The Annals of Statistics
2020, Vol. 48, No. 4, 2428–2454
https://doi.org/10.1214/19-AOS1893

MINIMAX ESTIMATION OF LARGE PRECISION MATRICES WITH BANDABLE CHOLESKY FACTOR

BY YU LIU* AND ZHAO REN**

*Department of Statistics, University of Pittsburgh, *yul125@pitt.edu; **zren@pitt.edu*

The last decade has witnessed significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this paper we focus on two commonly used types of parameter spaces and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found. Two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed. For the operator norm our optimal procedure is based on a novel local cropping estimator, targeting on all principle submatrices of the precision matrix, while for the Frobenius norm our optimal procedure relies on a delicate regression-based thresholding rule. Lepski's method is considered to achieve optimal adaptation. We further establish rate optimality in the nonparanormal model. Numerical studies are carried out to confirm our theoretical findings.

1. Introduction. Covariance matrix plays a fundamental role in many important multivariate statistical problems. They include the principal component analysis, linear and quadratic discriminant analysis, clustering analysis, regression analysis and conditional dependence relationship studies in graphical models. During the last two decades, with the advances of technology, it is very common that the datasets are high dimensional (the dimension p can be much larger than the sample size n) in many applications such as genomics, fMRI data, astrophysics, spectroscopic imaging, risk management, portfolio allocation and numerical weather forecasting [19, 24, 25, 33, 42, 48]. It has been well known that the sample covariance matrix performs poorly and can yield to invalid conclusions in the high-dimensional settings. For example, see [20, 28, 29, 43, 51, 52] for details on the limiting behaviors of the spectra of sample covariance matrices when both n and p increase.

To avoid the curse of dimensionality, certain structural assumptions are almost necessary in order to estimate the covariance matrix or its inverse, the precision matrix, consistently. In this paper we consider large precision matrix estimation with bandable Cholesky factor. Its connection with other structures are discussed at the end of the Introduction. Both the operator norm loss ($\|S\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Sx\|_2$) and the Frobenius norm loss ($\|S\|_{\mathrm{F}} = (\sum_{i,j} s_{ij}^2)^{\frac{1}{2}}$) are investigated.

Received February 2018; revised January 2019.
MSC2020 subject classifications. Primary 62H12; secondary 62F12, 62C20, 62G09.
Key words and phrases. Optimal rate of convergence, precision matrix, local cropping, Cholesky factor, minimax lower bound, thresholding, operator norm, Frobenius norm and adaptive estimation.

We begin with introducing the bandable Cholesky factor of the precision matrix. Assume that $\mathbf{X} = (X_1, \ldots X_p)^T$ is a centered, p-variate random vector with covariance matrix Σ. Let $\mathbf{a}_i = (a_{i1}, \ldots, a_{i(i-1)})^T$ be the coefficients of the population regression of X_i on its previous variables $\mathbf{X}_{1,i-1} = (X_1, X_2, \ldots, X_{i-1})^T$. In other words, $\hat{X}_i = \sum_{t=1}^{i-1} a_{it} X_t = \mathbf{X}_{1,i-1}^T \mathbf{a}_i$ is the linear projection of X_i on $\mathbf{X}_{1,i-1}$ in population (define $\hat{X}_1 = 0$). Set A as the lower triangular matrix with zeros on the diagonal and zero-padded coefficients $(\mathbf{a}_i^T, \mathbf{0})$ arranged in the rows. Denote the residual $\boldsymbol{\epsilon} = \mathbf{X} - \hat{\mathbf{X}} = (I - A)\mathbf{X}$ and $D = \mathsf{Var}(\boldsymbol{\epsilon})$. The regression theory implies the residuals are uncorrelated, and thus the matrix D is diagonal. The modified Cholesky decomposition of Ω is

$$\Omega = \Sigma^{-1} = (I - A)^T D^{-1} (I - A), \tag{1}$$

where $I - A$ is the Cholesky factor of Ω. There is a natural order on the variables based on the above Cholesky decomposition. Indeed, the well-known AR(k) model can be characterized by the k-banded Cholesky factor $A \equiv [a_{ij}]_{p\times p}$ of the precision matrix in which $a_{ij} = 0$ if $i - j > k$. Inspired by the auto-regression model, we consider the bandable structures imposed on the Cholesky factor. More specifically, for $M > 0$, $\eta > 1$ we define the parameter space $\mathcal{P}_\alpha(\eta, M)$ of precision matrices by

$$\begin{aligned}\mathcal{P}_\alpha(\eta, M) = \Big\{ &\Omega : \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ &\max_i \sum_{j<i-k} |a_{ij}| < Mk^{-\alpha}, k \in [p] \Big\}.\end{aligned} \tag{2}$$

Here, $\lambda_{\max}(\Omega)$, $\lambda_{\min}(\Omega)$ are the maximum and minimum eigenvalues of Ω and the index set $[p] = \{1, 2, \ldots, p\}$. We follow the convention that the sum over an empty set of indices is equal to zero when $i - k \le 1$. This parameter space was first proposed in [7]. The parameter α specifies how fast the sequence a_{ij} decays to zero as j goes away from i. The covariance matrix estimation problem has been extensively studied when a similar bandable structure is imposed on the covariance matrix (e.g., [7, 15]). Unlike the order in these bandable covariance matrices in which large distance $|i - j|$ implies nearly independence, the order in bandable Cholesky factor encodes a natural auto-regression interpretation in the sense that the coefficient a_{ij} is close to zero when $i - j > 0$ is large.

Although several approaches have been developed to estimate the precision matrix with bandable Cholesky factor, the optimality question remains mostly open, partially due to the following two reasons: (i) Intuitively, one would expect the minimax rate of convergence over $\mathcal{P}_\alpha(\eta, M)$ under the operator norm to be the same as that over the class of bandable covariance matrices with the same decay parameter α. Under sub-Gaussian assumptions [15] established the optimal rate of convergence $\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \asymp n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$ uniformly for all bandable covariance matrices $\Sigma = \Omega^{-1} = [\sigma_{ij}]_{p\times p}$ with bounded spectra such that $\max_i \sum_{|j-i|>k} |\sigma_{ij}| < Mk^{-\alpha}$, $k \in [p]$. To establish such a rate of convergence for $\mathcal{P}_\alpha(\eta, M)$, [34] provided a lower bound with the matching rate. However, we show a surprising result in this paper that estimation over $\mathcal{P}_\alpha(\eta, M)$ is a much harder task than that over bandable covariance matrices. Therefore, the lower bound in [34] is suboptimal, and all attempts on showing the same rate of convergence $n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$ intrinsically cannot succeed. (ii) From the methodological aspect, due to the regression interpretation of the Cholesky decomposition (1), almost all existing methods rely on an intermediate estimator of A obtained by running regularized regression of each variable against its previous variables $X_i \sim \sum_{j=1}^{i-1} a_{ij} X_j$. For instance, [7] estimated each row of A by fitting the banded regression model $X_i \sim \sum_{j=\max\{1,i-k\}}^{i-1} a_{ij} X_j$ with some bandwidth k. [53] used an AIC or BIC penalty

to pick the best bandwidth k. In addition, [27] proposed adding a lasso or ridge penalty while [36] proposed using a nested lasso penalty to the regression. See, for instance, [3, 34] for Bayesian approaches following the similar idea. The typical analysis for those estimation procedures in a row-wise fashion is to bound the operator norm by its matrix ℓ_1/ℓ_∞ norm. Although this analysis may provide optimal rates of convergence under the operator norm loss for some sparsity structure (see, i.e., [12, 16] for sparse covariance and precision matrices estimation), it might be suboptimal for the bandable structure, as seen in bandable covariance matrix estimation [7, 15]. Therefore, in order to obtain rate optimality over $\mathcal{P}_\alpha(\eta, M)$, a novel analysis or even a new estimation approach is expected.

Main results. With regard to the above two issues, we provide satisfactory solutions in this paper. We at the first time show that the rate of convergence under the operator norm over $\mathcal{P}_\alpha(\eta, M)$ is intrinsically slower than that over the counterpart class of bandable covariance matrices. This is achieved via a novel minimax lower bound construction. Moreover, in order to obtain a rate-optimal estimator, we propose a novel local cropping estimator which does not rely on any estimator of A and thus requires a new analysis. Our local cropping approach targets on accurate estimation of principal submatrices of the precision matrix under the operator norm which results in a tradeoff between one variance term and two bias terms. The name comes after the idea of estimating each principal submatrix of the precision matrix, which is to crop the center k by k submatrix of the inverse of $3k$ by $3k$ sample covariance matrix using their neighbors in two directions of the same size. (During the finalizing process of this paper, we realized that a similar estimator is independently proposed to estimate precision matrices with a different structure [26].) Since our procedure does not directly explore the structure on each row of A, the analysis of bias terms is much more involved, requiring a blockwise partition strategy. More details are discussed in Sections 2.1 and 3.1. In the end, besides $\mathcal{P}_\alpha(\eta, M)$, a similar type of classes of parameter spaces with bandable Cholesky factor is considered as well,

$$\begin{aligned}\mathcal{Q}_\alpha(\eta, M) = \big\{&\Omega : \eta^{-1} \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) < \eta, \\ &|a_{ij}| < M(i-j)^{-\alpha-1},\ j \in [i-1]\big\}.\end{aligned} \tag{3}$$

We further establish another surprising result. The optimal rates of convergence of two spaces, namely $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, are different under the operator norm. This remarkable distinction is different from the comparison of two similar types of parameter spaces for bandable covariance matrices in [15] and bandable Toeplitz covariance matrices in [13]. The contrast of minimax results on $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ is summarized in Theorem 1 below. We mainly focus on the high-dimensional setting, assuming that $\log p = O(n)$ and $n = O(p)$. Theorem 1 implies inconsistency when $\log p = O(n)$ is violated. In addition, one can easily obtain that when $n = O(p)$ is violated, the minimax rate becomes the smaller value between p/n and the one shown in Theorem 1 for each space.

THEOREM 1. *Under normality assumption the minimax risk of estimating the precision matrix Ω over the parameter space $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ given in (2) under the operator norm satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}. \tag{4}$$

The minimax risk of estimating the precision matrix Ω over the parameter space $\mathcal{Q}_\alpha(\eta, M)$ with $\alpha > 0$ given in (3) under the operator norm satisfies

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}. \tag{5}$$

Moreover, we also consider the minimax rates of convergence of precision matrix estimation under the Frobenius norm loss over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. This time, we prove that two types of spaces enjoy the same optimal rate of convergence. Together with the different rates of convergence under the operator norm loss, we demonstrate the intrinsic difference between operator norm and Frobenius norm. The Frobenius norm of a p by p matrix is defined as the ℓ_2 vector norm of all entries. Driven by this fact, our estimation approach is naturally obtained by optimally estimating A and D in (1) separately. Due to the decay structure in $\mathcal{P}_\alpha(\eta, M)$, which is defined in terms of nested ℓ_1 norm of each row of A, our estimator is based on regression with a delicate thresholding rule. The minimax procedure is motivated by wavelet nonparametric function estimation, although the space $\mathcal{P}_\alpha(\eta, M)$ cannot be exactly described by any Besov ball [11, 18]. We summarize the optimality result under the Frobenius norm in Theorem 2 below.

THEOREM 2. *Under normality assumption the minimax risk of estimating the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ given in* (2) *and* (3) *satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \asymp n^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{6}$$

Related literature. During the last decade various structural assumptions are imposed in literature of high-dimensional statistics in order to estimate the covariance/precision matrix consistently under various loss functions. While mostly driven by the specific scientific applications, popular structures include *ordered sparsity* (bandable covariance matrices, precision matrices with bandable Cholesky factor), *unordered sparsity* (sparse covariance matrices, sparse precision matrices) and other more complicated ones such as certain combination of sparsity and low rankness (spike covariance matrices, covariance with tensor product, latent graphical models). Many estimation procedures have been proposed accordingly to estimate high-dimensional covariance/precision matrices via taking advantages of these specific structures. For example, banding [7, 8, 54, 55] and tapering methods [15, 23] were developed to estimate bandable covariance matrices or precision matrices with bandable Cholesky factor; thresholding procedures were used in [6, 9, 21] to estimate sparse covariance matrices; penalized likelihood estimation [2, 17, 27, 32, 44, 46, 58] and penalized regression methods [10, 39, 45, 50, 57] are designed for sparse precision matrix estimation.

