The Annals of Statistics
2019, Vol. 47, No. 3, 1321–1350
https://doi.org/10.1214/18-AOS1716
© Institute of Mathematical Statistics, 2019

ESTIMATION OF LARGE COVARIANCE AND PRECISION MATRICES FROM TEMPORALLY DEPENDENT OBSERVATIONS

BY HAI SHU[1] AND BIN NAN[1]

University of Michigan and University of California, Irvine

We consider the estimation of large covariance and precision matrices from high-dimensional sub-Gaussian or heavier-tailed observations with slowly decaying temporal dependence. The temporal dependence is allowed to be long-range so with longer memory than those considered in the current literature. We show that several commonly used methods for independent observations can be applied to the temporally dependent data. In particular, the rates of convergence are obtained for the generalized thresholding estimation of covariance and correlation matrices, and for the constrained ℓ_1 minimization and the ℓ_1 penalized likelihood estimation of precision matrix. Properties of sparsistency and sign-consistency are also established. A gap-block cross-validation method is proposed for the tuning parameter selection, which performs well in simulations. As a motivating example, we study the brain functional connectivity using resting-state fMRI time series data with long-range temporal dependence.

1. Introduction. Let $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ be a sample of p-dimensional random vectors, each with the same mean $\boldsymbol{\mu}_p$, covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. It is well known that the sample covariance matrix is not a consistent estimator of $\boldsymbol{\Sigma}$ when p grows with n [3, 4]. When the sample observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent and identically distributed (i.i.d.), several regularization methods have been proposed for the consistent estimation of large $\boldsymbol{\Sigma}$, including thresholding [10, 17, 31, 61], block-thresholding [20], banding [11] and tapering [21]. Existing methods also include the Cholesky-based method [46, 62], penalized pseudo-likelihood method [48] and sparse matrix transform [23]. Consistent correlation matrix estimation can be obtained similarly from i.i.d. observations [31, 47].

The precision matrix $\boldsymbol{\Omega} = (\omega_{ij})_{p\times p}$, when it exists, is closely related to the partial correlations between the pairs of variables in a vector $\boldsymbol{X}$. Specifically, the partial correlation between X_i and X_j given $\{X_k, k \neq i, j\}$ is equal to $-\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$ [29]. Zero partial correlation means conditional independence between Gaussian

Received July 2017; revised December 2017.

[1] Supported in part by NIH Grant R01-AG036802 and NSF Grant DMS-1407142.

MSC2010 subject classifications. Primary 62H12; secondary 62H35.

Key words and phrases. Brain functional connectivity, correlation matrix, heavy tail, high-dimensional data, long memory, minimax optimal convergence rates, nonstationarity, sub-Gaussian tail, temporal dependence.

or nonparanormal random variables [51]. There is a rich literature on the estimation of large $\mathbf{\Omega}$ from i.i.d. observations. Various algorithms for the ℓ_1 penalized maximum likelihood method (ℓ_1-MLE) and its variants have been developed [5, 35, 44, 76], and related theoretical properties have been investigated by [48, 59, 60]. Methods of estimating $\mathbf{\Omega}$ column-by-column thus implementable with parallel computing include nodewise Lasso [54, 70], graphical Dantzig selector [75], constrained ℓ_1-minimization for inverse matrix estimation (CLIME) [18] and adaptive CLIME [19].

Recently, researchers become increasingly interested in estimating the large covariance and precision matrices from temporally dependent observations $\{\boldsymbol{X}_t : t = 1, \dots, n\}$; here, t denotes time. Such research is particularly useful in analyzing the resting-state functional magnetic resonance imaging (rfMRI) data to assess the brain functional connectivity [56, 64]. In such imaging studies, the number of brain nodes (voxels or regions of interest) p can be greater than the number of images n. The temporal dependence of time series $\boldsymbol{X}_t$ is traditionally dealt with by imposing the so-called strong mixing conditions [14]. To overcome the difficulties in computing strong mixing coefficients and verifying strong mixing conditions, [73] introduced a new type of dependence measure, the functional dependence measure, and recently applied it to the hard thresholding estimation of large covariance matrix and the ℓ_1-MLE type methods of large precision matrix [24]. The functional dependence measure may still be difficult to understand and to interpret for most data analysts. Practically, it is straightforward to describe the temporal dependence directly by using cross-correlations [15]. By imposing certain weak dependence conditions on the cross-correlation matrix of samples $\{\boldsymbol{X}_t\}_{t=1}^n$, [8, 9] extended the banding and tapering regularization methods for estimating covariance matrix.

