. To show that $\mathcal{F}_{\text{sub}}$ is a subclass of $\mathcal{HP}(c_{n,p}, M)$, we check the sparsity as well as the spectra of each element. First, it is easy to see that number of nonzero off-diagonal entries in Ω_m, $0 \leq m \leq m_*$ is no more than $c_{n,p}$ per row/column by its construction. Second, tedious calculations yield that the spectra of Ω_0 are in $[(1+b)^{-1}, (1-b)^{-1}]$ and the spectra of Ω_m are in $[(1+g)^{-1}, (1-g)^{-1}]$, where $g = \sqrt{b^2 + (c_{n,p}-1)\,a^2}$. Hence we obtain that the spectra of each element of $\mathcal{F}_{\text{sub}}$ are between $[1/M, M]$ by definitions of a and b. Thus $\mathcal{F}_{\text{sub}} \subset \mathcal{HP}(c_{n,p}, M)$.

The motivation of our construction is that a signal-to-noise ratio level a on each entry of the covariance matrix with the sparsity $c_{n,p}$ is able to accumulate to a level of $c_{n,p}a^2 \asymp (c_{n,p}\log p)/n$ on some entry of the precision matrix. Indeed, it is easy to check that

$$\inf_{1\leq m\leq m_*} \left|\omega_{11}^{(m)} - \omega_{11}^{(0)}\right| \geq C_{11}k_{n,p}a^2 \text{ and } \inf_{1\leq m\leq m_*} \left|\omega_{12}^{(m)} - \omega_{12}^{(0)}\right| \geq C_{12}k_{n,p}a^2.$$

Let $\mathbb{P}_{\Omega_m}$ denote the joint distribution of $X_1, \ldots, X_n$, i.i.d. copies of $N(0, \Omega_m^{-1})$, $0 \leq m \leq m_*$. It can be shown that $\left\|\mathbb{P}_{\Omega_0} \wedge \bar{\mathbb{P}}\right\| \geq C_0$. Finally, applying Lemma 4, together with above two facts, we obtain the lower bounds of estimating ω_{11} and ω_{12} as follows, which match the upper bounds attained by the corresponding estimators in Section 3.2.

$$\inf_{\tilde{\omega}_{11}} \sup_{\Omega\in\mathcal{F}_{\text{sub}}} \mathbb{E}_\Omega|\tilde{\omega}_{11} - \omega_{11}| \geq \frac{1}{2}l_{\min}\left\|\mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}}\right\| \geq \frac{C_{11}C_0\tau_1 k_{n,p}\log p}{2n},$$

$$\inf_{\tilde{\omega}_{12}} \sup_{\Omega\in\mathcal{F}_{\text{sub}}} \mathbb{E}_\Omega|\tilde{\omega}_{12} - \omega_{12}| \geq \frac{1}{2}l_{\min}\left\|\mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}}\right\| \geq \frac{C_{12}C_0\tau_1 k_{n,p}\log p}{2n}.$$

The Le Cam's method is applied to establish the second term of the minimax lower bound $\min\{\frac{\lambda_{n,p}^2 r_{n,p}}{n}, \lambda_{n,p}^2\}$ for estimating sparse spiked covariance matrices over the parameter space $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ defined in (7) under the squared spectral norm loss. To calculate the total variation affinity used in Le Cam's method, an interesting analysis on the symmetric random walk stopped at a hypergeometrically distributed time is established. For reasons of space, we omit the details. See Cai, Ma, and Wu (2015) for further details.

The Le Cam's method can also be used to obtain the lower bounds of other covariance/precision matrix estimation problems. For example, the lower bound $(\log p/n)^{1/2}$ is derived for bandable class $\mathcal{F}_\alpha(M_0, M)$ defined in (1) under the spectral norm by picking a collection of covariance/precision matrices with nonzero value only on the diagonal entries. See Cai et al. (2010). Le Cam's method is also applied in Cai and Zhou (2012a) to obtain the lower bound $c_{n,p}^2(\log p)/n$ for estimating sparse covariance matrices over the class $\mathcal{H}(c_{n,p})$ defined in (4) under the matrix ℓ_1 norm. It is worthwhile to point out that the lower bound under the spectral norm loss discussed later using Le Cam-Assouad's method immediately implies this lower bound under the matrix ℓ_1 norm. However, the construction using Le Cam's method is much easier when the loss function is the matrix ℓ_1 norm.

4.4. Application of Le Cam-Assouad's method to estimating sparse precision matrices

In Section 1.2, we introduced the classes of parameters $\mathcal{HP}(c_{n,p}, M)$ defined in (8) to model the sparse structure of precision matrices. We have seen that Sun and Zhang (2013) show that the estimator $\tilde{\Omega}^{SL}$ attains the rate of convergence $c_{n,p}^2(\log p)/n$ under the spectral norm over $\mathcal{HP}(c_{n,p}, M)$ in Theorem 13 of Section 3.1. Moreover, the estimator $\hat{\Omega}^{ANT}$ in Section 3.2 also attains the same rate. In this section, we will apply Le Cam-Assouad's method to show that these estimators are indeed rate optimal. The same technique was also used to show the rate-optimality of the ACLIME estimator proposed in Cai et al. (2012) over a different parameter space.

The Le Cam-Assouad's method was originally developed in Cai and Zhou (2012b) to establish a rate sharp lower bound for estimating sparse covariance matrices over the parameter space $\mathcal{H}(c_{n,p})$ defined in (4) under the squared spectral norm loss. It was shown that a lower bound in such a setting is $c_{n,p}^2(\log p)/n$. Hence the thresholding estimator defined in (13) attaining the rate of convergence $c_{n,p}^2(\log p)/n$ is indeed rate optimal over $\mathcal{H}(c_{n,p})$. The main idea for the lower bound proof is similar to that of estimating sparse precision matrices and is hence omitted here. See Cai and Zhou (2012b) for the detailed analysis.

Again, the key of the minimax lower bound proof is to carefully construct a finite collection of distributions in $\mathcal{HP}(c_{n,p}, M)$. We assume $p > c_1 n^\beta$ for some $\beta > 1$ and $c_1 > 0$. In the current setting of estimating sparse precision matrices over the class $\mathcal{HP}(c_{n,p}, M)$, the parameter subset $\mathcal{F}_{\text{sub}}$ is constructed as follows.