The fundamental difficulty of various covariance/precision matrices estimation problems have been carefully investigated in terms of the minimax risks under the operator norm loss among other losses, especially for those *ordered and unordered sparsity structures*. Specifically, for unordered structures, [16] considered the problems of optimal estimation of sparse covariance while [12] (see [45] as well) established the optimality results for estimating sparse precision matrices. For ordered structures [15] established the optimal rates of convergence over two types of bandable covariance matrices. In addition, with an extra Toeplitz structure [13] studied optimal estimation of two types of bandable Toeplitz covariance matrices. However, it was still largely unknown about the optimality results on estimating precision matrices with bandable Cholesky factor. See an exposure paper with discussion [14] and references therein on minimax results of covariance/precision matrix estimation under some other losses. In this paper we provide a solution to this open problem by establishing the optimal rates of convergence over two types of precision matrices with bandable Cholesky factor. *Thus, this paper completes the minimaxity results of all four sparsity structures commonly considered in literature*.

Organization of the paper. The rest of the paper is organized as follows. First, we propose our estimation procedures for precision matrix estimation in Section 2. In particular, local cropping estimators and regression-based thresholding estimators are designed for estimating precision matrices under the operator norm and the Frobenius norm respectively. Section 3 establishes the optimal rates of convergence under the operator norm for two commonly used types of parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. A striking difference between two spaces are revealed when considering operator norm loss. Section 4 considers rate-optimal estimation under the Frobenius norm. The results reveal that the fundamental difficulty of estimation for two parameter spaces are the same when considering Frobenius norm loss. Section 5 considers the adaptive estimation through a variation of Lepski's method under the operator norm. In Section 6 we extend the results to nonparanormal models for inverse correlation matrix estimation by applying local cropping procedure to rank-based estimators. Section 7 presents the numerical performance of our local cropping procedure to illustrate the difference between two parameter spaces by simulation studies. We also demonstrate the suboptimality of banding estimators. Discussion and all technical lemmas used in proofs of main results are relegated to the Supplementary Material [38].

Notation. We introduce some basic notations that will be used in the rest of the paper. $\mathbf{1}(\cdot)$ indicates the indicator function while $\mathbf{1}$ indicates the all-ones vector. $\mathsf{sgn}(\cdot)$ indicates the sign function. $\lfloor s \rfloor$ represents the largest integer which is no more than s. $\lceil s \rceil$ represents the smallest integer which is no less than s. Define $a_n \asymp b_n$ if there is a constant $C > 0$ independent of n such that $C^{-1} \le a_n/b_n \le C$. For any vector x, $\|x\|_p$ indicates its ℓ_p norm. For any p by q matrix $S = [s_{ij}]_{p\times q} \in \mathbb{R}^{p\times q}$, we use S^T to denote its transpose. The ℓ_p matrix norm is defined as $\|S\|_p = \sup_{\|x\|_p=1} \|Sx\|_p$. The ℓ_2 matrix norm is also called the operator norm or the spectral norm and denoted as $\|S\|_{\mathrm{op}}$. The Frobenius norm is defined as $\|S\|_{\mathrm{F}} = (\sum_{i,j} s_{ij}^2)^{\frac{1}{2}}$. $\lambda_{\max}(S)$ and $\lambda_{\min}(S)$ are the largest and smallest singular values of S when S is not symmetric. When S is a real symmetric matrix, $\lambda_{\max}(S)$ and $\lambda_{\min}(S)$ denote its largest and smallest eigenvalues. $\mathrm{row}_i(S)$ and $\mathrm{col}_i(S)$ indicate the ith row and column of matrix S. $a : b$ denotes the index set $\{a, a+1, \ldots, b\}$. $[p]$ is short for the set $1 : p$. For the random vector $\mathbf{X} \in \mathbb{R}^{p\times 1}$ and the data matrix $\mathbf{Z} \in \mathbb{R}^{n\times p}$, $\mathbf{X}_{a:b}$ and $\mathbf{Z}_{a:b}$ indicate the $(a : b)$-th columns of $\mathbf{X}^T$ and $\mathbf{Z}$. For any square matrix S, $\mathrm{diag}(S)$ denotes the diagonal matrix with diagonal entries being those on the main diagonal of S while, for any vector v, $\mathrm{diag}(v)$ denotes the diagonal matrix with diagonal entries being v. In the estimation procedure under the operator norm, we use the matrix notation in the form of $S_m^{(k)}$ to facilitate the proof; where S is always a square matrix, m indicates the location information and (k) indicates that the size of $S_m^{(k)}$ is k. Throughout the paper we denote by C a generic positive constant which may vary from place to place but only depends on α, η, M and, possibly, some sub-Gaussian distribution constant ρ in (17).

2. Methodologies. In this section we introduce our methodologies for estimating precision matrices over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ under both the operator norm and the Frobenius norm. Assume that $\mathbf{X} = (X_1, \ldots, X_p)^T$, a p-variate random vector with mean zero and precision matrix Ω_p. Our estimation procedures are based on its n i.i.d. copies $\mathbf{Z} \in \mathbb{R}^{n\times p}$. We write $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$, where each $\mathbf{Z}_i$ consists of n i.i.d. copies of X_i. Our estimation procedures are different under the operator norm and the Frobenius norm.

2.1. *Estimation procedure under the operator norm.* We focus on the estimation problem under the operator norm first. As we discussed in the Introduction, almost all existing methodologies [7, 27, 53] directly appeal the Cholesky decomposition of the precision matrix. They first estimate the Cholesky factor A and D by autoregression and then estimate

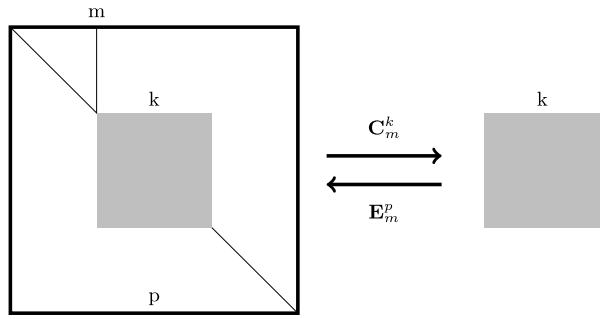

FIG. 1. *An illustration of the cropping operator and the expanding operator.*

the precision matrix according to $\Omega = (I - A)^T D^{-1}(I - A)$. The corresponding analysis in the row-wise fashion may not suitable for the operator norm loss. In this paper we propose a novel local cropping estimator which focuses on the estimation of Ω directly.

To facilitate the illustration of the estimation procedure, we define two matrix operators. The *cropping operator* is designed to crop the center block out of the matrix. For a p by p matrix $E \equiv [e_{ij}]_{p\times p}$, we define the $k \times k$ matrix $\mathbf{C}_m^k(E) \equiv [c_{ij}]_{k\times k}$, where $1 \le m \le p-k+1$, with

$$c_{ij} = e_{i+m-1,\, j+m-1}, \quad \text{when } 1 \le i, j \le k. \tag{7}$$

The parameter m indicates the location and k indicates the dimension. It is clear that $\mathbf{C}_m^k(E)$ is a principal submatrix of E. The *expanding operator* is designed to put a small matrix onto a large zero matrix. For a k by k matrix, $C \equiv [c_{ij}]_{k\times k}$, define the $p \times p$ matrix $\mathbf{E}_m^p(C) \equiv [e_{ij}]_{p\times p}$, where $1 \le m \le p - k + 1$, with

$$e_{ij} = c_{i-m+1,\, j-m+1}, \quad \text{when } m \le i, j \le m + k - 1, \quad \text{otherwise } e_{ij} = 0. \tag{8}$$

The parameter m indicates the location and p indicates the dimension. Note that for a k by k matrix C, we have $\mathbf{C}_m^k(\mathbf{E}_m^p(C)) = C$. An illustration of two operators is provided in Figure 1.

In addition, for technical reasons (of obtaining rates of convergence in expectation rather than in probability), we introduce a *projection operator*. For a real square matrix S, let the singular value decomposition of S be $S = U\Lambda V^T$ with $UU^T = I$, $VV^T = I$ and $\Lambda = \operatorname{diag}(\lambda_i)$. Let $\Lambda^* = \operatorname{diag}(\lambda_i^*)$, where $\lambda_i^* = \min\{\max\{\lambda_i, \eta^{-1}\}, \eta\}$, then define

$$\mathbf{P}_\eta(S) = U\Lambda^* V^T. \tag{9}$$

For a symmetric matrix S, we modify $\mathbf{P}_\eta(\cdot)$ a little bit and define $\mathbf{P}_\eta(S) = U\Lambda^* U^T$, where $S = U\Lambda U^T$ is its eigendecomposition. Since all eigenvalues of $\mathbf{P}_\eta(\cdot)$ are in the interval $[\eta^{-1}, \eta]$, $\mathbf{P}_\eta(S)$ is always invertible and positive definite.

We are ready to construct the local cropping estimator $\tilde{\Omega}_k^{\mathrm{op}}$ with bandwidth $k < p$. At a high level we first propose an estimator of each principal submatrix of size k and $2k$ in Ω using cropping and expanding operators. Then, we arrange over those local estimators to estimate Ω. Since the core idea of estimating those local estimators in our procedure is to crop the inverse of sample covariance matrix with a relatively larger size, we call $\tilde{\Omega}_k^{\mathrm{op}}$ in (12) *the local cropping estimator*.

Specifically, we first define an estimator $\tilde{\Omega}_m^{(k)}$ of the principal submatrix $\mathbf{C}_m^k(\Omega)$ at each location m. To this end, we select the sample covariance matrix with a relative larger size, in this case, $3k$. Let the modified local sample covariance matrix be

$$\tilde{\Sigma}_{m-k}^{(3k)} = \mathbf{P}_\eta\left(\mathbf{C}_{m-k}^{3k}\left(\frac{1}{n}\mathbf{Z}^T\mathbf{Z}\right)\right), \quad 2 - k \le m \le p. \tag{10}$$

We refer to Remark 1 for the treatment when the index is beyond the index set $[p]$. Note that the operator $\mathbf{P}_\eta(\cdot)$ guarantees $\tilde{\Sigma}_{m-k}^{(3k)}$ to be invertible. Then, we use the center part of its inverse to estimate $\mathbf{C}_m^k(\Omega)$, that is,

$$\tilde{\Omega}_m^{(k)} = \mathbf{C}_{k+1}^k\big((\tilde{\Sigma}_{m-k}^{(3k)})^{-1}\big). \tag{11}$$

Similarly, we can define local estimators of $\tilde{\Omega}_m^{(2k)}$ via replacing k by $2k$. Arranging over these estimators in the form of weighted sum, we obtain the estimator of Ω, that is,

$$\tilde{\Omega}_k^{\mathrm{op}} = \mathbf{P}_\eta\left(\frac{1}{k}\left(\sum_{m=2-2k}^{p} \mathbf{E}_m^p\big(\tilde{\Omega}_m^{(2k)}\big) - \sum_{m=2-k}^{p} \mathbf{E}_m^p\big(\tilde{\Omega}_m^{(k)}\big)\right)\right). \tag{12}$$

The operator $\mathbf{E}_m^p(\cdot)$ makes these local estimators in the correct places. The final step (12) is motivated by the analysis of optimal bandable covariance matrix estimation procedure proposed in [15]. Indeed, the optimal tapering estimator in [15] can be rewritten as a sum of many principal submatrices of the sample covariance matrix in a similar way as (12). In contrast, our estimator is not in a form of tapering the sample covariance matrix. However, in the analysis of our local cropping estimator in Section 3, the direct target of $\tilde{\Omega}_k^{\mathrm{op}}$ is a certain tapered population precision matrix with bandwidth k. There are natural bias and variance terms involved in the distance of $\tilde{\Omega}_k^{\mathrm{op}}$ and its direct target. Together with the bias of the tapered population precision matrix, our analysis involves two bias terms and one variance term which critically determine the optimal choice of bandwidth. It is worthwhile to mention that, although the local estimators in (12) overlap with each other significantly, the variance term is not influenced too much by the overlap due to a technique of rearranging all local estimators in the analysis. Please refer to the proof of Theorem 3 for further details.

In Section 3 we show that the local cropping estimator with an optimal choice of bandwidth would achieve the minimax risk under the operator norm over parameter spaces $\mathcal{P}_\alpha(\eta, M)$ in (2) and $\mathcal{Q}_\alpha(\eta, M)$ in (3). However, the optimal choices of bandwidth are fundamentally distinct between $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. Specifically, we show that the optimal bandwidth over $\mathcal{P}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha}}$ while that one over $\mathcal{Q}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha+1}}$.

REMARK 1. Of note, the estimator $\tilde{\Omega}_k^{\mathrm{op}}$ depends on $\mathbf{Z}_{2-4k}, \ldots, \mathbf{Z}_{p+4k-1}$. The index of variable is clear most of the time, while we need to be careful when it is close to the boundary. When the index is beyond the index set $[p]$, we shrink the size of the corresponding block by discarding the data with meaningless indexes. It can be shown that this shrinking operation would not change the theoretical properties of the final estimator.