A univariate stationary time series is said to be long-memory if its autocorrelation function $\rho(t)$ satisfies $\sum_{t=0}^{\infty} |\rho(t)| = \infty$, and short-memory otherwise [55]. The rfMRI data have been reported with long-memory in the scientific literature, for example, [41, 68]. However, the temporal dependence considered by [24] and that considered by [8, 9] do not cover any long-memory time series. Later we will illustrate that the rfMRI data example does not meet their restrictive temporal dependence conditions. Hence, it is important to show the applicability of the estimating methods for i.i.d. samples to this kind of data. In this article, we characterize the temporal dependence solely using the Frobenius norm and the spectral norm of the autocorrelation matrix of each time series in $\{\boldsymbol{X}_t\}_{t=1}^n$. Simple bounds of these norms clearly display the effect of temporal dependence on the convergence rates of our considered matrix estimators, allowing each time series to be long-memory or even to be nonstationary. So the rfMRI data can be well handled by our relaxed assumption (see Figure 1). To the best of our knowledge, this is the first work that investigates the estimation of large covariance and precision matrices from long-memory observations.

Note that the estimation of large correlation matrix was not considered by either [24] or [8, 9], which is a more interesting problem in, for example, the study of brain functional connectivity. It was considered in a recent work by [77] but under the assumption that all p time series have the same temporal decay rate, which is rather restrictive and often violated (see Figure 1 for an example of rfMRI data). Moreover, all four aforementioned articles assumed that $\boldsymbol{\mu}_p = (\mu_{pi})_{1 \le i \le p}$ is known, which may not be true in practice. Although the sample mean $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ entrywise converges to μ_{pi} in probability or even almost surely under some dependence conditions [15, 45], extra care will still be needed when true mean is replaced by sample mean in the matrix estimation, especially for long-memory, heavy-tailed or even nonstationary data. We consider unknown $\boldsymbol{\mu}_p$ in this article, and show that the mean estimation indeed affects our derived matrix convergence rates, particularly for data with heavy tail probabilities.

In this article, we study the generalized thresholding estimation [61] for covariance and correlation matrices, and the CLIME approach [18] and a ℓ_1-MLE type method called sparse permutation invariant covariance estimation (SPICE; [60]) for precision matrix. The convergence rates, sparsistency and sign-consistency are provided for temporally dependent data, potentially with long memory, which are generated from a linear spatiotemporal model with all basis random variables coming from sub-Gaussian, or generalized subexponential, or distributions with polynomial-type tails. We also establish the minimax optimal convergence rates of estimating covariance and correlation matrices for a certain class of temporally dependent sub-Gaussian data, including short-memory and some long-memory cases, and show that they can be achieved by the generalized thresholding method. Moreover, if the matrix ℓ_1 norm of the precision matrix is bounded by a constant for such data, then the CLIME estimator attains the minimax optimal rates for i.i.d. observations shown in [19].

The article is organized as follows. In Section 2, we introduce the useful temporal dependence bounds and the considered temporally dependent data generating mechanism. We provide the theoretical results for the estimation of covariance and correlation matrices in Section 3 and for the estimation of precision matrix in Section 4 for temporally dependent observations with sub-Gaussian tails. We consider extensions to data with generalized sub-exponential tails and polynomial-type tails in Section 5. In Section 6, we introduce a gap-block cross-validation method for the tuning parameter selection, evaluate the estimating performance via simulations and analyze a rfMRI data set for brain functional connectivity. The concentration inequalities that the proofs of the theoretical results are based on are given in the Appendix. Detailed proofs together with additional numerical considerations are provided in the Supplementary Material [65] due to the page limitation.

2. Temporal dependence. We start with a brief introduction of useful notation. For a real matrix $\mathbf{M} = (M_{ij})$, we define: the spectral norm $\|\mathbf{M}\|_2 =$

$[\varphi_{\max}(\mathbf{M}^\top\mathbf{M})]^{1/2}$, where $\varphi_{\max}$ is the largest eigenvalue, also φ_k and $\varphi_{\min}$ are the kth largest and the smallest eigenvalues, respectively, the Frobenius norm $\|\mathbf{M}\|_F = (\sum_i \sum_j M_{ij}^2)^{1/2}$, the matrix ℓ_1 norm $\|\mathbf{M}\|_1 = \max_j \sum_i |M_{ij}|$, the entrywise ℓ_1 norm $|\mathbf{M}|_1 = \sum_{i,j} |M_{ij}|$ and its off-diagonal version $|\mathbf{M}|_{1,\mathrm{off}} = \sum_{i \neq j} |M_{ij}|$ and the entrywise ℓ_∞ norm $|\mathbf{M}|_\infty = \max_{i,j} |M_{ij}|$.