Let $r = \lceil p/2 \rceil$ and let B be the collection of all vectors $(b_j)_{1 \le j \le p}$ such that $b_j = 0$ for $1 \le j \le p - r$ and $b_j = 0$ or 1 for $p - r + 1 \le j \le p$ under the constraint $\|b\|_0 = k = \lceil c_{n,p}/2 \rceil$. For each $b \in B$ and each $1 \le m \le r$, define a $p \times p$ matrix $\lambda_m(b)$ by making the mth row of $\lambda_m(b)$ equal to b and the rest of the entries 0. It is clear that the cardinality of B is $\binom{r}{k}$. Set $\Gamma = \{0, 1\}^r$. Hence each component b_i of $\lambda = (b_1, \ldots, b_r) \in \Lambda$ can be uniquely associated with a $p \times p$ matrix $\lambda_i(b_i)$. Now it is the time to define Λ as the set of all matrices λ with every column sum less than or equal to $2k$. Define $\Theta = \Gamma \otimes \Lambda$ and let $\epsilon_{n,p} \in \mathbb{R}$ be fixed, whose value will be chosen later. For each $\theta = (\gamma, \lambda) \in \Theta$ with $\gamma = (\gamma_1, \ldots, \gamma_r)$ and $\lambda = (b_1, \ldots, b_r)$, we associate θ with a precision matrix $\Omega(\theta)$ by

$$\Omega(\theta) = I + \epsilon_{n,p} \sum_{m=1}^{r} \gamma_m \lambda_m(b_m).$$

Finally we define the finite collection $\mathcal{F}_{\text{sub}}$ of precision matrices as

$$\mathcal{F}_{\text{sub}} = \left\{ \Omega(\theta) : \Omega(\theta) = I + \epsilon_{n,p} \sum_{m=1}^{r} \gamma_m \lambda_m(b_m),\ \theta = (\gamma, \lambda) \in \Theta \right\}.$$

Set $\epsilon_{n,p} = \upsilon((\log p)/n)^{1/2}$ for some sufficiently small positive υ. Now we can check that each $\Omega(\theta) \in \mathcal{F}_{\text{sub}}$ is diagonal dominated and further satisfies the bounded spectrum condition. Clearly for each $\Omega(\theta) \in \mathcal{F}_{\text{sub}}$, we have

$$\max_{1 \le i \le p} \sum_{j \ne i} \min \left\{ 1, \frac{|\omega_{ij}|}{\sqrt{(\log p)/n}} \right\} \le c_{n,p}.$$

Therefore, we obtain that $\mathcal{F}_{\text{sub}} \in \mathcal{HP}(c_{n,p}, M)$.

Let $X_1, \ldots, X_n$ be i.i.d. copies of $N(0, \Omega(\theta)^{-1})$ with $\theta \in \Theta$ and denote the joint distribution by $\mathbb{P}_\theta$. Applying Lemma 6 to the parameter space Θ indexing $\mathcal{F}_{\text{sub}}$ with $s = 2$ and metric d being the spectral norm, we have

$$\inf_{\hat{\Omega}} \max_{\Omega(\theta) \in \mathcal{F}_{\text{sub}}} 2^2 E_\theta \left\| \hat{\Omega} - \Omega(\theta) \right\|^2 \ge \alpha \cdot \frac{p}{4} \cdot \min_i \left\| \bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i} \right\| \tag{38}$$

where the per comparison loss α is defined in (35) and the mixture distributions $\bar{\mathbb{P}}_{0,i}$ and $\bar{\mathbb{P}}_{1,i}$ are defined as in (33). It can be shown that the per comparison loss $\alpha \ge \frac{(k\epsilon_{n,p})^2}{p}$ and the affinity $\min_i \left\| \bar{\mathbb{P}}_{0,i} \wedge \bar{\mathbb{P}}_{1,i} \right\| \ge c_1$ with a constant $c_1 > 0$. Plugging these two facts into equation (38), we obtain the desired minimax lower bound for estimating a sparse precision matrix over $\mathcal{HP}(c_{n,p}, M)$,

$$\inf_{\hat{\Omega}} \sup_{\mathcal{HP}(c_{n,p}, M)} \mathbb{E} \left\| \hat{\Omega} - \Omega(\theta) \right\|^2 \ge \frac{(k\epsilon_{n,p})^2}{p} \cdot \frac{c_1 p}{16} = c_2 c_{n,p}^2 \frac{\log p}{n}$$

for some constant $c_2 > 0$.

Besides covariance and precision matrix estimation problems, Le Cam-Assouad's method is also appropriate for lower bound proofs in other matrix estimation problems. One particular example is the volatility matrix estimation problem for high-dimensional diffusions. For more details, please refer to Tao et al. (2013).

4.5. Application of Fano's lemma to estimating toeplitz covariance matrices

In Section 2.2, we consider the problem of estimating Toeplitz covariance matrices. In particular, for estimation over the parameter space $\mathcal{FT}_\alpha(M_0, M)$ which is defined in terms of the smoothness of the spectral density f in (3), Cai, Ren, and Zhou (2013) showed that the tapering estimator $\hat{\Sigma}^{Toep}_{T,k_{To}}$ attains the optimal rate of convergence $(\log(np)/np)^{\frac{2\alpha}{2\alpha+1}}$ under the spectral norm. We briefly present a minimax lower bound argument, focusing on the use of Fano's lemma in this section.

Let $X_1, \ldots, X_n$ be i.i.d. copies of $N(0, \Sigma)$, where the covariance sequence $(\sigma_0, \sigma_1, \ldots, \sigma_p)$ is given via its corresponding spectral density $f = \frac{1}{2\pi}(\sigma_0 + 2\sum_{m=1}^{\infty} \sigma_m \cos mx)$ in $\mathcal{FT}_\alpha(M_0, M)$. There are two main steps in the lower bound argument. The first step is to construct a more informative model which is exactly equivalent to a Gaussian scale model under which one observes

$$Z_{ij} = S_p(f)^{1/2}\left(\frac{2\pi j}{2p-1}\right)\xi_{ij}, \text{ with } \xi_{ij} \overset{i.i.d.}{\sim} N(0,1), \tag{39}$$

for $|j| \le p-1$, and $i = 1, 2, \ldots n$. Here $S_p(f)(x) = \frac{1}{2\pi}(\sigma_0 + 2\sum_{m=1}^{p-1} \sigma_m \cos mx)$ is the partial sum of f with order p. The advantage of this more informative model is to make the analysis much easier. (see Cai, Ren, and Zhou (2013) for details). From now on we focus on this more informative model. The second step is to establish a minimax lower bound for this Gaussian scale model, which automatically provides a lower bound for the original model. We elaborate on the second step which mainly involves the construction of finite collection of spectral densities $\mathcal{F}_{\text{sub}} = \{f_0, f_1, \ldots f_{k_*/2}\} \subset \mathcal{FT}_\alpha(M_0, M)$ as follows.