2.2. *Estimation procedure under the Frobenius norm.* Under the Frobenius norm our estimation procedure is based on the Cholesky decomposition of the precision matrix (1). More specifically, we estimate the matrix A and D respectively by autoregression and then combine them to construct the estimator of Ω. The following estimation procedure applies to both the parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, as we will show that they enjoy the same optimal rate of convergence in Section 4.

Our estimator of the ith row of A is based on the regression of X_i against its previous variables. Unlike those existing methods [7, 27, 53] which rely on certain banding or penalized approaches for such a regression problem, we apply a thresholding procedure due to the decay structure in $\mathcal{P}_\alpha(\eta, M)$ which is defined in terms of nested ℓ_1 norm. To this end, we first regress X_i against $\mathbf{X}_{i-k_1:i-1} = (X_{i-k_1}, \ldots, X_{i-1})^T$ with bandwidth $k_1 = \lceil \frac{n}{c} \rceil$ with some sufficiently large $c > 0$. Recall that the $n \times 1$ matrix $\mathbf{Z}_i$ consists of n observations of X_i, and

the $n \times k_1$ matrix $\mathbf{Z}_{i-k_1:i-1}$ represents n observations of $\mathbf{X}_{i-k_1:i-1}$. The empirical regression coefficients are

$$(\hat{a}_{i(i-k_1)}, \ldots, \hat{a}_{i(i-1)})^T = \big(\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1}\big)^{-1} \mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_i. \tag{13}$$

We then further threshold the coefficients by taking advantages of the bandable structure of the Cholesky factor A. Specifically, we define $\hat{\mathbf{a}}_i^* \in \mathbb{R}^{i-1}$ with coordinate $\hat{a}_{ij}^*$ as follows:

$$\hat{a}_{ij}^* = \begin{cases} \hat{a}_{ij}, & \text{if } i - k_0 < j \le i-1, \\ \hat{a}_{ij} \mathbf{1}\big(|\hat{a}_{ij}| > \lambda_j\big), & \text{if } i - k_1 < j \le i - k_0, \\ 0, & \text{if } 1 \le j \le i - k_1, \end{cases} \tag{14}$$

where $k_0 = \lceil n^{\frac{1}{2\alpha+2}} \rceil$, the threshold level $\lambda_j = (\lceil \log_2^{i-j} - \log_2^{k_0} \rceil R)^{\frac{1}{2}}$ and $R = 8\eta \|(\mathbf{Z}_{i-k_1:i-1}^T \times \mathbf{Z}_{i-k_1:i-1})^{-1}\|_{\mathrm{op}}$. Note that we keep the last k_0 coefficients and apply an entrywise thresholding rule for which the thresholding level remains the same in each block and the size of each block doubles backward sequentially for the remaining coefficients in (14). Our procedure is inspired by the optimal estimation procedure over Besov balls for many nonparametric function estimation problems or, equivalently, the corresponding Gaussian sequence models (see [11] the reference therein). We emphasize that any linear estimator of the coefficients $(\hat{a}_{i(i-k_1)}, \ldots, \hat{a}_{i(i-1)})^T$ cannot yield to the optimal rates of convergence in our setting under the Frobenius norm.

Our estimator of $I - A$ can be constructed by arranging zero-padded $\hat{\mathbf{a}}_i^{*T}$, $i \in [p]$ accordingly with an identity matrix. Specifically, set the ijth entry of $\hat{A}^*$ as $\hat{a}_{ij}^*$ when $i \in [p]$, $j \in [i-1]$, otherwise as zero. We also need to bound the singular values of $(I - \hat{A}^*)$. To this end, we define

$$\widetilde{I - A} = \mathbf{P}_\eta\big(I - \hat{A}^*\big)$$

as our estimator of $(I - A)$, where $\mathbf{P}_\eta(\cdot)$ is defined in (9).

The estimation of D is based on the sample variances of those empirical residuals in the regression of X_i against $\mathbf{X}_{i-k_1:i-1} = (X_{i-k_1}, \ldots, X_{i-1})^T$. For each i, the sample variance of the empirical residual is

$$\hat{d}_i = \frac{1}{n - k_1} \mathbf{Z}_i^T (I - \mathbf{M}_i)^T (I - \mathbf{M}_i) \mathbf{Z}_i, \tag{15}$$

where $\mathbf{M}_i = \mathbf{Z}_{i-k_1:i-1} (\mathbf{Z}_{i-k_1:i-1}^T \mathbf{Z}_{i-k_1:i-1})^{-1} \mathbf{Z}_{i-k_1:i-1}^T$. Let $\hat{D} = \mathrm{diag}(\hat{d})$, where $\hat{d} = (\hat{d}_1, \ldots, \hat{d}_p)^T$. We define $\tilde{D} = \mathbf{P}_\eta(\hat{D})$ as our estimator of D.

Finally, define our estimator of Ω as

$$\tilde{\Omega}_k^{\mathrm{F}} = (\widetilde{I - A})^T \tilde{D}^{-1} (\widetilde{I - A}). \tag{16}$$

REMARK 2. For the parameter space $\mathcal{Q}_\alpha(\eta, M)$, a much simpler banding estimation scheme on the empirical regression coefficients is able to achieve the minimax risk. Set $k = \lceil n^{\frac{1}{2\alpha+2}} \rceil$. We use the empirical residuals and coefficients obtained by regressing each X_i against $\mathbf{X}_{i-k:i-1}$ to directly construct the estimators of A and D. It can be proved that this estimator achieves the minimax risk over the parameter space $\mathcal{Q}_\alpha(\eta, M)$.

3. Rate optimality under the operator norm. In this section we establish the optimal rates of convergence for estimating the precision matrix over the parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ given in (2) and (3) under the operator norm. We first derive the risk upper bound of the local cropping estimator in Section 3.1 over parameter space $\mathcal{P}_\alpha(\eta, M)$. We provide a matching risk lower bound by applying Assouad's lemma and Le Cam's method in Section 3.2 over $\mathcal{P}_\alpha(\eta, M)$. The establishment of the rate optimality over the parameter space $\mathcal{Q}_\alpha(\eta, M)$ is similar to the one over $\mathcal{P}_\alpha(\eta, M)$ which is provided in Section 3.3.

Throughout this section we assume that $\mathbf{X} = (X_1, \ldots, X_p)^T$ follows certain sub-Gaussian distribution with constant $\rho > 0$, that is,

$$\mathbb{P}\{|v^T(\mathbf{X} - \mathbb{E}\mathbf{X})| > t\} \le 2\exp(-t^2/(2\rho)), \tag{17}$$

for all $t > 0$ and all unit vectors $\|v\|_2 = 1$.

REMARK 3. The sub-Gaussian distribution is often assumed in high-dimensional statistical problems. In our settings this assumption is critical to derive the exponential-type concentration inequality for the quadratic terms of $\mathbf{X}$. When only certain moment conditions are posed, one can replace each local estimator in (11) by a Huber-type estimator proposed in [30, 40]. We leave the theoretical investigation in future works.

3.1. *Minimax upper bound under the operator norm over $\mathcal{P}_\alpha(\eta, M)$.* In this section we develop the following upper bound of our estimation procedure proposed in Section 2.1.

THEOREM 3. *When $\lceil n^{\frac{1}{2\alpha}} \rceil \le p$, the local cropping estimator defined in (12) of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$ given in (2) satisfies*

$$\sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le Ck^{-2\alpha+1} + C\frac{\log p + k}{n}.$$

When $k = \lceil n^{\frac{1}{2\alpha}} \rceil$, we have

$$\sup_{\mathcal{P}_\alpha(\eta, M)} \mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}.$$

The optimal choice of $k \asymp n^{\frac{1}{2\alpha}}$ is due to the bias-variance tradeoff. Combining Theorem 3 with the minimax lower bound derived in Section 3.2, we immediately obtain that the local cropping estimator is rate optimal.

PROOF. As we discussed in Section 2.1, the direct target of our local cropping estimator is certain tapered population precision matrix with bandwidth k which can be written as a weighted sum of many principal submatrices of the population precision matrix. We construct this corresponding tapered population precision matrix Ω_k^* as follows. Denote the precision matrix $\Omega \equiv [\omega_{ij}]_{p\times p}$. We define $\Omega_k^* \equiv [\omega_{ij}^*]_{p\times p}$ such that for $i, j \in [p]$,

$$\begin{aligned} &\omega_{ij}^* = m_{ij}\omega_{ij}, \\ &\text{where } m_{ij} = \max\left\{0, 2 - \frac{1}{k}|i-j|\right\} - \max\left\{0, 1 - \frac{1}{k}|i-j|\right\}. \end{aligned} \tag{18}$$

The following lemma elucidates the decomposition of this tapered precision matrix Ω_k^*.

LEMMA 1. *The Ω_k^* defined in (18) can be written as*

$$\Omega_k^* = \frac{1}{k}\left(\sum_{m=2}^{2k+1}\left(\sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}_{m+2kj}^{p}(\mathbf{C}_{m+2kj}^{2k}(\Omega))\right) - \sum_{m=2}^{k+1}\left(\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\mathbf{C}_{m+2kj}^{k}(\Omega))\right)\right).$$

The proof of Lemma 1 can be found in [15] (refer to the proof of Lemma 1 with covariance matrix therein replaced by the precision matrix) and thus omitted. Define

$$\tilde{\Omega}_k^* = \frac{1}{k}\left(\sum_{m=2}^{2k+1}\left(\sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}_{m+2kj}^{p}(\tilde{\Omega}_{m+2kj}^{(2k)})\right) - \sum_{m=2}^{k+1}\left(\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\tilde{\Omega}_{m+kj}^{(k)})\right)\right).$$

It is easy to check $\tilde{\Omega}_k^{\mathrm{op}} = \mathbf{P}_\eta(\tilde{\Omega}_k^*)$. Since the eigenvalues of Ω are in the interval $[\eta^{-1}, \eta]$, the operator $\mathbf{P}_\eta(\cdot)$ would not increase the risk much. Indeed, according to (B.1) in Lemma B.1 of [38], we have

$$\begin{aligned}\mathbb{E}\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 &\le 4\mathbb{E}\|\tilde{\Omega}_k^* - \Omega\|_{\mathrm{op}}^2 \\ &\le 8\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\mathrm{op}}^2 + 8\|\Omega_k^* - \Omega\|_{\mathrm{op}}^2.\end{aligned} \tag{19}$$

The following lemma bounds the bias between our direct target Ω_k^* and the population precision matrix.

LEMMA 2. *For Ω in the parameter space $\mathcal{P}_\alpha(\eta, M)$ defined in (2) with $\alpha > \frac{1}{2}$, Ω_k^* defined in (18), we have*

$$\|\Omega_k^* - \Omega\|_{\mathrm{op}}^2 \le Ck^{-2\alpha+1}.$$

REMARK 4. Unlike existing methods, our procedure does not directly utilize the decay structure of Cholesky factor. Consequently, the proof of Lemma 2 is involved and requires a blockwise partition strategy.

Then, we turn to the analysis of $\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\mathrm{op}}^2$:

$$\begin{aligned}&\mathbb{E}\|\tilde{\Omega}_k^* - \Omega_k^*\|_{\mathrm{op}}^2 \\ &\quad\le 2\mathbb{E}\left(\frac{1}{k}\sum_{m=2}^{2k+1}\left\|\sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}_{m+2kj}^{p}(\tilde{\Omega}_{m+2kj}^{(2k)}) - \sum_{j=-1}^{\lfloor p/2k\rfloor}\mathbf{E}_{m+2kj}^{p}(\mathbf{C}_{m+2kj}^{2k}(\Omega))\right\|_{\mathrm{op}}\right)^2 \\ &\qquad + 2\mathbb{E}\left(\frac{1}{k}\sum_{m=2}^{k+1}\left\|\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\tilde{\Omega}_{m+kj}^{(k)}) - \sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\mathbf{C}_{m+kj}^{k}(\Omega))\right\|_{\mathrm{op}}\right)^2.\end{aligned} \tag{20}$$

These two terms can be bounded in the same way; we only focus on the second term:

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{k}\sum_{m=2}^{k+1}\left\|\sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\tilde{\Omega}_{m+kj}^{(k)}) - \sum_{j=-1}^{\lfloor p/k\rfloor}\mathbf{E}_{m+kj}^{p}(\mathbf{C}_{m+kj}^{k}(\Omega))\right\|_{\mathrm{op}}\right)^2 \\ &\quad\le \mathbb{E}\left(\max_m\left\|\sum_{j=-1}^{\lfloor p/k\rfloor}\left(\mathbf{E}_{m+kj}^{p}(\tilde{\Omega}_{m+kj}^{(k)}) - \mathbf{E}_{m+kj}^{p}(\mathbf{C}_{m+kj}^{k}(\Omega))\right)\right\|_{\mathrm{op}}^2\right)\end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{E}\Big(\max_{m,j}\big\|\tilde{\Omega}^{(k)}_{m+kj} - \mathbf{C}^{k}_{m+kj}(\Omega)\big\|^2_{\mathrm{op}}\Big) \\
&\le 2\mathbb{E}\Big(\max_{m\in[p]}\big\|\tilde{\Omega}^{(k)}_{m} - \mathbf{C}^{k}_{k+1}((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1})\big\|^2_{\mathrm{op}}\Big) \\
&\quad + 2\Big(\max_{m\in[p]}\big\|\mathbf{C}^{k}_{k+1}((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}) - \mathbf{C}^{k}_{m}(\Omega)\big\|^2_{\mathrm{op}}\Big),
\end{aligned} \tag{21}
$$

where we further have variance term and bias term of local estimators. For the variance term in (21) we further have

$$
\begin{aligned}
&\big\|\tilde{\Omega}^{(k)}_{m} - \mathbf{C}^{k}_{k+1}((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1})\big\|_{\mathrm{op}} \\
&\quad\le \big\|(\tilde{\Sigma}^{(3k)}_{m-k})^{-1} - (\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}\big\|_{\mathrm{op}} \\
&\quad\le \eta^2\big\|\tilde{\Sigma}^{(3k)}_{m-k} - \mathbf{C}^{3k}_{m-k}(\Omega^{-1})\big\|_{\mathrm{op}} \\
&\quad\le 2\eta^2\left\|\mathbf{C}^{3k}_{m-k}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T\right) - \mathbf{C}^{3k}_{m-k}(\Omega^{-1})\right\|_{\mathrm{op}}.
\end{aligned} \tag{22}
$$

The last two inequalities hold because of the fact that the eigenvalues of $\tilde{\Sigma}^{(3k)}_{m-k}$ and $\mathbf{C}^{3k}_{m-k}(\Omega^{-1})$ are in the interval $[\eta^{-1}, \eta]$, and Lemma B.1 of [38]. The following concentration inequality of sample covariance matrix facilitates our proof.