Denote $\mathrm{vec}(\mathbf{M}) = (\boldsymbol{M}_1^\top, \dots, \boldsymbol{M}_n^\top)^\top$, where $\boldsymbol{M}_j$ is the jth column of $\mathbf{M}$. Write $\mathbf{M} \succ 0$ when $\mathbf{M}$ is positive definite. Denote the trace and the determinant of a square matrix $\mathbf{M}$ by $\mathrm{tr}(\mathbf{M})$ and $\det(\mathbf{M})$, respectively. Denote the Kronecker product by $\otimes$. Write $x_n \asymp y_n$ if $x_n = O(y_n)$ and $y_n = O(x_n)$, and denote $x_n \sim y_n$ if $x_n/y_n \to 1$ as $n \to \infty$. Define $\lceil x \rceil$ and $\lfloor x \rfloor$ to be the smallest integer $\geq x$ and the largest integer $\leq x$, respectively. Let $\mathbb{1}(A)$ be the indicator function of event A, $x_+ = x\mathbb{1}(x \geq 0)$ and $\mathrm{sign}(x) = \mathbb{1}(x \geq 0) - \mathbb{1}(x \leq 0)$. Let $A := B$ denote that A is defined to be B. Denote $X \overset{d}{=} Y$ if X and Y have the same distribution. Denote $\mathbf{1}_n = (1, 1, \dots, 1)^\top$ with length n and $\mathbf{I}_{n\times n}$ to be the $n\times n$ identity matrix. If without further notification, a constant is independent of n and p. Throughout the rest of the article, we assume $p \to \infty$ as $n \to \infty$ and only use $n \to \infty$ in the asymptotic arguments.

2.1. *Useful bounds for temporal dependence.* Let $\mathbf{X}_{p\times n} := (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$, where each column $\boldsymbol{X}_i$ follows a distribution with the same covariance matrix $\boldsymbol{\Sigma} = (\sigma_{k\ell})_{p\times p}$ and correlation matrix $\mathbf{R} = (\rho_{k\ell})_{p\times p}$. Let $\boldsymbol{X}_{[1]}, \dots, \boldsymbol{X}_{[p]}$ be the p row vectors of $\mathbf{X}_{p\times n}$, and $\mathbf{R}_{[k]} = (\rho_{[k]}^{ij})_{n\times n}$ be the correlation matrix of $\boldsymbol{X}_{[k]}$, that is, the autocorrelation matrix of the kth time series. For all k, we have the following inequalities:

$$1 \leq \frac{1}{n}\|\mathbf{R}_{[k]}\|_F^2 \leq \|\mathbf{R}_{[k]}\|_2 \leq \|\mathbf{R}_{[k]}\|_1 \leq n, \tag{1}$$

where the second inequality follows from

$$\frac{1}{n}\mathrm{tr}(\mathbf{R}_{[k]}^2) = \frac{1}{n}\sum_{i=1}^n \varphi_i^2(\mathbf{R}_{[k]}) \leq \frac{1}{n}\varphi_{\max}(\mathbf{R}_{[k]})\sum_{i=1}^n \varphi_i(\mathbf{R}_{[k]}) = \frac{1}{n}\|\mathbf{R}_{[k]}\|_2 \,\mathrm{tr}(\mathbf{R}_{[k]}),$$

and the third inequality is obtained from Corollary 2.3.2 in [37].

We quantify the temporal dependence of $\boldsymbol{X}_{[k]}$ using the Frobenius norm and the spectral norm of its autocorrelation matrix $\mathbf{R}_{[k]}$, and define g_F and g_2 such that

$$\max_{1\leq k\leq p} \frac{1}{n}\|\mathbf{R}_{[k]}\|_F^2 \leq g_F, \qquad \max_{1\leq k\leq p} \|\mathbf{R}_{[k]}\|_2 \leq g_2. \tag{2}$$

From (1), we set $1 \leq g_F \leq g_2 \leq n$. Particularly, we can set $g_2 = 1$ if all the p time series are white noise processes, and $g_F = n$ if every pair of data points in one of the time series are perfectly correlated or anticorrelated. Later we will show that the convergence rates of considered estimators are nicely characterized by the bounds g_F and g_2, which is particularly useful in obtaining convergence results for long-memory data.

Note that we do not consider cross-correlations between the multiple time series, neither assume any specific temporal decay model or stationarity for autocorrelations within each individual time series. From the proofs provided in the Supplementary Material [65], we can see that those information does not contribute to the convergence rate calculations once g_F and g_2 are provided.