Define $f_0 = M_0/2$ and f_i as follows,

$$f_i = f_0 + \tau\epsilon_{n,p}^{\alpha}\left[A\left(\frac{x - \epsilon_{n,p}(i-0.5)}{\epsilon_{n,p}}\right) + A\left(\frac{x + \epsilon_{n,p}(i-0.5)}{\epsilon_{n,p}}\right)\right], \ \epsilon_{n,p} = 2\pi/k_* \tag{40}$$

where $i = 1, 2, \ldots, k_*/2$ with $k_* = \left\lfloor (np/\log(np))^{\frac{1}{2\beta+1}} \right\rfloor$, and $A(u) = \exp(-\frac{1}{1-4u^2})1_{\{|2u|<1\}}$. It is easy to check that each distribution in our collection $f_i \in \mathcal{FT}_\alpha(M_0, M)$ by setting $\tau > 0$ sufficiently small, noting $A \in C^\infty(\mathbb{R}) \cap \mathcal{FT}_\alpha\left(e^{-1}, 1/2\right)$. Therefore, we have $\mathcal{F}_{\text{sub}} \subset \mathcal{FT}_\alpha(M_0, M)$.

Now we apply Lemma 7 to the parameter space $\mathcal{F}_{\text{sub}}$ with the distance $d = \|\cdot\|_\infty$, the sup-norm. Careful calculation implies that the assumption (36) in Lemma 7 is satisfied with $m_* = k_*/2$ and $\mathbb{P}_{f_i}$ the probability distribution of the Gaussian scale model in (39) indexed by $f_i \in \mathcal{F}_{\text{sub}}$. Hence we obtain that there exist some positive constants c_i, $i = 1, 2, 3$ such that

$$\begin{aligned}\inf_{\tilde{f}} \sup_{\mathcal{F}_{\text{sub}}} \mathbb{E}\left\|\tilde{f} - f\right\|_\infty^2 &\ge c_1 \min_{i \ne j} \|f_i - f_j\|_\infty^2 \\ &\ge c_2\left(\tau\epsilon_{n,p}^{\beta}\right)^2 \ge c_3\left(np/\log(np)\right)^{-\frac{2\beta}{1+2\beta}}.\end{aligned} \tag{41}$$

Finally, a close connection between autocovariance matrix and spectral density function implies that, for our construction of $\mathcal{F}_{\text{sub}}$, it can be shown that

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{FT}_\alpha(M_0,M)} \mathbb{E}\left\|\hat{\Sigma}-\Sigma\right\|^2 \geq c_0 \inf_{\tilde{f}} \sup_{\mathcal{F}_{\text{sub}}} \mathbb{E}\left\|\tilde{f}-f\right\|_\infty^2. \tag{42}$$

Hence the minimax lower bound for estimating a Toeplitz covariance matrix over the collection $\mathcal{FT}_\alpha(M_0, M)$ is obtained by putting (41) and (42) together.

For estimating sparse spiked covariance matrices over the parameter space $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ defined in (7) under the squared spectral norm loss, Fano's lemma is also used to obtain the first term in the minimax lower bound $\min\{\frac{(\lambda_{n,p}+1)c_{n,p}}{n}\log\frac{ep}{c_{n,p}}, \lambda_{n,p}^2\}$ in Theorem 4. See Theorem 12 in Section 2.4 for the method attaining this minimax optimal rate. Note that this term does not depend on $r_{n,p}$ and is based on a sparse vector estimation argument only. Hence the analysis is standard given the packing number of sparse vectors under the vector ℓ_2 norm and we omit the proof in this paper. See Theorem 4 in Cai et al. (2015) (also Theorem 2 in Birnbaum et al. (2013)) for further details.

Besides covariance and precision matrix estimation problems, Fano's lemma has also been used in other matrix estimation problems. For example, Rohde and Tsybakov (2011) applied it in a trace regression model to provide a lower bound of low-rank matrix estimation under the Frobenius norm.

5. Discussions

We have considered optimal estimation of high-dimensional covariance and precision matrices under various structural assumptions. Minimax rates of convergence are established and rate-optimal adaptive procedures are constructed. For ease of presentation, we have so far assumed that the random vector X is centered. As mentioned in the introduction, this assumption is not essential. We will elaborate on this point here. The estimators introduced in the previous sections are positive definite with high probability, but not guaranteed so for a given sample. It is sometimes desirable to have estimators that are guaranteed to be positive (semi)-definite. We will show that a simple additional step will lead to a desirable estimator with the same theoretical guarantees. We also discuss in this section a related problem, hypothesis testing on the covariance structure.

5.1. Non-centered case

Suppose $\mathbb{E}(X) = \mu$ with μ unknown. In this case, μ can be estimated by the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$. We can then apply the corresponding procedures such as banding, tapering, thresholding or regression to the sample covariance matrix $\hat{\Sigma}_n = \mathbf{X}'\mathbf{X}/n - \hat{\mu}\hat{\mu}'$. In fact, all the results remain the same in the unknown mean case except those for estimating the Toeplitz covariance matrix.

All the rate-optimal procedures introduced so far are translation invariant, hence we can assume $\mu = 0$. Those covariance estimators, which are directly

based on the sample covariance matrix, now depends on $\hat{\Sigma}_n = \mathbf{X}'\mathbf{X}/n - \hat{\mu}\hat{\mu}'$. Compared to the term $\mathbf{X}'\mathbf{X}/n$ which is the sample covariance matrix when the mean is known, the term $\hat{\mu}\hat{\mu}'$ usually yields a negligible estimation error. In particular, note that $\mathbb{E}\hat{\mu}\hat{\mu}' = \Sigma/n$ since $\mu = 0$. Clearly the contribution of this term entrywise is negligible with respect to the noise level $n^{-1/2}$. Globally, the contribution of $\hat{\mu}\hat{\mu}'$ is usually also negligible following an analysis that is similar to that of $\mathbf{X}'\mathbf{X}/n$, as long as the optimal rate is not faster than $n^{-1/2}$. This can be made rigorous for those covariance estimators in Section 2 except the estimator of Toeplitz covariance matrices. See, for example, Remark 1 in Cai et al. (2010). In the Toeplitz case, the effective sample size for estimating each autocovariance σ_i is far larger than n but there are only n samples for estimating its mean. The issue due to the unknown mean is no longer negligible unless extra information on μ is imposed. For example, under the condition that all coordinates of the mean are equal to some constant c_u, we can estimate it using all np samples by $\hat{c}_u = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} X_j^{(i)}$. Then the Toeplitz estimator depends on $\mathbf{X}'\mathbf{X}/n - \hat{c}_u\mathbf{1}\mathbf{1}'$, where the second term $\hat{c}_u \sim O_p((np)^{-1/2})$ is of higher order and can be ignored. In this setting, it can be shown that the results of Toeplitz covariance estimators remain valid.