LEMMA 3. *For the observations* $\mathbf{Z}$ *following certain sub-Gaussian distribution with constant* ρ *and precision matrix* Ω, *we have*

$$
\mathbb{E}\left(\max_{m\in[p]}\left\|\mathbf{C}^{3k}_{m-k}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^T\right) - \mathbf{C}^{3k}_{m-k}(\Omega^{-1})\right\|^2_{\mathrm{op}}\right) \le C\frac{\log p + k}{n}.
$$

Lemma 3 is an extension of the result in Chapter 2 of [47]. Its proof can be found in Lemma 3 of [15].

Combining Lemma 3, (21) and (22), we have

$$
\mathbb{E}\Big(\max_{m\in[p]}\big\|\tilde{\Omega}^{(k)}_{m} - \mathbf{C}^{k}_{k+1}((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1})\big\|^2_{\mathrm{op}}\Big) \le C\frac{\log p + k}{n}. \tag{23}
$$

We turn to bounding the bias term of local estimator in (21).

LEMMA 4. *Assume that* $\Omega \in \mathcal{P}_\alpha(\eta, M)$ *defined in* (2) *with* $\alpha > \frac{1}{2}$. *Then, we have*

$$
\big\|\mathbf{C}^{k}_{k+1}((\mathbf{C}^{3k}_{m-k}(\Omega^{-1}))^{-1}) - \mathbf{C}^{k}_{m}(\Omega)\big\|^2_{\mathrm{op}} \le Ck^{-2\alpha+1}.
$$

Lemma 4, together with (23), (21) and (20), implies that

$$
\mathbb{E}\|\tilde{\Omega}^{*}_{k} - \Omega^{*}_{k}\|^2_{\mathrm{op}} \le C\frac{\log p + k}{n} + Ck^{-2\alpha+1}. \tag{24}
$$

Plugging Lemma 2 and (24) into (19), we finish the proof of Theorem 3. □

3.2. *Minimax lower bound under the operator norm over* $\mathcal{P}_\alpha(\eta, M)$. Theorem 3 in Section 3.1 proves that the local cropping estimator defined in (12) attains the convergence rate of $n^{\frac{-2\alpha+1}{2\alpha}} + \frac{\log p}{n}$. In this section we establish the following matching lower bound which proves the rate optimality of the local cropping estimator.

THEOREM 4. *The minimax risk of estimating the precision matrix* Ω *over* $\mathcal{P}_\alpha(\eta, M)$ *defined in* (2) *under the operator norm with* $\alpha \geq \frac{1}{2}$ *satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{32}\left(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}\right), \tag{25}$$

where $0 < \tau < \min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

REMARK 5. Theorems 3 and 4 together show the minimax risk for estimating the precision matrices over $\mathcal{P}_\alpha(\eta, M)$ stated in (4) of Theorem 1. It is worthwhile to notice that there is no consistent estimator over $\mathcal{P}_\alpha(\eta, M)$ under the operator norm, when $\alpha \leq \frac{1}{2}$.

PROOF. The lower bound of parameter space $\mathcal{P}_\alpha(\eta, M)$ can be established by the lower bounds over its subsets. We construct two subsets, $\mathcal{P}_1$ and $\mathcal{P}_2$, and calculate the lower bound over those two subsets separately. Let τ be a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

First, we construct $\mathcal{P}_1$. Set $k = \min\{\lceil n^{\frac{1}{2\alpha}} \rceil, \frac{p}{2}\}$. Set the index set $\Theta = \{0, 1\}^k$, that is, for any $\theta \equiv \{\theta_i\}_{1 \leq i \leq k} \in \Theta$, each θ_i is either 0 or 1. Then, we define the $k \times k$ matrix $A_k^*(\theta) \equiv [a_{ij}]_{k\times k}$ with $a_{ij} = \tau n^{-\frac{1}{2}} \theta_i \mathbf{1}(j = k)$ and

$$A(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ A_k^*(\theta) & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)\times(p-2k)} \end{bmatrix}.$$

We then define $\mathcal{P}_1$ as the collection of 2^k matrices indexed by Θ,

$$\mathcal{P}_1 = \big\{\Omega(\theta) : \Omega(\theta) = \big(I_p - A(\theta)\big)^T \big(I_p - A(\theta)\big), \theta \in \Theta\big\}. \tag{26}$$

Next, we construct $\mathcal{P}_2$ as the collection of the diagonal matrices in the following equation:

$$\begin{aligned} \mathcal{P}_2 = \big\{&\Omega(m) \equiv \big[w_{ij}(m)\big]_{p\times p} : \\ &w_{ij}(m) = \big(\mathbf{1}(i = j) + \tau a^{\frac{1}{2}} \mathbf{1}(i = j = m)\big)^{-1}, m \in 0 : p\big\}, \end{aligned} \tag{27}$$

where $a = \min\{\frac{\log p}{n}, 1\}$.

LEMMA 5. $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *are subsets of* $\mathcal{P}_\alpha(\eta, M)$.

Note that we assume $\log p = O(n)$ and $n = O(p)$. Without loss of generality, we further assume $\log p < n < p$. For any estimator $\tilde{\Omega}$ based on n i.i.d. observations, we establish the lower bounds over those two subsets in Sections 3.2.1 and 3.2.2, respectively,

$$\sup_{\mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1} \min\left\{n^{\frac{1}{2\alpha}}, \frac{p}{2}\right\} \geq \frac{\tau^2}{16} n^{-\frac{2\alpha-1}{2\alpha}}, \tag{28}$$

$$\sup_{\mathcal{P}_2} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \geq \frac{\tau^2}{16} n^{-1} \min\{\log p, n\} \geq \frac{\tau^2}{16} \frac{\log p}{n}. \tag{29}$$

According to Lemma 5, $(\mathcal{P}_1 \cup \mathcal{P}_2) \subset \mathcal{P}_\alpha(\eta, M)$. Therefore, we obtain

$$\begin{aligned} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 &\geq \max\Big\{\sup_{\mathcal{P}_1} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2, \sup_{\mathcal{P}_2} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2\Big\} \\ &\geq \frac{\tau^2}{32}\left(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}\right) \end{aligned}$$

which completes the proof of Theorem 4. □

We introduce some further notation before establishing (28) using Assouad's lemma in Section 3.2.1 and (29) using Le Cam's method in Section 3.2.2. Let $H(\theta,\theta')=\sum_{i=1}^{k}|\theta_i-\theta_i'|$ be the Hamming distance on $\{0,1\}^k$ which is the number of different elements between θ and θ'. The total variation affinity $\|P\wedge Q\|=\int p\wedge q\ d\mu$, where p and q are the density functions of two probability measure P and Q with respect to any common dominating measure μ.

3.2.1. *Assouad's lemma in proof of* (28). Assouad's lemma [1] is a powerful tool to provide the lower bound over distributions indexed by the hypercube $\Theta=\{0,1\}^k$. Let P_θ be the distribution generated from observations indexed by $\Omega(\theta)$. The proof of Lemma 6 can be found in [56] and thus omitted.

LEMMA 6 (Assouad). *Let $\tilde{\Omega}$ be an estimator based on observations from a distribution in the collection $\{P_\theta,\theta\in\Theta\}$, where $\Theta=\{0,1\}^k$. Then,*

$$\sup_{\theta\in\Theta}2^2\mathbb{E}_\theta\|\tilde{\Omega}-\Omega(\theta)\|_2^2\geq\min_{H(\theta,\theta')\geq1}\frac{\|\Omega(\theta)-\Omega(\theta')\|_2^2}{H(\theta,\theta')}\frac{k}{2}\min_{H(\theta,\theta')=1}\|P_\theta\wedge P_{\theta'}\|.$$

Applying Assouad's lemma to the subset $\mathcal{P}_1$, we have the following results.

LEMMA 7. *Let P_θ be the joint distribution of n i.i.d. observations from $N(0,\Omega(\theta)^{-1})$, where $\Omega(\theta)\in\mathcal{P}_1$ defined in (26). Then,*

$$\min_{H(\theta,\theta')=1}\|P_\theta\wedge P_{\theta'}\|\geq0.5.$$

LEMMA 8. *Consider all $\Omega(\theta)\in\mathcal{P}_1$ defined in (26). Then,*

$$\min_{H(\theta,\theta')\geq1}\frac{\|\Omega(\theta)-\Omega(\theta')\|_2^2}{H(\theta,\theta')}\geq\big(\tau n^{-\frac{1}{2}}\big)^2.$$

Lemmas 6, 7 and 8 together imply the desired (28), with the choice $k=\lceil n^{\frac{1}{2\alpha}}\rceil$. The proofs of the above lemmas can be found in the Supplementary Material [38].

3.2.2. *Le Cam's method in proof of* (29). Le Cam's method can be used to establish the lower bound via testing a single distribution against a convex hull of distributions. Set $r=\inf_{m\in[p]}\|\Omega(0)-\Omega(m)\|_{\mathrm{op}}^2$. Let P_i be the distribution generated from observations indexed by $\Omega(i)$, where $0\leq i\leq p$. Define $\bar{P}=\sum_{m=1}^{p}P_m$. The proof of the following lemma can be found in [56] and thus omitted.

LEMMA 9 (Le Cam). *Let $\tilde{\Omega}$ be an estimator based on observations from a distribution in the collection $\{P_i,0\leq i\leq p\}$. Then,*

$$\sup_{0\leq m\leq p}\mathbb{E}\|\tilde{\Omega}-\Omega(m)\|_{\mathrm{op}}^2\geq\frac{1}{2}r\|P_0\wedge\bar{P}\|.$$

Applying Le Cam's method to $\mathcal{P}_2$, we obtain that $r=(\frac{\tau a^{\frac{1}{2}}}{1+\tau a^{\frac{1}{2}}})^2\geq\frac{1}{4}\tau^2a$ and the following results.

LEMMA 10. *Let P_m be the joint distribution of n i.i.d. observations from $N(0,\Omega(m)^{-1})$, where $\Omega(m)\in\mathcal{P}_2$ defined in (27). Then,*

$$\|P_0\wedge\bar{P}\|>\frac{7}{8}.$$

Combining the above results in Lemmas 9 and 10, we obtain the desired (29), that is,

$$\sup_{0\le m\le p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_{\mathrm{op}}^2\ge\frac{7}{64}\tau^2 a\ge\frac{\tau^2}{16}\min\biggl\{\frac{\log p}{n},1\biggr\}.$$

3.3. *Rate optimality under the operator norm over* $\mathcal{Q}_\alpha(\eta,M)$.

3.3.1. *Minimax upper bound.* In this section we establish the upper bound of the proposed local cropping estimator over $\mathcal{Q}_\alpha(\eta,M)$. Compared to that over $\mathcal{P}_\alpha(\eta,M)$, the analysis here involves smaller bias terms which lead to a different optimal bandwidth $k\asymp n^{\frac{1}{2\alpha+1}}$.