Here are two special examples of practical interests.

CASE 1 (High-dimensional short-memory dependence). Recall that a univariate stationary time series is said to be short-memory if its autocorrelation function satisfies $\sum_{t=0}^{\infty} |\rho(t)| < \infty$. We extend the "short-memory" concept to multivariate time series that are allowed to be nonstationary by the property

$$\max_{1 \le k \le p} \|\mathbf{R}_{[k]}\|_1 < \infty \qquad \text{as } n \to \infty. \tag{3}$$

Thus from (1) we can set $g_2 < \infty$ as $n \to \infty$.

CASE 2 (Polynomial-dominated decay (PDD) model). We say $\mathbf{X}_{p \times n}$ has PDD temporal dependence if

$$\max_{1 \le k \le p} |\rho_{[k]}^{ij}| \le C_0 |i - j|^{-\alpha} \qquad \text{for all } i \ne j \tag{4}$$

with some positive constants C_0 and α. We can then set the bounds

$$g_F = 2C_0^2 H_{\lceil n/2 \rceil}^{(2\alpha)} + 1 \quad \text{and} \quad g_2 = 2C_0 H_{\lceil n/2 \rceil}^{(\alpha)} + 1 \ge \max_{1 \le k \le p} \|\mathbf{R}_{[k]}\|_1 \tag{5}$$

with the generalized harmonic number (see (25), (26) in [26])

$$H_n^{(\alpha)} = \sum_{k=1}^{n} k^{-\alpha} < 1 + \begin{cases} \dfrac{n^{1-\alpha} - 1}{1 - \alpha}, & \alpha \ne 1; \\ \log n, & \alpha = 1. \end{cases} \tag{6}$$

The model is short-memory in the sense of (3) when $\alpha > 1$, but allows an individual time series to be long-memory when $0 < \alpha \le 1$. It is worth noting that the fractional Gaussian noise [52, 69] and the autoregressive fractionally integrated moving average process [38, 43] are classical examples of stationary univariate time series with autocorrelation function $\rho(t) \sim Ct^{-\alpha}$ as $t \to \infty$ with $C \ne 0$ and $\alpha \in (0, 1) \cup (1, 2)$.

2.2. *Comparisons to existing work.* For banding and tapering estimators of $\mathbf{\Sigma}$, [8] considered a weak temporal dependence $\max_{a_n \le |i-j| \le n} |\mathbf{\Lambda}^{ij}|_\infty = O(n^{-2}a_n)$, where $\mathbf{\Lambda}^{ij} = (\Lambda_{k\ell}^{ij})_{p \times p}$ with $\Lambda_{k\ell}^{ij}$ satisfying $\mathrm{Cov}(X_{ki}, X_{\ell j}) = \Lambda_{k\ell}^{ij} \sigma_{k\ell}$, $a_n \sqrt{\log p / n} = o(1)$, and $\{a_n\}_{n \ge 1}$ is a nondecreasing sequence of nonnegative integers. That $a_n \sqrt{\log p / n} = o(1)$ implies $a_n = o(\sqrt{n})$. Thus, $|\mathbf{\Lambda}^{ij} : |i - j| = \lfloor \sqrt{n} \rfloor|_\infty \le$

$\max_{a_n \le |i-j| \le n} |\mathbf{\Lambda}^{ij}|_\infty = O(n^{-2}a_n) = o(n^{-3/2})$. Then $\sum_{|i-j|=0}^{\infty} |\rho_{[k]}^{ij}| < \infty$ for any given k, which means that the time series cannot be long-memory, not even $\rho_{[k]}^{ij} \asymp |i-j|^{-\alpha}$ with $\alpha \in (0,3]$. Bhattacharjee and Bose [9] extended the banding and tapering techniques to the estimation of $\mathrm{Cov}(\boldsymbol{X}_j, \boldsymbol{X}_{j+k})$, $k \ge 0$, for the stationary infinite-order moving average model. It is easy to show that their time series cannot be long-memory.

Chen et al. [24] considered the hard thresholding estimation of $\mathbf{\Sigma}$ and an ℓ_1-MLE type estimation of $\mathbf{\Omega}$ using the functional dependence measure of [73]. Without loss of generality, assume that the first row of $\mathbf{X}_{p\times n}$, $\{X_{1t}\}$, is a stationary process with autocovariance function $\gamma_1(t)$. Following their setup by letting $E(X_{1t}) = 0$, we have $\gamma_1(t) = E(X_{11}X_{1,t+1})$. By the argument in the proof of Theorem 1 in [74] together with Lyapunov's inequality [12] and Theorem 1 of [73], one can see that their model requires $\sum_{t=0}^{\infty} |\gamma_1(t)| < \infty$, which indicates that $\{X_{1t}\}$ cannot be long-memory.