For the sparse precision matrix classes, we introduced rate optimal matrix estimators under the spectral norm in Section 3.1 and rate optimal estimators of each entry ω_{ij} in Section 3.2. Both estimators are derived through a regression approach, motivated by the conditional distribution $N\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}X_{A^c}, \Omega_{A,A}^{-1}\right)$ of $X_A|X_{A^c}$, where either the index set A is a singleton in Equation (16) or $A = \{i, j\}$ in Equation (20). The corresponding regression model of the data matrix can be written as $\mathbf{X}_A = \mathbf{X}_{A^c}\beta + \epsilon_A$. Then the analysis of both estimators involves the scaled Lasso in Equations (17) and (21). When taking the unknown mean $\hat{\mu}$ into account, the analysis of scaled Lasso is applied to the data matrix as follows

$$\mathbf{X}_A - \hat{\mu}_A\mathbf{1} = \left(\mathbf{X}_{A^c} - \hat{\mu}_{A^c}\mathbf{1}\right)\beta + \epsilon_A - \bar{\epsilon}_A\mathbf{1},$$

where $\bar{\epsilon}_A$ is the sample mean of the noise vector ϵ_A. Note that the extra sample mean terms introduced above have a higher order, for example, $\bar{\epsilon}_A \sim O_p(n^{-1/2})$. As a consequence, the contribution of these terms is also negligible and all the results remain valid. For more details, see, for instance, the discussion section in Ren et al. (2015).

5.2. Positive (semi-)definiteness

In many applications, the positive (semi-) definiteness of the covariance or precision matrix estimator is usually required. Although nearly all estimators of both covariance and precision matrices we surveyed in the current paper are symmetric, they are not guaranteed to be positive (semi-) definite. Under the mild condition that the population covariance is nonsingular, it follows from the consistency results that those estimators are positive definite with high probability. Whenever an estimator $\hat{A}$ is not positive semi-definite, a simple extra

step can make the final estimator $\hat{A}_+$ positive semi-definite and also achieve the optimal rate of convergence.

Write the eigen-decomposition of the estimator $\hat{A}$ as

$$\hat{A} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{v}_i \hat{v}_i',$$

where eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ and $\hat{v}_i$'s are the corresponding eigenvectors. Define the final estimator

$$\hat{A}_+ = \hat{A} + |\hat{\lambda}_p| I_p \cdot I\{\hat{\lambda}_p < 0\},$$

where I_p is the p-dimensional identity matrix and $I\{\hat{\lambda}_p < 0\}$ is the indicator function that $\hat{A}$ is negative definite. Then with A being the target covariance or precision matrix and λ_p being its smallest eigenvalue, we have

$$\begin{aligned} \|\hat{A}_+ - A\| &\leq \|\hat{A} - A\| + |\hat{\lambda}_p| \cdot I\{\hat{\lambda}_p < 0\} \\ &\leq \|\hat{A} - A\| + |\hat{\lambda}_p - \lambda_p| \cdot I\{\hat{\lambda}_p < 0\} \\ &\leq 2\|\hat{A} - A\|. \end{aligned}$$

Clearly $\hat{A}_+$ is positive semi-definite and enjoys the same rate of convergence as that of $\hat{A}$. This simple idea has appeared in some papers on matrix estimation. See, e.g., El Karoui (2008). Another advantage of this final procedure is that $\hat{A}_+$ has the same desirable structure of $\hat{A}$ such as bandable, sparse or Toeplitz structure.

5.3. Hypothesis testing for the covariance structure

In addition to estimation, there have been considerable recent developments on testing high-dimensional covariance structure. Unlike the estimation problems, an asymptotic null distribution of a test statistic is required explicitly such that the significance level of the test can be controlled. The asymptotic analysis can be very delicate. Various testing methods have been proposed including likelihood ratio test in Bai et al. (2009) and Jiang et al. (2012), largest eigenvalue test in Johnstone (2001), Soshnikov (2002) and Peche (2009), Frobenius distance test in Ledoit and Wolf (2002), Srivastava (2005), Birke and Dette (2005), and Chen et al. (2010), and maximum entrywise deviation test in Jiang (2004), Zhou (2007), Liu et al. (2008), Li et al. (2010), Li et al. (2012), Cai and Jiang (2011), Shao and Zhou (2014), and Cai, Liu, and Xia (2013). But unlike estimation problems there are only few optimality results on testing, see Baik et al. (2005), El Karoui (2007), Cai and Ma (2013) and Onatski et al. (2013). To show the optimality of a test, asymptotic power functions are needed under alternatives to match the lower bound. We now briefly survey some recent developments on testing the covariance structure.

Testing identity We mainly focus on the problem of testing $H_0 : \Sigma = I$. A slightly more general testing problem is that of testing sphericity $H_0 : \Sigma = \sigma^2 I$ for some unknown σ^2. For reasons of space, we omit the details on testing sphericity. Four types of test statistics have been proposed and studied in the literature: likelihood-based, largest eigenvalue-based, Frobenius distance-based and maximum entrywise deviation-based statistics.

In the classical fixed p regime, it is very natural to consider the likelihood ratio test where the test statistic $L_n^{LR} = \mathrm{tr}(\hat{\Sigma}_n) - \log\det(\hat{\Sigma}_n) - p$ weakly converges to $\chi^2_{p(p+1)/2}$ under H_0, see i.e. Anderson (2003). In the high-dimensional setting, the Chi-squared limiting null distribution is no longer valid. Bai et al. (2009) proposed a corrected LRT with the Gaussian limiting null distribution in the regime $p, n \to \infty$ and $p/n \to c \in (0, 1)$ with known population mean $\mathbb{E}(X) = 0$. Recently, Zheng et al. (2015) extended it to the setting with unknown population mean by applying a substitution principle established for the Central Limit Theorem for linear spectral statistics of sample covariance matrices with unknown mean. Jiang et al. (2012) further extended it to the regime $p < n \to \infty$ and $p/n \to c = 1$ with known mean.