THEOREM 5. *When* $\lceil n^{\frac{1}{2\alpha+1}}\rceil\le p$, *the local cropping estimator defined in* (12) *of the precision matrix* Ω *over* $\mathcal{Q}_\alpha(\eta,M)$ *given in* (3) *satisfies*

$$\sup_{\mathcal{Q}_\alpha(\eta,M)}\mathbb{E}\bigl\|\tilde{\Omega}_k^{\mathrm{op}}-\Omega\bigr\|_{\mathrm{op}}^2\le Ck^{-2\alpha}+C\frac{\log p+k}{n}.$$

When $k=\lceil n^{\frac{1}{2\alpha+1}}\rceil$, *we have*

$$\sup_{\mathcal{Q}_\alpha(\eta,M)}\mathbb{E}\bigl\|\tilde{\Omega}_k^{\mathrm{op}}-\Omega\bigr\|_{\mathrm{op}}^2\le Cn^{-\frac{2\alpha}{2\alpha+1}}+C\frac{\log p}{n}.$$

PROOF. We employ the same proof strategy as that of Theorem 3. Only two lemmas bounding bias terms need to be replaced. We only emphasize the differences here.

We replace Lemma 2 in the proof by Lemma 11 which bounds the distance of the population precision matrix and its tapered one.

LEMMA 11. *For* Ω *in the parameter space* $\mathcal{Q}_\alpha(\eta,M)$ *defined in* (3), Ω_k^* *is defined in* (18), *we have*

$$\bigl\|\Omega_k^*-\Omega\bigr\|_{\mathrm{op}}^2\le Ck^{-2\alpha}.$$

In addition, we replace Lemma 4 by Lemma 12 which bounds the bias term of each local estimator.

LEMMA 12. *For* $\Omega\in\mathcal{Q}_\alpha(\eta,M)$ *defined in* (2) *with* $\alpha>0$, *we have*

$$\bigl\|\mathbf{C}_{k+1}^k\bigl((\mathbf{C}_{m-k}^{3k}(\Omega^{-1}))^{-1}\bigr)-\mathbf{C}_m^k(\Omega)\bigr\|_{\mathrm{op}}^2\le Ck^{-2\alpha}.$$

The remaining part of the proof remains the same, including a similar upper bound for the variance term stated in Lemma 3. Therefore, we complete our proof. □

3.3.2. *Minimax lower bound.*

THEOREM 6. *The minimax risk for estimating the precision matrix* Ω *over* $\mathcal{Q}_\alpha(\eta,M)$ *defined in* (3) *under the operator norm with* $\alpha>0$ *satisfies*

$$\inf_{\tilde{\Omega}}\sup_{\mathcal{Q}_\alpha(\eta,M)}\mathbb{E}\|\tilde{\Omega}-\Omega\|_{\mathrm{op}}^2\ge\frac{\tau^2}{32}\biggl(n^{-\frac{2\alpha}{2\alpha+1}}+\frac{\log p}{n}\biggr).\tag{30}$$

REMARK 6. Theorems 5 and 6 together show that the minimax risk for estimating the precision matrices over $\mathcal{Q}_\alpha(\eta,M)$ stated in (5) of Theorem 1. In contrast to $\mathcal{P}_\alpha(\eta,M)$, the optimal rate of convergence over $\mathcal{Q}_\alpha(\eta,M)$ is faster. In particular, rate-optimal local cropping estimators are always consistent as long as $\alpha>0$.

PROOF. To establish the lower bound for $\mathcal{Q}_\alpha(\eta, M)$ in which the decay of a_{ij} is in the entrywise fashion, we repeat the proof scheme in Section 3.2 with a few changes. Let τ be a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

Set $k = \min\{\lceil n^{\frac{1}{2\alpha+1}} \rceil, \frac{p}{2}\}$ and the index set $\Theta = \{0,1\}^k$, that is, for any $\theta \in \Theta$, $\theta \equiv \{\theta_i\}_{1\le i\le k}$, each θ_i is either 0 or 1. Define the $k \times k$ matrix $B_k^*(\theta) \equiv [b_{ij}]_{k\times k}$ with $b_{ij} = \tau(nk)^{-\frac{1}{2}}\theta_i$. Define

$$B(\theta) = \begin{bmatrix} 0_{k\times k} & 0_{k\times k} & 0_{k\times(p-2k)} \\ B_k^*(\theta) & 0_{k\times k} & 0_{k\times(p-2k)} \\ 0_{(p-2k)\times k} & 0_{(p-2k)\times k} & 0_{(p-2k)\times(p-2k)} \end{bmatrix}.$$

We construct the collection of 2^k matrices as

$$\mathcal{P}_3 = \big\{\Omega(\theta) : \Omega(\theta) = \big(I_p - B(\theta)\big)^T\big(I_p - B(\theta)\big), \theta \in \Theta\big\}. \tag{31}$$

LEMMA 13. *$\mathcal{P}_3$ is a subset of $\mathcal{Q}_\alpha(\eta, M)$.*

Let P_θ be the joint distribution of n i.i.d. observations from $N(0, \Omega(\theta)^{-1})$, where $\Omega(\theta) \in \mathcal{P}_3$ defined in (31). Parallel to Lemmas 7 and 8 and the lower bound (28) for $\mathcal{P}_1$, we establish the following lower bound for $\mathcal{P}_3$.

LEMMA 14. *Consider all $\Omega(\theta) \in \mathcal{P}_3$ defined in (31). Then,*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \ge 0.5, \tag{32}$$

$$\min_{H(\theta,\theta')\ge 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_2^2}{H(\theta,\theta')} \ge \big(\tau n^{-\frac{1}{2}}\big)^2. \tag{33}$$

According to Assouad's lemma, for any estimator $\tilde{\Omega}$ based on n i.i.d. observations, we have

$$\sup_{\mathcal{P}_3} \mathbb{E}\|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \ge \frac{\tau^2}{16} n^{-1} \min\Big\{n^{\frac{1}{2\alpha+1}}, \frac{p}{2}\Big\}. \tag{34}$$

It is easy to show $(\mathcal{P}_3 \cup \mathcal{P}_2) \subset \mathcal{Q}_\alpha(\eta, M)$, where $\mathcal{P}_2$ is defined in (27). Therefore, combining (34) and (29), we complete the proof of Theorem 6. □

REMARK 7. The estimation of the covariance matrix Σ is of significant importance as well. We propose the estimator of Σ by inverting our estimator $\tilde{\Omega}_k^{\text{op}}$ given in (12). The results and the analysis given in Section 3 can be used to establish the minimax optimality of our estimator under the operator norm. According to the inequality $\|(\tilde{\Omega}_k^{\text{op}})^{-1} - \Sigma\|_{\text{op}} \le \|(\tilde{\Omega}_k^{\text{op}})^{-1}\|_{\text{op}}\|\tilde{\Omega}_k^{\text{op}} - \Omega\|_{\text{op}}\|\Omega^{-1}\|_{\text{op}}$ and the fact that both $\|\tilde{\Omega}_k^{\text{op}})^{-1}\|_{\text{op}}$ and $\|\Omega^{-1}\|_{\text{op}}$ are bounded by η, we establish the upper bound of our estimator $(\tilde{\Omega}_k^{\text{op}})^{-1}$. Furthermore, considering the analog between the covariance matrix and the precision matrix in the subset $\mathcal{P}_1$ and $\mathcal{P}_2$ defined in (26) and (27), the matching lower bound can be proved by a similar argument in Section 3.2. Therefore, we have the following rate optimality of estimating the covariance matrix under the operator norm, which can be achieved by estimator $(\tilde{\Omega}_k^{\text{op}})^{-1}$,

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\big\|\tilde{\Omega}^{-1} - \Omega^{-1}\big\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n},$$

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\big\|\tilde{\Omega}^{-1} - \Omega^{-1}\big\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}.$$

4. Rate optimality under the Frobenius norm. In this section we establish that the optimal rates of convergence for estimating the precision matrix over the parameter spaces $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ are identical under the Frobenius norm. Intuitively, estimating precision matrix under the Frobenius norm is equivalent to estimating each row of Cholesky factor A under the ℓ_2 vector norm. Consequently, it is not a surprise to see that $\mathcal{Q}_\alpha(\eta, M)$ and $\mathcal{P}_\alpha(\eta, M)$ enjoy the same optimal rates here. Since $\mathcal{Q}_\alpha(\eta, \alpha M) \subset \mathcal{P}_\alpha(\eta, M)$, one immediately obtains that

$$\inf_{\tilde{\Omega}} \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2 \geq \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,\alpha M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_{\mathrm{F}}^2. \tag{35}$$

In order to show (6) in Theorem 2, it suffices to establish the upper bound over the parameter space $\mathcal{P}_\alpha(\eta, M)$ and the matching lower bound over the parameter space $\mathcal{Q}_\alpha(\eta, M)$. We assume that $\mathbf{X}$ follows the p-variate Gaussian distribution, with mean zero and precision matrix Ω in this section.

4.1. *Minimax upper bound under the Frobenius norm.* In this section we establish the following risk upper bound of the regression-based thresholding estimation procedure we proposed in Section 2.2 under the Frobenius norm over $\mathcal{P}_\alpha(\eta, M)$.

THEOREM 7. *Assume $\lceil n^{\frac{1}{2\alpha+2}} \rceil \leq p$. The estimator defined in (16) of the precision matrix Ω over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, \alpha M)$ given in (2) and (3) with $k = \lceil n^{\frac{1}{2\alpha+2}} \rceil$ satisfies*

$$\sup_{\mathcal{Q}_\alpha(\eta,\alpha M)} \frac{1}{p}\mathbb{E}\big\|\tilde{\Omega}_k^{\mathrm{F}} - \Omega\big\|_{\mathrm{F}}^2 \leq \sup_{\mathcal{P}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\big\|\tilde{\Omega}_k^{\mathrm{F}} - \Omega\big\|_{\mathrm{F}}^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{36}$$

PROOF. We focus on the second inequality since the first one is trivial. Note that $\tilde{\Omega}_k^{\mathrm{F}} = \widetilde{(I-A)}^T \tilde{D}^{-1} \widetilde{(I-A)}$ according to (16) while $\Omega = (I-A)^T D^{-1}(I-A)$. The risk upper bound can be controlled by bounding $\widetilde{I-A} - (I-A)$ and $\tilde{D} - D$. To this end, we first provide some properties of our estimator.

LEMMA 15. *Assume that $\mathbf{X}$ follows the p-variate Gaussian distribution with mean zero and precision matrix $\Omega = (I-A)^T D^{-1}(I-A)$ which belongs to parameter space $\mathcal{P}_\alpha(\eta, M)$ defined in (2). For any fixed i, d_i is the ith diagonal of D, $\mathbf{a}_i \in \mathbb{R}^{i-1}$ corresponds the ith row of the lower triangle in A. $\hat{d}_i$ is defined in (15), and $\hat{\mathbf{a}}_i^* \in \mathbb{R}^{i-1}$ corresponds with the ith row of the lower triangle in $\hat{A}^*$ defined in (14). Then, we have*

$$\mathbb{E}|\hat{d}_i - d_i|^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}},$$

$$\mathbb{E}\big\|\hat{\mathbf{a}}_i^* - \mathbf{a}_i\big\|_2^2 \leq Cn^{-\frac{2\alpha+1}{2\alpha+2}}.$$

We are ready to establish the upper bounds of $\widetilde{I-A} - (I-A)$ and $\tilde{D} - D$ separately. Note that $\|\tilde{D}^{-1}\|_{\mathrm{op}} \leq \eta$ and $\|D^{-1}\|_{\mathrm{op}} \leq \eta$ which is due to Lemma B.2 of [38]. Therefore, Lemma B.1 of [38] yields $\mathbb{E}\|\tilde{D} - D\|_{\mathrm{F}}^2 \leq 4\mathbb{E}\|\hat{D} - D\|_{\mathrm{F}}^2$, which further implies that

$$\begin{aligned}\frac{1}{p}\mathbb{E}\big\|D^{-1} - \tilde{D}^{-1}\big\|_{\mathrm{F}}^2 &\leq \frac{1}{p}\mathbb{E}\big\|\tilde{D}^{-1}\big\|_{\mathrm{op}}^2 \|\tilde{D} - D\|_{\mathrm{F}}^2 \big\|D^{-1}\big\|_{\mathrm{op}}^2 \\ &\leq 4\eta^4 \frac{1}{p}\mathbb{E}\|\hat{D} - D\|_{\mathrm{F}}^2 \\ &\leq 4\eta^4 \frac{1}{p}\sum_i \mathbb{E}|\hat{d}_i - d_i|^2.\end{aligned}$$

Together with Lemma 15, it follows that

$$\frac{1}{p}\mathbb{E}\|D^{-1}-\tilde{D}^{-1}\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{37}$$