Zhou [77] considered estimating a separable covariance $\mathrm{Cov}(\boldsymbol{X}_{pn}) = \mathbf{A} \otimes \mathbf{B}$, where $\boldsymbol{X}_{pn} := \mathrm{vec}(\mathbf{X}_{p\times n})$. Her model implies the same autocorrelation coefficients $\{\rho_{[k]}^{ij}\}_{1\le i,j\le n}$ for all k, indicating a rather restrictive model with homogeneous decay rate for all p time series.

Now take a look at the rfMRI data example of a single subject which will be further analyzed in Section 6.3. The data set consists of 1190 temporal brain images. We consider 907 functional brain nodes in each image. All node time series have passed the Priestley–Subba Rao test for stationarity [57], the generalized Jarque–Bera test for Gaussianity [2] and the Hinich's bispectral test for linearity [42]. Hence the linear spatiotemporal model that we will define in the next subsection with sub-Gaussian tails seems adequate for the data. There are 134 time series detected as long-memory by the GPH test [36]. All these tests are conducted with a significant level of 0.05 for the p-values adjusted by the false discovery rate controlling procedure of [7]. Hence, the weak temporal dependence models of [8, 9, 24] do not apply to these long-memory time series. For node k, its autocorrelation function $\rho_k(t) := \rho_{[k]}^{ij}$, $t = |i-j|$, can be approximated by the sample autocorrelation function $\hat{\rho}_k(t)$. Figure 1 shows that the rfMRI data approximately satisfy the PDD model (4) with $C_0 = 1$ and $\alpha = 0.30$ since $\max_{1\le k\le p} |\hat{\rho}_k(t)| \le t^{-0.30}$. The figure also illustrates the estimated autocorrelation functions for two randomly selected brain nodes, which clearly have different patterns, indicating that the assumption of homogeneous decay rates for all time series in [77] does not hold.

2.3. *Data generating mechanism.* Throughout the article, we assume that the vectorized data are obtained from the linear spatiotemporal model

$$\boldsymbol{X}_{pn} := \mathrm{vec}(\mathbf{X}_{p\times n}) = \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}, \tag{7}$$

where $\mathbf{H} = (h_{ij})_{pn\times m}$ is a real deterministic matrix, $\boldsymbol{\mu}_{pn} = \mathbf{1}_n \otimes \boldsymbol{\mu}_p$, and the random vector $\boldsymbol{e} = (e_1, \dots, e_m)^\top$ consists of m independent (not necessarily i.i.d.)