Johnstone (2001) established the limiting distribution of the largest eigenvalue of the sample covariance matrix in the case of Gaussian distribution with the identity covariance matrix, following the work of Johansson (2000) in which a limit theorem for the largest eigenvalue of a complex Gaussian sample covariance matrix was proved. It is shown that its scaled limiting law is the Tracy-Widom distribution, assuming $p/n \to c \in (0, \infty)$. See The result immediately yields a test for $H_0 : \Sigma = I$ in the Gaussian case by using the largest eigenvalue of the sample covariance matrix as the test statistic. Johnstone's result was extended by Soshnikov (2002) under a sub-Gaussian assumption which is valid only for $n - p = o(p^{1/3})$ and by El Karoui (2003) which allows $p/n \to c \in [0, \infty]$ under the Gaussian assumption. In a later work, Peche (2009) further extended Johnstone's work to the regime $p/n \to c \in [0, \infty]$ with moment requirements and no Gaussianity assumption. Ma (2012) studied the convergence rate of Tracy-Widom approximation. See, for example, Lee and Schnelli (2014) and Knowles and Yin (2014) for some recent development.

Frobenius distance-based test was originally proposed by John (1971) and Nagao (1973) in the fixed p regime. The Frobenius distance between Σ and I is

$$\frac{1}{p}\mathrm{tr}\left\{(\Sigma - I)^2\right\} = \frac{1}{p}\mathrm{tr}\left(\Sigma^2\right) - \frac{2}{p}\mathrm{tr}(\Sigma) + 1, \tag{43}$$

which is zero if and only if H_0 holds. Nagao (1973) replaced Σ in (43) with the sample covariance matrix to obtain the test statistic $V = \mathrm{tr}(\hat{\Sigma}_n - I)^2/p$ while John (1971) proposed a similar statistic to test sphericity. In the regime $p/n \to c \in (0, \infty)$, Ledoit and Wolf (2002) showed that V is inconsistent and proposed a modification, which has a Gaussian limiting null distribution. Birke and Dette (2005) further investigated and modified their test statistic in the extreme cases where $p/n \to c \in \{0, \infty\}$. Note that expression (43) is a function of the first two moments of the spectra of Σ. Based on this idea, Srivastava (2005) constructed similar test statistics in the restricted regime $n = O(p^\delta)$ with $\delta \leq 1$.

In the high-dimensional setting, Chen et al. (2010) also investigated the testing problem with the Frobenius distance (43). However, instead of plugging in the sample covariance matrix to estimate $\left(\mathrm{tr}(\Sigma), \mathrm{tr}(\Sigma^2)\right)$, a U-statistic is applied to derive more accurate and reliable estimators $(\hat{T}_{1,n}, \hat{T}_{2,n})$ of $\left(\mathrm{tr}\left(\Sigma\right), \mathrm{tr}\left(\Sigma^2\right)\right)$. Chen et al. (2010) also provided a lower bound for the asymptotic power function. In particular, as long as $\|\Sigma - I\|_F \sqrt{n/p} \to \infty$, the test is consistent. In a separate paper, Cai and Ma (2013) showed that this procedure is indeed optimal under some natural alternatives. Similar results are obtained for testing sphericity as well.

Several tests based on the maximum entrywise deviations for testing the hypothesis $H_0 : R = I$, where R is the correlation matrix, have been proposed and studied in the literature. Jiang (2004) studied the asymptotic distribution of the test statistic

$$L_n = \max_{1 \le i < j \le p} |\hat{\rho}_{i,j}|, \tag{44}$$

where $\hat{\rho}_{ij}$ is the sample correlation between X_i and X_j. In particular, the Gumbel distribution is derived as the null limiting distribution of L_n

$$\lim_{n\to\infty} \mathbb{P}\left(nL_n^2 - 4\log p + \log\log p \le y\right) = \exp\left\{-\frac{1}{\sqrt{8\pi}} e^{-y/2}\right\}, \quad -\infty < y < \infty, \tag{45}$$

under the moment condition $\mathbb{E}\,|X_i|^{30+\varepsilon} < \infty$ and the regime $p/n \to \gamma \in (0, \infty)$. Jiang's work attracted considerable attention. However, the moment condition and asymptotic regime seem too restrictive. Zhou (2007) reduced the moment condition to $x^6\mathbb{P}\left(|X_1X_2| \ge x\right) \to 0$ and Liu et al. (2008) further weakened it to $x^6/\log^3 x\mathbb{P}\left(|X_1X_2| \ge x\right) \to 0$ as a special case. In general, their result allowed wide regime $cn^\alpha \le p \le Cn^\alpha$ but the moment condition also depends on α. Liu et al. (2008) also introduced some "intermediate" approximation to approximate the test statistic L_n with a much faster rate of convergence. In comparison, the convergence of L_n to the Gumbel distribution has a typical slow rate $\log\log n/\log n$ in the regime $cn^\alpha \le p \le Cn^\alpha$. Li et al. (2010) and Li et al. (2012) further showed that some moment condition is necessary. More specifically, in the bounded ratio regime $\lim p/n \in (c, C)$, if X_1 has finite second moment, then $\mathbb{E}\,|X_1|^\beta < \infty$ for all $\beta < 6$ is a necessary condition such that limiting distribution (45) holds. Cai and Jiang (2011) and Shao and Zhou (2014) generalized the polynomial regime and push it to the ultra-high-dimensional case $\log p = o(n^{\beta_\alpha})$. Shao and Zhou (2014) showed that under the moment condition $\mathbb{E}\exp\left(t\,|X_{1,1}|^\alpha\right) < \infty$ for some $t > 0$ and $\alpha \in (0, 1]$, the necessary and sufficient conditions for establishing (45) in the ultra-high-dimensional setting in terms of the optimal β_α is that $\beta_\alpha = 4/\left(4 - \alpha\right)$.

Testing more general covariance structures Hypothesis testing for other covariance structures has also been considered in the literature. These include (i) banded structure with $H_0\left(k\right) : \Sigma$ is k banded, (ii) bandable structure and (iii) Toeplitz structure. Here a matrix $\Sigma = (\sigma_{ij})$ is called k banded if $\sigma_{ij} = 0$ for all pairs (i, j) such that $|i - j| \ge k$.