Next, we turn to prove that $\frac{1}{p}\mathbb{E}\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}$. Lemma B.1 of [38] implies

$$\frac{1}{p}\mathbb{E}\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2 \le \frac{4}{p}\mathbb{E}\|\hat{A}^*-A\|_{\mathrm{F}}^2 \le \frac{4}{p}\sum_i \mathbb{E}\|\hat{\mathbf{a}}_i^*-\mathbf{a}_i\|_2^2.$$

Combining above equation with Lemma 15, we have

$$\frac{1}{p}\mathbb{E}\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2 \le Cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{38}$$

At last, we derive the risk upper bound of our estimator. It is clear that $\|\widetilde{I-A}\|_{\mathrm{op}} \le \eta$, $\|\tilde{D}^{-1}\|_{\mathrm{op}} \le \eta$. According to Lemma B.2 of [38], $\|I-A\|_{\mathrm{op}} \le \eta$, $\|D^{-1}\|_{\mathrm{op}} \le \eta$. Combining these facts with (37) and (38), we have

$$\begin{aligned}
\frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_{\mathrm{F}}^2 \le{}& \frac{3}{p}\mathbb{E}\big(\|I-A\|_{\mathrm{op}}^2\|D^{-1}\|_{\mathrm{op}}^2\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2 \\
&+\|I-A\|_{\mathrm{op}}^2\|D^{-1}-\tilde{D}^{-1}\|_{\mathrm{F}}^2\|\widetilde{I-A}\|_{\mathrm{op}}^2 \\
&+\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2\|\tilde{D}^{-1}\|_{\mathrm{op}}^2\|\widetilde{I-A}\|_{\mathrm{op}}^2\big) \\
\le{}& 6\eta^4\frac{1}{p}\mathbb{E}\|\widetilde{(I-A)}-(I-A)\|_{\mathrm{F}}^2+3\eta^4\frac{1}{p}\mathbb{E}\|D^{-1}-\tilde{D}^{-1}\|_{\mathrm{F}}^2 \\
\le{}& Cn^{-\frac{2\alpha+1}{2\alpha+2}}.
\end{aligned}$$

Therefore, we finish the proof of Theorem 7. □

4.2. *Minimax lower bound under the Frobenius norm.* In this section we establish the matching lower bound $n^{-\frac{2\alpha+1}{2\alpha+2}}$ over parameter spaces $\mathcal{P}_\alpha(\eta,M)$ and $\mathcal{Q}_\alpha(\eta,M)$.

THEOREM 8. *The minimax risk for estimating the precision matrix Ω over $\mathcal{P}_\alpha(\eta,M)$ and $\mathcal{Q}_\alpha(\eta,\alpha M)$ under the Frobenius norm satisfies*

$$\inf_{\tilde{\Omega}}\sup_{\mathcal{P}_\alpha(\eta,M)}\frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_F^2 \ge \inf_{\tilde{\Omega}}\sup_{\mathcal{Q}_\alpha(\eta,\alpha M)}\frac{1}{p}\mathbb{E}\|\tilde{\Omega}-\Omega\|_F^2 \ge \frac{\tau^2}{32}n^{-\frac{2\alpha+1}{2\alpha+2}}.$$

REMARK 8. The minimax risk for estimating the precision matrices over $\mathcal{P}_\alpha(\eta,M)$ and $\mathcal{Q}_\alpha(\eta,M)$ under the Frobenius norm in Theorem 2 immediately follows from Theorems 7 and 8.

PROOF. It is sufficient to establish the lower bound over $\mathcal{Q}_\alpha(\eta,M)$ since the first inequality immediately follows from (35). We construct a least favorable subset in $\mathcal{Q}_\alpha(\eta,M)$. Without loss of generality, we assume $\frac{p}{2k}$ is an integer where $k=\min\{\lceil n^{\frac{1}{2\alpha+2}}\rceil,\frac{p}{2}\}$. Define the index set $\Theta'=\{0,1\}^{\frac{kp}{2}}$. For each $\theta\in\Theta'$, we further denote it as $\frac{p}{2k}$ many k^2 dimensional vectors, that is, $\theta=\{\theta(s)\}_{1\le s\le\lceil\frac{p}{2k}\rceil}$, where $\theta(s)_{ij}$ is equal to 0 or 1. For such an index θ, there is a corresponding $p\times p$ block diagonal matrix $C(\theta)$ such that each $k\times k$ block

$C_s(\theta(s)) \equiv [c(s)_{ij}]_{k\times k}$, where $c(s)_{ij} = \tau n^{-\frac{1}{2}}\theta(s)_{ij}$, $s \in \lceil \frac{p}{2k} \rceil$. We set τ as a positive constant which is less than $\min\{M, \frac{1}{4}\eta^{-1}, \eta^{\frac{1}{2}} - 1\}$.

$$C(\theta) = \begin{bmatrix} \begin{matrix} 0_k & 0_k \\ C_1(\theta(1)) & 0_k \end{matrix} & 0_{2k} & \cdots & 0_{2k} \\ 0_{2k} & \begin{matrix} 0_k & 0_k \\ C_2(\theta(2)) & 0_k \end{matrix} & \cdots & 0_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2k} & 0_{2k} & \cdots & \begin{matrix} 0_k & 0_k \\ C_{\lceil \frac{p}{2k} \rceil}\left(\theta\left(\left\lceil \frac{p}{2k} \right\rceil\right)\right) & 0_k \end{matrix} \end{bmatrix}.$$

Finally, we define the subset of $\mathcal{Q}_\alpha(\eta, M)$ indexed by Θ' as follows:

$$\mathcal{P}_4 = \{\Omega(\theta) : \Omega(\theta) = (I_p - C(\theta))^T (I_p - C(\theta)), \theta \in \Theta'\}. \tag{39}$$

LEMMA 16. *$\mathcal{P}_4$ is a subset of $\mathcal{Q}_\alpha(\eta, M)$.*

Applying Lemma 6 to $\mathcal{P}_4$, we obtain that

$$\begin{aligned} &\inf_{\tilde{\Omega}} \max_{\theta\in\Omega(\Theta')} 2^2 \mathbb{E}_\theta \|\tilde{\Omega} - \Omega(\theta)\|_F^2 \\ &\quad \geq \min_{H(\theta,\theta')\geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_F^2}{H(\theta,\theta')} \frac{kp}{4} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|. \end{aligned} \tag{40}$$

LEMMA 17. *Let P_θ be the joint distribution of n i.i.d. observations from $N(0, \Omega(\theta)^{-1})$, where $\Omega(\theta) \in \mathcal{P}_4$ defined in (39). Then,*

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq 0.5 \tag{41}$$

and

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Omega(\theta) - \Omega(\theta')\|_F^2}{H(\theta,\theta')} \geq \tau^2 n^{-1}. \tag{42}$$

Applying Lemma 17 into (40), we obtain

$$\begin{aligned} \inf_{\tilde{\Omega}} \sup_{\mathcal{Q}_\alpha(\eta,M)} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_F^2 &\geq \inf_{\tilde{\Omega}} \sup_{\mathcal{P}_4} \frac{1}{p}\mathbb{E}\|\tilde{\Omega} - \Omega\|_F^2 \\ &\geq \frac{\tau^2}{32} n^{-1} \min\left\{ n^{\frac{1}{2\alpha+2}}, \frac{p}{2} \right\}, \end{aligned}$$

which completes the proof of Theorem 8, noting that $n < p$. □

5. Adaptive estimation. To achieve the minimax rates in Theorem 1 under the operator norm, the local cropping estimator $\tilde{\Omega}_k^{\text{op}}$ requires the knowledge of smoothness parameter α as the optimal choice of bandwidth $k = \lceil n^{\frac{1}{2\alpha}} \rceil$ and $k = \lceil n^{\frac{1}{2\alpha+1}} \rceil$ over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, respectively. In this section we consider adaptive estimation, where the goal is to construct a single procedure which is minimax rate optimal simultaneously over each parameter space $\mathcal{P}_\alpha(\eta, M)$ ($\alpha > 1/2$) and $\mathcal{Q}_\alpha(\eta, M)$ ($\alpha > 0$). Throughout this section we assume that $\mathbf{X}$ follows certain sub-Gaussian distribution defined in (17).

Recall that for each k, the local cropping estimator $\tilde{\Omega}_k^{\mathrm{op}}$ is defined in (12). Without the knowledge of α, the bandwidth k needs to be picked in a data-driven fashion. Motivated by Lepski's methods for nonparametric function estimation problems [35], we select the bandwidth $\hat{k}$ through the following procedure:

$$\hat{k} = \min\left\{k \in \mathcal{H} : \|\tilde{\Omega}_k^{\mathrm{op}} - \tilde{\Omega}_l^{\mathrm{op}}\|_{\mathrm{op}}^2 \le C_L \frac{\log p + l}{n}, \text{ for all } l \ge k\right\}, \tag{43}$$

where $\mathcal{H} = \{1, 2, \ldots \lceil \frac{n}{\log p} \rceil\}$ and $C_L > 0$ is a sufficiently large constant. If the set that is minimized over is empty, we use the convention $\hat{k} = \lceil \frac{n}{\log p} \rceil$. The adaptive local cropping estimator $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ enjoys the following theoretical guarantee and, thus, is adaptive minimax rate optimal.

THEOREM 9. *Assume* $\log p = O(n)$, $n = O(p)$. *Then, the adaptive estimator* $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ *with* $\hat{k}$ *defined in* (43) *of the precision matrix* Ω *over* $\mathcal{P}_\alpha(\eta, M)$ *with* $\alpha > \frac{1}{2}$ *satisfies*

$$\sup_{\mathcal{P}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}.$$

In addition, the adaptive estimator $\tilde{\Omega}_{\hat{k}}^{\mathrm{op}}$ *over* $\mathcal{Q}_\alpha(\eta, M)$ *with* $\alpha > 0$ *satisfies*

$$\sup_{\mathcal{Q}_\alpha(\eta,M)} \mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le C n^{-\frac{2\alpha}{2\alpha+1}} + C\frac{\log p}{n}.$$

PROOF. We only show the upper bound over $\mathcal{P}_\alpha(\eta, M)$ with $\alpha > \frac{1}{2}$. The proof over space $\mathcal{Q}_\alpha(\eta, M)$ with $\alpha > 0$ can be shown similarly and thus omitted.

Set the oracle bandwidth $k^* = \lceil n^{\frac{1}{2\alpha}} \rceil$. For any $\Omega \in \mathcal{P}_\alpha(\eta, M)$, we decompose the risk as follows:

$$\mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le 2\mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 + 2\mathbb{E}\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2. \tag{44}$$

Since k^* is deterministic, we immediately obtain from Theorem 3 that

$$\mathbb{E}\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n} \tag{45}$$

which controls the second term of the risk decomposition (44).

We turn to bound the first term of (44). Due to the definition of $\hat{k}$ and k^*, we have that on the event $\{\hat{k} \le k^*\}$,

$$\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 \le C_L \frac{\log p + k^*}{n} \le C n^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n}. \tag{46}$$

It suffices to show that $\hat{k} \le k^*$ with high probability. The following lemma, a probability version of Theorem 3, facilitates our proof of this claim.

LEMMA 18. *Assume* $\lceil n^{\frac{1}{2\alpha}} \rceil \le p$. *Then, for any constant* $C_1 > 0$, *there exists a sufficiently large constant* $C > 0$ *irrelevant of* α *such that the local cropping estimator defined in* (12) *satisfies*

$$\mathbb{P}\left(\|\tilde{\Omega}_k^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \le C k^{-2\alpha+1} + C\frac{\log p + k}{n}\right) > 1 - \exp(-C_1(\log p + k)),$$

simultaneously for each $k \in \mathcal{H}$ *and each* $\Omega \in \mathcal{P}_\alpha(\eta, M)$ *with* $\alpha > \frac{1}{2}$.

Notice that for any l, we have $\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \tilde{\Omega}_l^{\mathrm{op}}\|_{\mathrm{op}}^2 \le 2\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 + 2\|\Omega - \tilde{\Omega}_l^{\mathrm{op}}\|_{\mathrm{op}}^2$. Thus,

$$
\begin{aligned}
&\mathbb{P}(\hat{k} > k^*) \\
&\le \sum_{l \ge k^*} \mathbb{P}\left(\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \tilde{\Omega}_l^{\mathrm{op}}\|_{\mathrm{op}}^2 > C_L \frac{\log p + l}{n} \right) \\
&\le \sum_{l \ge k^*} \left(\mathbb{P}\left(\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 > \frac{C_L}{4} \frac{\log p + k^*}{n} \right) \right. \\
&\qquad \left. + \mathbb{P}\left(\|\tilde{\Omega}_l^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 > \frac{C_L}{4} \frac{\log p + l}{n} \right) \right) \\
&\le n\left(\exp\left(-C_1(\log p + k^*)\right) + \exp\left(-C_1(\log p + l)\right)\right) \\
&\le n^{-1}\eta^{-2}.
\end{aligned} \tag{47}
$$

We have used the fact $k^* \le l$ and the definition of k^* in the inequalities above, noting that a sufficiently large $C_1 > 0$ can be picked to guarantee the last inequality holds. The second to last inequality holds because of Lemma 18 and a sufficiently large C_L. Therefore, we have shown that the event $\hat{k} \le k^*$ holds with probability at least $1 - n^{-1}\eta^{-2}$.