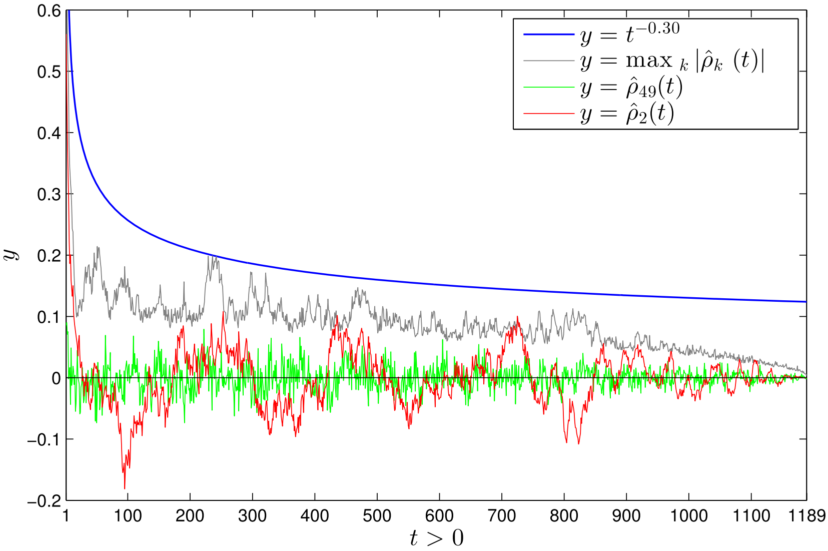

FIG. 1. *Sample autocorrelations of brain nodes.*

random variables satisfying $E(e_i) = 0$ and $E(e_i^2) = 1$ for all i. We allow $m = \infty$ by requiring that for each i, $\sum_{j=1}^{m} h_{ij} e_j$ converges both almost surely and in mean square when $m \to \infty$. A sufficient and necessary condition for both modes of convergence is $\sum_{j=1}^{\infty} h_{ij}^2 < \infty$ for every i (see Theorem 8.3.4 and its proof in [1]). Under these two modes of convergence, it can be shown that $E(\mathbf{H}\boldsymbol{e}) = \mathbf{H}E(\boldsymbol{e})$ and $\operatorname{Cov}(\mathbf{H}\boldsymbol{e}) = \mathbf{H}\operatorname{Cov}(\boldsymbol{e})\mathbf{H}^\top$ (see Proposition 2.7.1 in [15]). Hence, for either finite or infinite m, we have $E(\boldsymbol{X}_{pn}) = \boldsymbol{\mu}_{pn}$ and $\operatorname{Cov}(\boldsymbol{X}_{pn}) = \mathbf{H}\mathbf{H}^\top$ with all n submatrices of dimension $p \times p$ on the diagonal equal to $\boldsymbol{\Sigma}$ and temporal correlations determined by the off-diagonal submatrices. In filtering theory, matrix $\mathbf{H}$ is said to be a linear spatiotemporal coloring filter [33, 53], which generates the output $\boldsymbol{X}_{pn}$ by introducing both spatial and temporal dependence in the input independent variables $e_1, \dots, e_m$. We will use $\mathbf{X}_{p\times n}$ and $\boldsymbol{X}_{pn}$ exchangeably.

The following are two examples of (7) which are often seen in the literature. In particular, two processes used in analyzing fMRI data, that is, the multivariate fractional Gaussian noise [27] and the vector autoregressive model [39], are special cases of these two examples, respectively.

EXAMPLE 1 (Gaussian data). Assume that $\boldsymbol{X}_{pn}$ has a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{pn}, \boldsymbol{\Delta})$. Then $\boldsymbol{\Delta} = \mathbf{H}\mathbf{H}$ with a symmetric real matrix $\mathbf{H}$. If $\boldsymbol{\Delta} \succ 0$, then $\boldsymbol{X}_{pn} = \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}$ with $\boldsymbol{e} = \mathbf{H}^{-1}(\boldsymbol{X}_{pn} - \boldsymbol{\mu}_{pn}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{pn\times pn})$. If $\boldsymbol{\Delta}$ is singular, then $\boldsymbol{X}_{pn}$ has a degenerate multivariate Gaussian distribution, and can be expressed as $\boldsymbol{X}_{pn} \overset{d}{=} \mathbf{H}\boldsymbol{e} + \boldsymbol{\mu}_{pn}$ with any $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{pn\times pn})$. In fact, replacing "$=$" in (7) by "$\overset{d}{=}$" does not affect the theoretical results.

EXAMPLE 2 (Moving average processes). Consider the processes

$$X_j = \sum_{\ell=0}^{L} \mathbf{B}_\ell e_{j-\ell} \qquad \text{with } 0 \le L \le \infty, \tag{8}$$

where the case with $L = \infty$ is well-defined in the sense of entrywise almost-sure convergence and mean-square convergence, $\{\mathbf{B}_\ell\}$ are $p \times p$ real deterministic matrices and $e_j = (e_{1j}, e_{2j}, \dots, e_{pj})^\top$ is a vector with independent zero-mean and unit-variance entries $\{e_{ij}, 1 \le i \le p, -\infty \le j \le n\}$. Since every X_{ij} is a linear combination of $\{e_{st}\}$, we always can find a matrix $\mathbf{H}$ such that $X_{pn} = \mathbf{H}e$ with $e = (e_{1-L}^\top, e_{2-L}^\top, \dots, e_n^\top)^\top$. It is well known that any causal vector autoregressive moving average process of the form $X_j - \mathbf{A}_1 X_{j-1} - \cdots - \mathbf{A}_a X_{j-a} = e_j + \mathbf{M}_1 e_{j-1} + \cdots + \mathbf{M}_b e_{j-b}$ with finite nonnegative integers a and b, and real deterministic matrices $\{\mathbf{A}_i, \mathbf{M}_k\}$, can be written in the form of (8) with $L = \infty$ (see page 418 in [15]). Model (8) with $L = \infty$ is widely studied in recent literature of high dimensional time series (see, e.g., [9, 24, 25, 50]).

We will consider the following three types of moment conditions for the basis random variables $e_1, \dots, e_m$ in (7), corresponding to sub-Gaussian, generalized subexponential, and polynomial-type tails, respectively. Let Z be a random variable, and K, ϑ and η_k be positive constants.

(C1) *Sub-Gaussian tails*: For all $k \ge 1$, we have $(E|Z|^k)^{1/k} \le K k^{1/2}$.