Cai and Jiang (2011) and Shao and Zhou (2014) considered testing k banded structure. To test this hypothesis $H_0(k)$, analogous to the definition of L_n, Cai and Jiang (2011) proposed the following test statistic

$$L_{n,k} = \max_{|i-j|\geq k} |\hat{\rho}_{i,j}|.$$

It can be shown that $L_{n,k}$ has the same limiting Gumbel distribution in (45) as long as most correlations are bounded away from 1 and $k = o(p^\tau)$ for some small τ. Recently, Xiao and Wu (2011) established a more general result which allows dependent entries over large range of p. Instead of the $L_{n,k}$, a self-normalized version of maximum entrywise deviation is constructed in Xiao and Wu (2011) as the new test statistic,

$$M_n = \max_{1\leq i\leq j\leq p} |\sigma_{ij} - \hat{\sigma}_{ij}| / \sqrt{\hat{\tau}_{ij}},$$

where $\tau_{ij} = \mathrm{Var}(X_i X_j)$ and $\hat{\tau}_{ij}$ is the empirical counterpart. In different regimes $p = O(n^\alpha)$ and $p = o\left(\exp n^\beta\right)$, M_n is also shown to weakly converge to the Gumbel distribution. This result in turn allows testing all three structures (i), (ii), and (iii) listed above. Qiu and Chen (2012) constructed an unbiased estimator of $\Sigma_{|i-j|\geq k}\sigma_{ij}^2$ via certain U-statistic to test banded covariance structure $H_0(k)$, motivated by the Frobenius distance-based tests in Chen et al. (2010). A lower bound of asymptotic power function is also established.

6. Some open problems

Although much progress has been made on estimation of structured high-dimensional covariance and precision matrices, there are still many open problems. We conclude the paper with a brief discussion on a few interesting open problems.

6.1. Optimality for covariance matrix estimation under Schatten q norm

In addition to the matrix ℓ_ω norm and Frobenius norm considered in this paper, the Schatten q norm, which is unitarily invariant, is another commonly used matrix norm and considered in many statistical problems, including trace regression, low-rank matrix recovery and density matrix estimation in quantum tomography. The Schatten q norm is the vector ℓ_q norm of the spectra. Denote the singular values of Σ by $\{\lambda_i\}$, $i = 1, \ldots, p$. The Schatten q norm is defined by

$$\|\Sigma\|_{S_q} = \left(\sum_{i=1}^{p} \lambda_i^q\right)^{1/q}.$$

When $q = 2$, it coincides with the Frobenius norm and when $q = \infty$, it becomes the spectral norm. Estimating a covariance or precision matrix under the general

Schatten q norm is an interesting open problem. So far the corresponding optimality results for covariance and precision matrix estimation under the general Schatten q norm remain unknown. The major difficulty lies in the establishment of a rate-sharp minimax lower bound. We believe that new technical tools are needed to solve this problem.

6.2. Lower Bound via packing number

Fano's lemma is a standard tool for deriving the minimax lower bounds. It relies on a good bound for the cardinality of a suitable packing set for a given parameter space. So far little is known about the packing number for the class of sparse matrices under the spectral norm. Consider the following general sparse matrix class $\mathcal{S}(k)$ in which there are at most k ones in each row and each column

$$\mathcal{S}(k) = \left\{A_{p\times p} = (a_{ij}) : a_{ij} = 0 \text{ or } 1, \max\left\{\|A\|_{\ell_1}, \|A\|_{\ell_\infty}\right\} \le k\right\}.$$

We conjecture that there exists a "good" packing set $\mathcal{S}_{\text{sub}}(k) \subset \mathcal{S}(k)$ under the spectral norm such that for any $A_1, A_2 \in \mathcal{S}_{\text{sub}}(k)$, we have $\|A_1 - A_2\| \ge \epsilon k$ and $\log|\mathcal{S}_{\text{sub}}(k)| > \epsilon kp\log(p/k)$ for some small constant ϵ. If this statement is true, then the standard Fano's lemma can be applied to obtain the minimax lower bound under the spectral norm for estimating sparse covariance/precision/volatility matrix. As a consequence, the proof of those lower bounds in literature using Le Cam-Assouad's method introduced in Section 4 can be unified and simplified. In addition, we point out that in Theorems 3 and 5, the optimality results are obtained under the assumption $c_{n,p} \le C\sqrt{n/(\log p)^3}$ rather than the natural one $c_{n,p} \le C\sqrt{n/(\log p)}$, which is due to the Le Cam-Assouad's method employed to prove the lower bound. Therefore if the conjecture is true, we can obtain more complete optimality results without such an unpleasant sparsity condition.

Similarly, a good bound on the packing number for the class of bandable covariance matrices and the class of sparse covariance matrices under the general Schatten q norm would be very helpful for the establishment of the minimax lower bound for estimating the corresponding covariance matrices under the Schatten q norm.

6.3. Optimal estimation of matrix functionals

Many high-dimensional statistical inference problems such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) require knowledge of certain aspects, i.e., functionals, of the covariance structure, instead of the whole matrix itself. So estimation of the functionals of covariance/precision matrices is an important problem. The most common practice for matrix functional estimation is the plug-in approach: first estimating the whole matrix in a certain optimal way and then plugging-in the estimator to estimate the corresponding functional. This often leads to a sub-optimal solution.

Despite recent progress on optimality of estimating covariance and precision matrices under various matrix norms, there have been few optimality results on estimation of functionals of the covariance matrices. Cai, Liang, and Zhou (2015) obtained the limiting distribution of the log determinant of the sample covariance matrix in the Gaussian setting and applied the result to establish the optimality for estimation of the differential entropy, which is a functional of the covariance matrix. The problem of optimally estimating an individual entry of a sparse precision matrix discussed in Section 3.2 can also be viewed as estimating a functional.

Given two independent samples, $X^{(1)}, \ldots, X^{(n_1)} \overset{iid}{\sim} N_p(\mu_1, \Omega^{-1})$ and $Y^{(1)}, \ldots, Y^{(n_2)} \overset{iid}{\sim} N_p(\mu_2, \Omega^{-1})$, an important functional to estimate is $\Omega(\mu_1 - \mu_2)$. This is motivated by the linear discriminant analysis. In the ideal case when the parameters μ_1, μ_2 and Ω are known, for a new observation Z drawn with equal probability from either $N_p(\mu_1, \Omega^{-1})$ or $N_p(\mu_2, \Omega^{-1})$, the optimal classification rule is Fisher's linear discriminant rule which classifies Z to class 1 if and only if $(Z - (\mu_1 + \mu_2)/2)' \Omega (\mu_1 - \mu_2) > 0$. In applications, the parameters μ_1, μ_2 and Ω are unknown and it is a common practice to estimate them separately and then plug in. This approach has been shown to be inefficient as the discriminant depends on the parameters primarily through the functional $\Omega (\mu_1 - \mu_2)$. Cai and Liu (2011b) introduced a constrained ℓ_1 minimization method for estimating the functional $\Omega (\mu_1 - \mu_2)$ directly and proposed a classification method based on the estimator. A similar approach was also used in Mai et al. (2012). Although direct estimators of $\Omega (\mu_1 - \mu_2)$ have been proposed and used for classification, the optimality of the estimation problem remains unknown. See also El Karoui and Kösters (2011) for further discussions. It is of significant interest to study the problem of optimal estimation of the functional $\Omega (\mu_1 - \mu_2)$ under certain sparsity assumptions.