In the end, combining (44)–(47), we obtain that for any $\Omega \in \mathcal{P}_\alpha(\eta, M)$,

$$
\begin{aligned}
&\mathbb{E}\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 \\
&\le 2\mathbb{E}\|\tilde{\Omega}_{k^*}^{\mathrm{op}} - \Omega\|_{\mathrm{op}}^2 + 2\mathbb{E}\left(\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 : \hat{k} \le k^*\right) \\
&\quad + 2\mathbb{E}\left(\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 : \hat{k} > k^*\right) \\
&\le Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n} + 8\eta^2\mathbb{P}(\hat{k} > k^*) \\
&\le Cn^{-\frac{2\alpha-1}{2\alpha}} + C\frac{\log p}{n} + 8n^{-1} \\
&\le C\left(n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}\right),
\end{aligned}
$$

where we also used that $\|\tilde{\Omega}_{\hat{k}}^{\mathrm{op}} - \tilde{\Omega}_{k^*}^{\mathrm{op}}\|_{\mathrm{op}}^2 \le 4\eta^2$ in the second inequality. Therefore, we complete the proof. □

6. An extension to nonparanormal distributions. In this section we extend the minimax framework to the nonparanormal model. Assume that $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ follows the p-variate Gaussian distribution with covariance matrix Σ. Instead of n i.i.d. copies $\mathbf{X}_1, \mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$, we only observe their transformations. Specifically, we denote the transformed variables of $\mathbf{X}$ by $\mathbf{Y} = (f_1(X_1), f_2(X_2), \ldots, f_p(X_p))^T$, where each f_i is some unknown strictly increasing function. Then, our observation is $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)^T \in \mathbb{R}^{n \times p}$, where each $\mathbf{Y}_i$ is the transformed $\mathbf{X}_i$. This is a form of the Gaussian copula model [5] or the nonparanormal model [37]. To avoid the identifiability issue, we set $\mathrm{diag}(\Sigma) = I$ which makes Σ the correlation matrix. Here, we consider the same structural assumption as in previous sections on the inverse of the correlation matrix which is denoted by Ω. Based on $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ and defined in (2) and (3), the following two types of parameter spaces are of interest:

$$
\mathcal{P}'_\alpha(\eta, M) = \left\{ \{\Omega, \{f_i\}\} : \begin{array}{l} \mathrm{diag}(\Omega^{-1}) = I, \Omega \in \mathcal{P}_\alpha(\eta, M); \\ f_i \text{ is strictly increasing}, i \in [p] \end{array} \right\} \tag{48}
$$

and

$$\mathcal{Q}'_\alpha(\eta, M) = \left\{ \{\Omega, \{f_i\}\} : \begin{array}{c} \operatorname{diag}(\Omega^{-1}) = I, \Omega \in \mathcal{Q}_\alpha(\eta, M); \\ f_i \text{ is strictly increasing}, i \in [p] \end{array} \right\}. \tag{49}$$

Our goal is to estimate the latent correlation structure, the inverse of the correlation matrix Ω, using the observation $\mathbf{Z}$. We establish the minimax risk of estimating Ω over the parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$ under the operator norm in the following theorem.

THEOREM 10. *Assume* $\log p = O(n)$, $n = O(p)$. *Then, for the nonparanormal model, the minimax risk of estimating* Ω *under the operator norm over* $\mathcal{P}'_\alpha(\eta, M)$ *with* $\alpha > \frac{1}{2}$ *satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\{\Omega, \{f_i\}\} \in \mathcal{P}'_\alpha(\eta, M)} \mathbb{E} \|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha-1}{2\alpha}} + \frac{\log p}{n}. \tag{50}$$

The minimax risk of estimating Ω *under the operator norm over* $\mathcal{Q}'_\alpha(\eta, M)$ *satisfies*

$$\inf_{\tilde{\Omega}} \sup_{\{\Omega, \{f_i\}\} \in \mathcal{Q}'_\alpha(\eta, M)} \mathbb{E} \|\tilde{\Omega} - \Omega\|_{\text{op}}^2 \asymp n^{-\frac{2\alpha}{1+2\alpha}} + \frac{\log p}{n}. \tag{51}$$

Finally, we introduce our rate-optimal estimation procedure over the parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$ under the operator norm. The approach to estimate the inverse of the correlation matrix in nonparanormal model is almost the same as the estimation scheme of the precision matrix under the operator norm in Section 2.1, except that the sample covariance matrix needs to be replaced by its rank-based nonparametric variant via Kendall's tau (τ) [31] or Spearman's correlation coefficient rho (ρ) [49]. Rank-based estimators are widely applied in the nonparanormal model. Progress has been made in this field during the last decade, especially for high-dimensional statistics. For instance, see [41] for bandable correlation matrix estimation, [4] for Gaussian graphical models and [22] for multitask regression via Cholesky decomposition.

Kendall's tau is defined as

$$\hat{\tau}_{ij} = \frac{2}{n(n-1)} \sum_{1 \le k_1 < k_2 \le n} \operatorname{sgn}(Z_{k_1 i} - Z_{k_2 i}) \operatorname{sgn}(Z_{k_1 j} - Z_{k_2 j}).$$

Then, define

$$\hat{\Sigma}^\tau = \left[\sin\left(\frac{\pi}{2} \hat{\tau}_{ij} \right) \right]_{p \times p}. \tag{52}$$

Spearman's rho is defined as

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^n (r_{ki} - (n+1)/2)(r_{kj} - (n+1)/2)}{\sqrt{\sum_{k=1}^n (r_{ki} - (n+1)/2)^2 \sum_{k=1}^n (r_{kj} - (n+1)/2)^2}},$$

where r_{ij} is the rank of Z_{ij} among $Z_{1j}, Z_{2j}, \ldots, Z_{nj}$. Define

$$\hat{\Sigma}^\rho = \left[2 \sin\left(\frac{\pi}{6} \hat{\rho}_{ij} \right) \right]_{p \times p}. \tag{53}$$

It is well known that both $\hat{\Sigma}^\tau$ and $\hat{\Sigma}^\rho$ are unbiased estimators of the population correlation matrix Σ. We adopt almost the same estimation procedure proposed in Section 2.1 but replace $\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ in (10) with either $\hat{\Sigma}^\tau$ or $\hat{\Sigma}^\rho$. In this way we construct the nonparametric local cropping estimators $\tilde{\Omega}_k^\tau$ and $\tilde{\Omega}_k^\rho$ in place of $\tilde{\Omega}_k^{\text{op}}$ in (12). Note that the optimal choices of the bandwidth k are picked differently over two types of parameter spaces $\mathcal{P}'_\alpha(\eta, M)$ and $\mathcal{Q}'_\alpha(\eta, M)$, as we did over $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ in Section 2.1. To provide some technical insights, we rely on some recent results in [41] to bound the variance of each local estimator in (10) under the operator norm which is the key to establish the upper bounds in Theorem 10.

7. Numerical studies. In this section we turn to the numerical performance of the proposed rate-optimal estimators under the operator norm for $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ defined in (2) and (3) to further illustrate the fundamental difference of $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$. In addition, we compare them with the banding estimator proposed in [6] which is based on the auto-regression between variables. Specifically, for a given bandwidth $k < n$, the banding estimator is defined as $\tilde{\Omega}^{BL} = (I - \tilde{A}^{BL})^T(\tilde{D}^{BL})^{-1}(I - \tilde{A}^{BL})$. Here, the ith row of the lower triangular matrix $\tilde{A}^{BL}$ is the vector $\hat{\mathbf{a}}_i$ in (13), that is, the least square estimates of the coefficients for the regression of X_i against $\mathbf{X}_{i-k:i-1}$. The ith entry of the diagonal matrix $\tilde{D}^{BL}$ is the estimate of the residual variance for the regression of X_i against $\mathbf{X}_{i-k:i-1}$.

7.1. *Simulation in $\mathcal{Q}_\alpha(\eta, M)$ under the operator norm.* We first focus on the parameter space $\mathcal{Q}_\alpha(\eta, M)$ and compare the performance of local cropping estimator and the banding estimator. Specifically, we generate the precision matrix in the following form:

$$\Omega = (I - A)^T D^{-1}(I - A), \quad A \equiv [a_{ij}]_{p\times p}, D = I_p,$$

where $a_{ij} = -(i - j)^{-\alpha-1}$ when $i > j$; otherwise, $a_{ij} = 0$. It is easy to check that $\Omega \in \mathcal{Q}_\alpha(\eta, 1)$ with some large $\eta > 0$. The simulation is done with a range of parameter values for p, n, α. Specifically, the decay rate α ranges from 0.5 to 2 with a step of 0.5, the sample size n ranges from 500 to 4000, the dimension p ranges from 500 to 2000.

In this setting we compare our local cropping estimator (denoted as cropping.Q.) with the banding estimator (denoted as BL) proposed in [6]. According to [6], the bandwidth of banding estimator is chosen as $k \asymp (n/\log p)^{1/(2\alpha+2)}$. The optimal bandwidth over $\mathcal{Q}_\alpha(\eta, M)$ is $k \asymp n^{1/(2\alpha+1)}$. In the simulation the bandwidth of BL estimator is $\lfloor (n/\log p)^{1/(2\alpha+2)} \rfloor$ and the bandwidth of crop.Q is $\lfloor n^{1/(2\alpha+1)} \rfloor$.

Table 1 reports the average errors of the banding estimator (BL) and local cropping estimator (crop.Q) under the operator norm over 100 replications. The smaller errors in each experiment are highlighted in boldface. Figure 2 displays the boxplots of the errors of BL and crop.Q.

It can be seen from Table 1 that crop.Q outperforms BL in most cases with a few exceptions when n is small. As the sample size increases, the average errors of both methods decrease

TABLE 1
The average errors under the operator norm of the banding estimator (BL) and the local cropping estimator (crop.Q) over 100 replications

p	n	$\alpha = 0.5$		$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2$	
		crop.Q	BL	crop.Q	BL	crop.Q	BL	crop.Q	BL
500	500	**4.68**	5.44	**1.64**	2.38	1.18	**1.16**	0.93	**0.81**
	1000	**3.29**	4.89	**1.17**	1.72	**0.82**	1.08	**0.66**	0.69
	2000	**2.47**	4.45	**0.89**	1.33	**0.59**	0.69	**0.48**	0.59
	4000	**1.84**	3.80	**0.62**	1.07	**0.41**	0.64	**0.34**	0.53
1000	500	**4.96**	5.74	**1.75**	2.40	1.30	**1.19**	0.99	**0.84**
	1000	**3.43**	5.19	**1.24**	1.74	**0.86**	1.10	**0.68**	0.70
	2000	**2.58**	4.75	**0.93**	1.35	**0.62**	0.71	**0.51**	0.60
	4000	**1.93**	4.10	**0.66**	1.33	**0.44**	0.65	**0.36**	0.55
2000	500	**5.14**	5.97	**1.85**	2.41	1.33	**1.21**	1.06	**0.89**
	1000	**3.58**	5.41	**1.30**	1.76	**0.90**	1.12	0.72	**0.71**
	2000	**2.69**	4.97	**0.98**	1.37	**0.65**	0.73	**0.54**	0.62
	4000	**2.01**	4.32	**0.69**	1.34	**0.45**	0.66	**0.38**	0.55

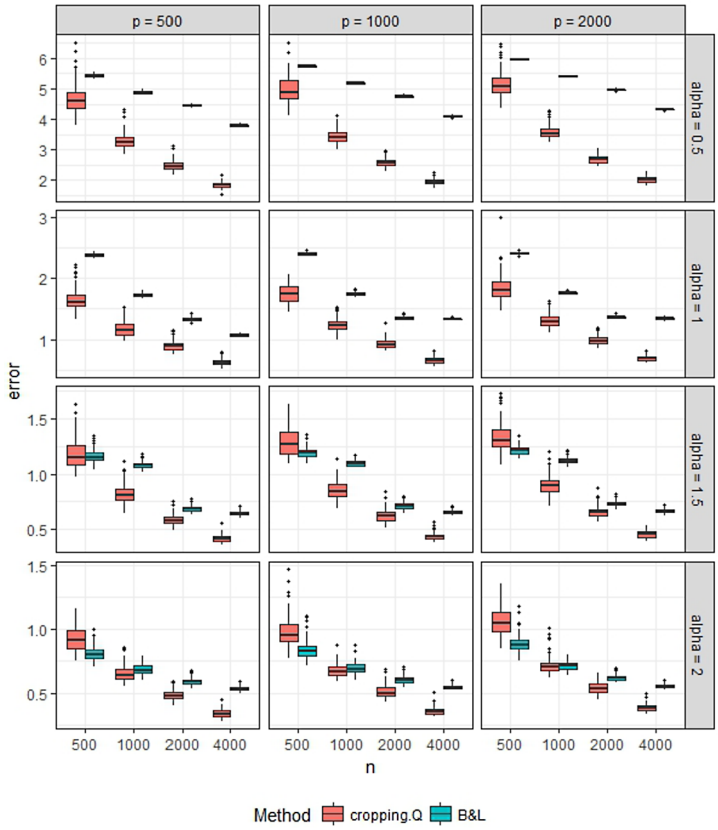

FIG. 2. *The boxplot of the errors from the local cropping estimator with the optimal bandwidth in* $\mathcal{Q}_\alpha(\eta, M)$ (*cropping.Q*) *and the banding estimator* (*BL*) *over* 100 *replications*.

which matches our intuition. In addition, the dimension p has minor effect on the errors of both estimators which is partially reflected by the optimal rates (dominating term $n^{-\frac{2\alpha}{2\alpha+1}}$) obtained in Theorem 1. For each fixed dimension p, the superiority crop.Q over BL becomes more significant as the sample size n increases which implies that BL estimator is indeed suboptimal.