(C2) *Generalized subexponential tails*: For some $\vartheta \in (0, 2)$ and all $k \ge \vartheta$, we have $(E|Z|^k)^{1/k} \le K(k/\vartheta)^{1/\vartheta}$.

(C3) *Polynomial-type tails*: For some $k \ge 4$, we have $(E|Z|^k)^{1/k} \le \eta_k$.

We do not consider condition (C1) as a special case of condition (C2) by setting $\vartheta = 2$ due to the fact that sharper convergence rates can be obtained under (C1). We can apply the Hanson–Wright inequality to (C1) [63], but are not able to extend it to (C2) because the moment generating function of Z^2 is no longer finite in an open interval around zero (see Proposition 7.23 and inequality (7.32) in [34], and Lemma 5.5 in [71]). Conditions (C1) and (C2) can be equivalently written as $P(|Z| \ge u) \le 2\exp(-u^\vartheta/C)$ with some constant $C > 0$ for all $u \ge 0$, where for the former $\vartheta = 2$. Condition (C3) implies $P(|Z| \ge u) \le \eta_k^k/u^k$ for all $u > 0$. Conversely, if $P(|Z| \ge u) = O(u^{-k'})$ with some $k' \in (0, k)$ as $u \to \infty$, then $(E|Z|^k)^{1/k} < \infty$.

3. Estimation of covariance and correlation matrices for sub-Gaussian data.

Consider the ℓ_q-ball sparse covariance matrices [10, 61]

$$\mathcal{U}(q, c_p, v_0) = \left\{ \mathbf{\Sigma} : \max_{1\le i\le p} \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_p, \max_{1\le i\le p} \sigma_{ii} \le v_0 \right\}, \tag{9}$$

and the corresponding correlation matrices

$$\mathcal{R}(q, c_p) = \left\{ \mathbf{R} : \max_{1 \le i \le p} \sum_{j=1}^{p} |\rho_{ij}|^q \le c_p \right\}, \tag{10}$$

where constants $v_0 > 0$ and $0 \le q < 1$. For any thresholding parameter $\tau \ge 0$, define a generalized thresholding function [61] by $s_\tau : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions for all $z \in \mathbb{R}$: (i) $|s_\tau(z)| \le |z|$; (ii) $s_\tau(z) = 0$ for $|z| \le \tau$; (iii) $|s_\tau(z) - z| \le \tau$. Such defined generalized thresholding function covers many widely used thresholding functions, including hard thresholding $s_\tau^H(z) = z\mathbb{1}(|z| > \tau)$, soft thresholding $s_\tau^S(z) = \mathrm{sign}(z)(|z| - \tau)_+$, smoothly clipped absolute deviation and adaptive lasso thresholdings. See details about these examples in [61]. We define the generalized thresholding estimators of $\mathbf{\Sigma}$ and $\mathbf{R}$, respectively, by

$$S_{\tau_1}(\widehat{\mathbf{\Sigma}}) = \big(s_{\tau_1}(\hat{\sigma}_{ij})\big)_{p \times p} \quad \text{and} \quad S_{\tau_2}(\widehat{\mathbf{R}}) = \big(s_{\tau_2}(\hat{\rho}_{ij})\mathbb{1}(i \ne j) + \mathbb{1}(i = j)\big)_{p \times p},$$

where $\widehat{\mathbf{\Sigma}} := (\hat{\sigma}_{ij})_{p \times p}$ is the sample covariance matrix given by

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top - \bar{\boldsymbol{X}} \bar{\boldsymbol{X}}^\top \tag{11}$$

with $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$, and $\widehat{\mathbf{R}} := (\hat{\rho}_{ij})_{p \times p} = (\hat{\sigma}_{ij}/\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}})_{p \times p}$ is the sample correlation matrix. Define

$$u_1 = \max\{(\log p) g_2/n, [(\log p) g_F/n]^{1/2}\}. \tag{12}$$

Then we have the following results.

THEOREM 1. *Suppose that $\mathbf{X}_{p \times n}$ is generated from* (7) *with all e_i satisfying condition* (C1) *with the same K. Uniformly on $\mathbf{\Sigma} \in \mathcal{U}(q, c_p, v_0)$ and $\{\mathbf{R}_{[k]}\}_{k=1}^{p}$ subject to* (2), *for sufficiently large constant $M_1 > 0$ depending only on v_0 and K, if $\tau_1 = M_1 u_1$ and $u_1 = o(1)$, then*

$$|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty = O_P(u_1),$$

$$\|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2 = O_P(c_p u_1^{1-q}), \tag{13}$$

$$\frac{1}{p}\|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2 = O_P(c_p u_1^{2-q}). \tag{14}$$