References

Agarwal, A., S. Negahban, and M. J. Wainwright (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics 40*(2), 1171–1197. MR2985947

Amini, A. A. and M. J. Wainwright (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics 37*(5), 2877–2921. MR2541450

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). Wiley. MR1990662

Assouad, P. (1983). Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique 296*(23), 1021–1024. MR0777600

Bai, Z., D. Jiang, J.-F. Yao, and S. Zheng (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics 37*(6B), 3822–3840. MR2572444

Baik, J., G. Ben Arous, and S. Péché (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability 33*(5), 1643–1697. MR2165575

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research 9*, 485–516. MR2417243

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*(4), 1535–1567. MR3357870

Berthet, Q. and P. Rigollet (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics 41*(4), 1780–1815. MR3127849

Bickel, P. J. and E. Levina (2004). Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli 10*(6), 989–1010. MR2108040

Bickel, P. J. and E. Levina (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227. MR2387969

Bickel, P. J. and E. Levina (2008b). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604. MR2485008

Bickel, P. J. and Y. Ritov (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 381–393. MR1065550

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732. MR2533469

Birke, M. and H. Dette (2005). A note on testing the covariance matrix for large dimension. *Statistics & Probability Letters 74*(3), 281–289. MR2189467

Birnbaum, A., I. M. Johnstone, B. Nadler, and D. Paul (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics 41*(3), 1055–1084. MR3113803

Cai, T. T. and T. Jiang (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics 39*(3), 1496–1525. MR2850210

Cai, T. T., T. Liang, and H. H. Zhou (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *Journal of Multivariate Analysis 137*, 161–172. MR3332804

Cai, T. T. and W. Liu (2011a). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684. MR2847949

Cai, T. T. and W. Liu (2011b). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577. MR2896857

Cai, T. T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization ap-

proach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607. MR2847973

Cai, T. T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*(501), 265–277. MR3174618

Cai, T. T., W. Liu, and H. H. Zhou (2012). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *arXiv preprint arXiv:1212.2882*.

Cai, T. T. and Z. Ma (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli 19*(5B), 2359–2388. MR3160557

Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics 41*(6), 3074–3110. MR3161458

Cai, T. T., Z. Ma, and Y. Wu (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields 161*(3-4), 781–815. MR3334281

Cai, T. T., Z. Ren, and H. H. Zhou (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields 156*(1-2), 101–143. MR3055254

Cai, T. T. and M. Yuan (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics 40*(4), 2014–2042. MR3059075

Cai, T. T., C. H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics 38*(4), 2118–2144. MR2676885

Cai, T. T. and H. H. Zhou (2012a). Minimax estimation of large covariance matrices under ℓ_1 norm. *Statistica Sinica 22*(4), 1319–1349. MR3027084

Cai, T. T. and H. H. Zhou (2012b). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics 40*(5), 2389–2420. MR3097607

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011). Robust principal component analysis? *Journal of the ACM (JACM) 58*(3), 11. MR2811000

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–304. MR0736050

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics 40*(4), 1935–1967. MR3059067

Chen, S. X., L. X. Zhang, and P. S. Zhong (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association 105*(490), 810–819. MR2724863

Chen, X., M. Xu, and W. B. Wu (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*(6), 2994–3021. MR3161455

d'Aspremont, A., O. Banerjee, and L. El Ghaoui (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications 30*(1), 56–66. MR2399568

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review 49*(3), 434–448. MR2353806

Davidson, K. R. and S. J. Szarek (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces 1*, 317–366. MR1863696

Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis 7*(1), 1–46. MR0264450

Donoho, D. L. and R. C. Liu (1991). Geometrizing rates of convergence, II. *The Annals of Statistics 19*(2), 633–667. MR1105839

El Karoui, N. (2003). On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*.

El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability 35*(2), 663–714. MR2308592

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics 36*(6), 2717–2756. MR2485011

El Karoui, N. and H. Kösters (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*.

Engle, R. and M. Watson (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association 76*(376), 774–781.

Fan, J. (1991). On the estimation of quadratic functionals. *The Annals of Statistics 19*(3), 1273–1294. MR1126325

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197. MR2472991

Fan, J., Y. Liao, and M. Mincheva (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics 39*(6), 3320–3356. MR3012410

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680. MR3091653

Franaszczuk, P., K. Blinowska, and M. Kowalczyk (1985). The application of parametric multichannel spectral estimates in the study of electrical brain activity. *Biological Cybernetics 51*(4), 239–247.

Friedman, J., T. Hastie, H. Hofling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302–332. MR2415737

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

Friston, K. J., P. Jezzard, and R. Turner (1994). Analysis of functional MRI time-series. *Human Brain Mapping 1*(2), 153–171.

Fuhrmann, D. R. (1991). Application of Toeplitz covariance estimation to adaptive beamforming and detection. *IEEE Transactions on Signal Processing 39*, 2194–2198.

Furrer, R. and T. Bengtsson (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis 98*(2), 227–255. MR2301751

Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics 15*(3), 502–523. MR2291261

Gao, C., Z. Ma, and H. H. Zhou (2014). Sparse CCA: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.

Gaspari, G. and S. E. Cohn (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society 125*(554), 723–757.

Goldfarb, D. and G. Iyengar (2003). Robust portfolio selection problems. *Mathematics of Operations Research 28*(1), 1–38. MR1961265

Golubev, G. K., M. Nussbaum, and H. H. Zhou (2010). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *The Annals of Statistics 38*(1), 181–214. MR2589320

Grenander, U. and G. Szegö (1958). *Toeplitz Forms and Their Applications*, Volume 321. Univ of California Press. MR0094840

Hamill, T. M., J. S. Whitaker, and C. Snyder (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review 129*(11), 2776–2790.

Houtekamer, P. L. and H. L. Mitchell (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review 129*(1), 123–137.

Hsieh, C.-J., I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pp. 2330–2338.

Hsieh, C.-J., M. A. Sustik, I. Dhillon, P. Ravikumar, and R. Poldrack (2013). BIG & QUIC: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*, pp. 3165–3173.

Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85–98. MR2277742

Javanmard, A. and A. Montanari (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *Information Theory, IEEE Transactions on 60*(10), 6522–6554. MR3265038

Jiang, D., T. Jiang, and F. Yang (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference 142*(8), 2241–2256. MR2911842

Jiang, T. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *The Annals of Applied Probability 14*(2), 865–880. MR2052906

Johansson, K. (2000). Shape fluctuations and random matrices. *Communications in Mathematical Physics 209*(2), 437–476. MR1737991

John, S. (1971). Some optimal multivariate tests. *Biometrika 58*(1), 123–127. MR0275568

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics 29*(2), 295–327. MR1863961

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486), 682–693. MR2751448

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547. MR2002634

Knowles, A. and J. Yin (2014). Anisotropic local laws for random matrices. *arXiv preprint arXiv:1410.3516*.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics 37*(6B), 4254–4278. MR2572459

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics 1*(1), 38–53. MR0334381

Ledoit, O. and M. Wolf (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics 30*(4), 1081–1102. MR1926169

Lee, J. O. and K. Schnelli (2014). Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *arXiv preprint arXiv:1409.4979*.

Li, D., W. D. Liu, and A. Rosalsky (2010). Necessary and sufficient conditions for the asymptotic distribution of the largest entry of a sample correlation matrix. *Probability Theory and Related Fields 148*(1-2), 5–35. MR2653220

Li, D., Y. Qi, and A. Rosalsky (2012). On Jiang's asymptotic distribution of the largest entry of a sample correlation matrix. *Journal of Multivariate Analysis 111*, 256–270. MR2944420

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics 41*(6), 2948–2978. MR3161453

Liu, W.-D., Z. Lin, and Q.-M. Shao (2008). The asymptotic distribution and Berry–Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability 18*(6), 2337–2366. MR2474539

Ma, Z. (2012). Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices. *Bernoulli 18*(1), 322–359. MR2888709

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801. MR3099121

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42. MR2899661

McMurry, T. L. and D. N. Politis (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis 31*(6), 471–482. MR2732601

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*(3), 1436–1462. MR2278363

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics 36*(6), 2791–2817. MR2485013

Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics 1*(4), 700–709. MR0339405

Onatski, A., M. Moreira, and M. Hallin (2013). Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics 41*(3), 1204–1231. MR3113808

Pang, H., H. Liu, and R. Vanderbei (2014). The FASTCLIME package for linear programming and large-scale precision matrix estimation in R. *The Journal of Machine Learning Research 15*(1), 489–493.

Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *The Annals of Mathematical Statistics 28*(2), 329–348. MR0088833

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica 17*(4), 1617. MR2399865

Peche, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields 143*(3-4), 481–516. MR2475670

Qiu, Y. and S. X. Chen (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *The Annals of Statistics 40*(3), 1285–1314. MR3015026

Quah, D. (2000). Internet cluster emergence. *European Economic Review 44*(4), 1032–1044.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980. MR2836766

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *The Annals of Statistics 43*(3), 991–1026. MR3346695

Ren, Z. and H. H. Zhou (2012). Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics 40*(4), 1989–1996. MR3059072

Rigollet, P. and A. B. Tsybakov (2012). Comment: Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica 22*(4), 1358–1367. MR3027087

Rohde, A. and A. B. Tsybakov (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics 39*(2), 887–930. MR2816342

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360. MR0429063

Ross, S. A. (1977). The capital asset pricing model (CAPM), short-sale restrictions and related issues. *The Journal of Finance 32*(1), 177–183.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515. MR2417391

Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186. MR2504372

Rudelson, M. and S. Zhou (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on 59*(6), 3434–3447. MR3061256

Runge, J., V. Petoukhov, and J. Kurths (2014). Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate 27*(2), 720–739.

Samarov, A. (1977). Lower bound on the risk for spectral density estimates. *Problemy Peredachi Informatsii 13*(1), 67–72. MR0494767

Shao, Q.-M. (1999). A Cramér type large deviation result for Student's t-statistic. *Journal of Theoretical Probability 12*(2), 385–398. MR1684750

Shao, Q.-M. and W.-X. Zhou (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *The Annals of Probability 42*(2), 623–648. MR3178469

Shen, D., H. Shen, and J. Marron (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis 115*, 317–333. MR3004561

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis 99*(6), 1015–1034. MR2419336

Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics 108*(5-6), 1033–1056. MR1933444

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society 35*(2), 251–272. MR2328427

Sun, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898. MR2999166

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research 14*(1), 3385–3418. MR3144466

Tao, M., Y. Wang, and H. H. Zhou (2013). Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors. *The Annals of Statistics 41*(4), 1816–1864. MR3127850

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer. MR2724359

van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360–1392. MR2576316

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202. MR3224285

Vanderberghe, L., S. Boyd, and S. P. Wu (1998). Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications 19*(2), 499–533. MR1614078

Visser, H. and J. Molenaar (1995). Trend estimation and regression analysis in climatological time series: An application of structural time series models and the Kalman filter. *Journal of Climate 8*(5), 969–979.

Vu, V. Q. and J. Lei (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics 41*(6), 2905–2947. MR3161452

Wachter, K. W. (1976). Probability plotting points for principal components. In *Ninth Interface Symposium Computer Science and Statistics*, pp. 299–308. Prindle, Weber and Schmidt, Boston.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability 6*(1), 1–18. MR0467894

Wang, T., Q. Berthet, and R. J. Samworth (2014). Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology 5*(11), R92.

Witten, D., J. Friedman, and N. Simon (2011). New insights and faster computations for the graphical Lasso. *Journal of Computational and Graphical Statistics 20*(4), 892–900. MR2878953

Witten, D., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America 102*(40), 14150–14154. MR2172215

Wu, W. B. and M. Pourahmadi (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika 90*(4), 813–844. MR2024760

Wu, W. B. and M. Pourahmadi (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica 19*(4), 1755–1768. MR2589209

Xiao, H. and W. B. Wu (2011). Simultaneous inference of covariances. *arXiv preprint arXiv:1109.0524*.

Xiao, H. and W. B. Wu (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics 40*(1), 466–493. MR3014314

Xiao, L. and F. Bunea (2014). On the theoretic and practical merits of the banding estimator for large covariance matrices. *arXiv preprint arXiv:1402.0844*.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*(5), 1564–1599. MR1742500

Ye, F. and C.-H. Zhang (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research 11*, 3519–3540. MR2756192

Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, 423–435. MR1462963

Yuan, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286. MR2719856

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35. MR2367824

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242. MR3153940

Zhang, C.-H. and T. Zhang (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science 27*(4), 576–593. MR3025135

Zheng, S., Z. Bai, and J. Yao (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *The Annals of Statistics 43*(2), 546–591. MR3316190

Zhou, W. (2007). Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Transactions of the American Mathematical Society 359*(11), 5345–5363. MR2327033

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286. MR2252527