7.2. *Simulation in* $\mathcal{P}_\alpha(\eta, M)$ *under the operator norm*. We demonstrate the fundamental difference between two types of parameter space $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$ by numerical studies in this section. Of note, although local cropping estimators proposed in (12) are rate-optimal over both $\mathcal{P}_\alpha(\eta, M)$ and $\mathcal{Q}_\alpha(\eta, M)$, the corresponding optimal choices of bandwidth are distinct. We generate precision matrices in the following way to guarantee that Ω is always in $\mathcal{P}_\alpha(\eta, M)$ but not in $\mathcal{Q}_\alpha(\eta, M)$ with some fixed η and M. Considering

$$\Omega = (I - A)^T D^{-1}(I - A), \quad A \equiv [a_{ij}]_{p\times p}, D = I_p,$$

where the first column of A is $a_{i1} = -2(i-1)^{-\alpha}$, $2 \le i \le p$. The remaining entries are all zeros. It is easy to check that $\Omega \in \mathcal{P}_\alpha(\eta, 2)$ with some large $\eta > 0$. The simulation is carried out with a similar range of values for p, n, α as in Section 7.1. Note that the consistent estimator exists only if $\alpha > 0.5$. Therefore, in this setting, the decay rate α varies among 1, 1.5 and 2.

The optimal choice of bandwidth of local cropping estimator over $\mathcal{P}_\alpha(\eta, M)$ is $k \asymp n^{\frac{1}{2\alpha}}$ which is different from the one of crop.Q. We denote this rate-optimal estimator in $\mathcal{P}_\alpha(\eta, M)$

TABLE 2

The average errors under the operator norm of the banding estimator (BL) and the local cropping estimators (crop.P & crop.Q) over 100 *replications*

p	n	$\alpha = 1$			$\alpha = 1.5$			$\alpha = 2$		
		crop.P	crop.Q	BL	crop.P	crop.Q	BL	crop.P	crop.Q	BL
500	500	1.50	**1.18**	2.32	**0.66**	0.73	0.86	**0.52**	0.65	0.53
	1000	1.09	**0.96**	1.80	**0.47**	0.56	0.83	**0.38**	0.56	0.45
	2000	0.83	**0.80**	1.53	**0.35**	0.43	0.55	**0.27**	0.32	0.41
	4000	**0.64**	0.68	1.33	**0.26**	0.35	0.54	**0.19**	0.24	0.38
1000	500	1.50	**1.20**	2.36	**0.68**	0.74	0.91	**0.57**	0.68	0.59
	1000	1.12	**0.98**	1.82	**0.49**	0.58	0.81	**0.39**	0.55	0.46
	2000	0.84	**0.81**	1.54	**0.37**	0.44	0.55	**0.27**	0.32	0.41
	4000	**0.65**	0.68	1.52	**0.26**	0.35	0.53	**0.19**	0.24	0.38
2000	500	1.51	**1.21**	2.39	**0.69**	0.75	0.96	**0.62**	0.71	0.63
	1000	1.16	**1.00**	1.81	**0.51**	0.60	0.84	**0.39**	0.56	0.46
	2000	0.85	**0.81**	1.55	**0.39**	0.44	0.56	**0.27**	0.33	0.41
	4000	**0.65**	0.69	1.70	**0.26**	0.35	0.53	**0.19**	0.24	0.38

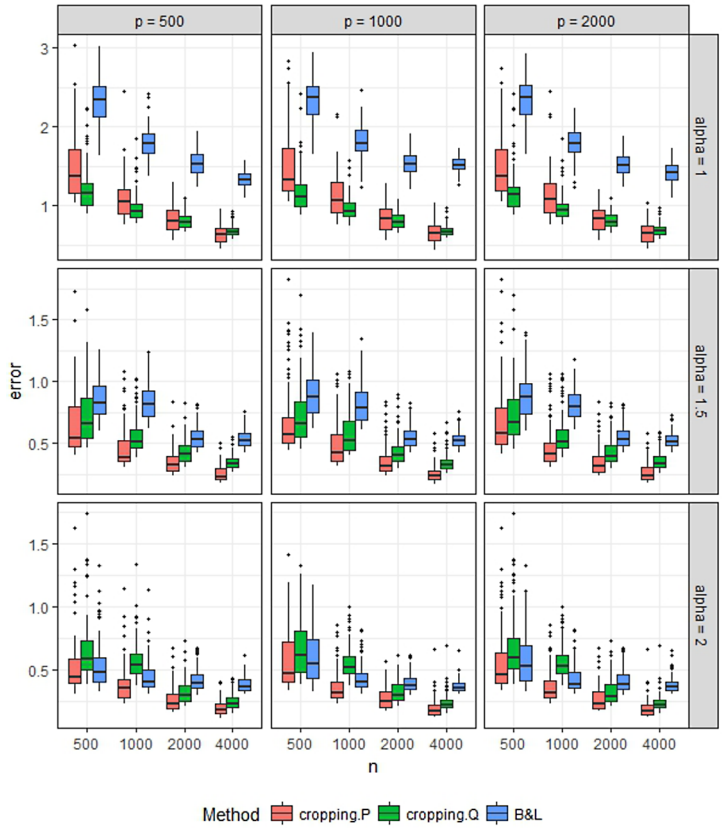

FIG. 3. *The boxplot of the errors from the local cropping estimator with the optimal bandwidth in* $\mathcal{P}_\alpha(\eta, M)$ *(cropping.P), the local cropping estimator with the optimal bandwidth in* $\mathcal{Q}_\alpha(\eta, M)$ *(cropping.Q) and the banding estimator (BL) over* 100 *replications.*

by crop.P. In the simulation the bandwidth of crop.P is $\lfloor n^{\frac{1}{2\alpha}} \rfloor$. We also include BL estimator as a reference.

Table 2 reports the average errors of the three procedures, crop.P, crop.Q and BL, under the operator norm over 100 replications. The smallest errors in each experiment are highlighted in boldface. Figure 3 plots the boxplots of their errors for $p = 500, 1000, 2000$.

Since Ω always belongs to $\mathcal{P}_\alpha(\eta, M)$ but not $\mathcal{Q}_\alpha(\eta, M)$, the estimator crop.Q is suboptimal and thus expected to have an inferior performance. Table 2 shows this point, that is, for fixed p and α, the advantage of crop.P is more obvious as n increases. Especially, crop.P outperforms the other two estimators when $n = 4000$. We also see a similar pattern as in Table 1 that p has minor effect on the errors of all the estimators.

Acknowledgments. This work was supported in part by NSF Grant DMS-1812030, an AMS Simons Travel Grant and the Central Research Development Fund at the University of Pittsburgh.

SUPPLEMENTARY MATERIAL

Supplement to "Minimax estimation of large precision matrices with bandable Cholesky factor" (DOI: 10.1214/19-AOS1893SUPP; .pdf). In this supplement, we provide key lemmas in the proofs of those main theorems, additional numerical studies as well as some discussions.

REFERENCES

[1] ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.* **296** 1021–1024. MR0777600

[2] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[3] BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. MR3273620 https://doi.org/10.1214/14-EJS945

[4] BARBER, R. F. and KOLAR, M. (2018). ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *Ann. Statist.* **46** 3422–3450. MR3852657 https://doi.org/10.1214/17-AOS1663

[5] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[6] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

[7] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[8] BIEN, J., BUNEA, F. and XIAO, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. MR3538709 https://doi.org/10.1080/01621459.2015.1058265

[9] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560

[10] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973 https://doi.org/10.1198/jasa.2011.tm10155

[11] CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494 https://doi.org/10.1214/11-STS355

[12] CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. MR3476606 https://doi.org/10.1214/13-AOS1171

[13] CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. MR3055254 https://doi.org/10.1007/s00440-012-0422-7

[14] Cai, T. T., Ren, Z. and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172 https://doi.org/10.1214/15-EJS1081

[15] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885 https://doi.org/10.1214/09-AOS752

[16] Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998

[17] d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568 https://doi.org/10.1137/060670985

[18] Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. MR1400080 https://doi.org/10.1006/acha.1996.0017

[19] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** 14863–14868.

[20] El Karoui, N. (2003). On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. Preprint. Available at arXiv:math/0309355.

[21] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559

[22] Fan, J., Xue, L. and Zou, H. (2016). Multitask quantile regression under the transnormal model. *J. Amer. Statist. Assoc.* **111** 1726–1735. MR3601731 https://doi.org/10.1080/01621459.2015.1113973

[23] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255. MR2301751 https://doi.org/10.1016/j.jmva.2006.08.003

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

[25] Heyer, M. H. and Schloerb, F. P. (1997). Application of principal component analysis to large-scale spectral line imaging studies of the interstellar medium. *Astrophys. J.* **475** 173.

[26] Hu, A. and Negahban, S. (2017). Minimax estimation of bandable precision matrices. In *Advances in Neural Information Processing Systems* 4893–4901.

[27] Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742 https://doi.org/10.1093/biomet/93.1.85

[28] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[29] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[30] Ke, Y., Minsker, S., Ren, Z., Sun, Q. and Zhou, W.-X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.*. To appear.

[31] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

[32] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

[33] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance* **10** 603–621.

[34] Lee, K. and Lee, J. (2017). Estimating large precision matrices via modified Cholesky decomposition. Preprint. Available at arXiv:1707.01143.

[35] Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. MR1147167 https://doi.org/10.1137/1136085

[36] Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* **2** 245–263. MR2415602 https://doi.org/10.1214/07-AOAS139

[37] Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[38] Liu, Y. and Ren, Z. (2020). Supplement to "Minimax estimation of large precision matrices with bandable Cholesky factor." https://doi.org/10.1214/19-AOS1893SUPP.

[39] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[40] Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

[41] Mitra, R. and Zhang, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Preprint. Available at arXiv:1403.6195.

[42] PADMANABHAN, N., WHITE, M., ZHOU, H. H. and O'CONNELL, R. (2016). Estimating sparse precision matrices. *Mon. Not. R. Astron. Soc.* **460** 1567–1576.
[43] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[44] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766 https://doi.org/10.1214/11-EJS631
[45] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286
[46] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391 https://doi.org/10.1214/08-EJS176
[47] SAULIS, L. and STATULEVIČIUS, V. A. (1991). *Limit Theorems for Large Deviations*. *Mathematics and Its Applications* (*Soviet Series*) **73**. Kluwer Academic, Dordrecht. MR1171883 https://doi.org/10.1007/978-94-011-3530-6
[48] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. MR2183942 https://doi.org/10.2202/1544-6115.1175
[49] SPEARMAN, C. (1904). Spearman's rank correlation coefficient. *Am. J. Psychol.* **15** 72–101.
[50] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466
[51] WACHTER, K. W. (1976). Probability plotting points for principal components. In *Ninth Interface Symposium Computer Science and Statistics* 299–308. Prindle, Weber and Schmidt, Boston, MA.
[52] WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6** 1–18. MR0467894 https://doi.org/10.1214/aop/1176995607
[53] WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760 https://doi.org/10.1093/biomet/90.4.831
[54] WU, W. B. and POURAHMADI, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19** 1755–1768. MR2589209
[55] XIAO, L. and BUNEA, F. (2014). On the theoretic and practical merits of the banding estimator for large covariance matrices. Preprint. Available at arXiv:1402.0844.
[56] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963
[57] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856
[58] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018