Moreover, if $p \ge n^c$ for some constant $c > 0$, then with sufficiently large M_1 additionally depending on c and q, we have

$$E(|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|_\infty^2) = O(u_1^2),$$

$$E(\|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_2^2) = O(c_p^2 u_1^{2-2q}),$$

$$\frac{1}{p}E(\|S_{\tau_1}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F^2) = O(c_p u_1^{2-q}).$$

REMARK 1. When $(\log p)/n = o(1)$, if $u_1 = O(\sqrt{(\log p)/n})$, which is true when $g_2 < \infty$ that holds for short-memory multivariate time series satisfying (3), then the in-probability convergence rates given in (13) and (14) are the same rates as those for i.i.d. observations given in [10] and [61]. The same in-probability convergence rates are also obtained by [6], Proposition 5.1, for certain short-memory stationary Gaussian data using the hard thresholding method.

REMARK 2. For the PDD temporal dependence given in (4), by (5) and (6), together with $u_1 = o(1)$, we have

$$u_1 \lesssim \begin{cases} \sqrt{(\log p)/n}, & \alpha > 1; \\ \max\{[(\log p)(\log n)]/n, \sqrt{(\log p)/n}\}, & \alpha = 1; \\ \max\{(\log p)/n^{\alpha}, \sqrt{(\log p)/n}\}, & \alpha \in (1/2, 1); \\ \max\{(\log p)/n^{1/2}, \sqrt{[(\log p)(\log n)]/n}\}, & \alpha = 1/2; \\ (\log p)/n^{\alpha}, & \alpha \in (0, 1/2). \end{cases}$$

Here, $x_n \lesssim y_n$ denotes $x_n = O(y_n)$. Note that the case with $\alpha > 1$ is short-memory in the sense of (3). When $\alpha = 1$ and $(\log n)\sqrt{(\log p)/n} = O(1)$, or when $\alpha \in (1/2, 1)$ and $(\log p)^{1/2} n^{1/2-\alpha} = O(1)$, which allows some individual time series to be long-memory, we also have $u_1 = O(\sqrt{(\log p)/n})$, yielding the same convergence rates as in the case with i.i.d. data.

THEOREM 2 (Sparsistency and sign-consistency). *Under the conditions for the convergence in probability given in Theorem 1, we have $s_{\tau_1}(\hat{\sigma}_{ij}) = 0$ for all (i, j) where $\sigma_{ij} = 0$ with probability tending to 1. If further assume all nonzero entries of $\mathbf{\Sigma}$ satisfy $|\sigma_{ij}| \geq 2\tau_1$, then we have $\operatorname{sign}(s_{\tau_1}(\hat{\sigma}_{ij})) = \operatorname{sign}(\sigma_{ij})$ for all (i, j) where $\sigma_{ij} \neq 0$ with probability tending to 1.*

COROLLARY 1. *Theorems 1 and 2 hold with $\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}, \hat{\sigma}_{ij}, \sigma_{ij}, \mathcal{U}(q, c_p, v_0), \tau_1$, and M_1 replaced by $\widehat{\mathbf{R}}, \mathbf{R}, \hat{\rho}_{ij}, \rho_{ij}, \mathcal{R}(q, c_p), \tau_2$, and M_2, respectively, where M_2 does not depend on v_0.*

We now provide the minimax optimal rates for estimating covariance and correlation matrices over certain sets of distributions of $\mathbf{X}_{p\times n}$, including the short-memory case (3) and some long-memory cases (4) with $\alpha \in (1/2, 1]$.

THEOREM 3 (Minimax rates). *Let $K_G = \sup_{k\geq 1} \sqrt{2/k}[\Gamma(\frac{1+k}{2})/\sqrt{\pi}]^{1/k}$, where $\Gamma(x)$ is the gamma function. Let $\mathcal{P}_1(q, c_p, v_0, g_F, g_2, K, \kappa)$ be the set of distributions of $\mathbf{X}_{p\times n}$ generated from (7) with all e_i satisfying (C1) with constant $K \geq K_G$, $\mathbf{\Sigma} \in \mathcal{U}(q, c_p, v_0)$, and $\{\mathbf{R}_{[k]}\}_{k=1}^{p}$ subject to (2), where the constant $\kappa \geq 1$*

is used in setting $u_1 \le \kappa\sqrt{(\log p)/n}$. Let $\mathcal{P}_2(q, c_p, g_F, g_2, K, \kappa)$ be the corresponding set to $\mathcal{P}_1$ with $\mathbf{\Sigma} \in \mathcal{U}(q, c_p, v_0)$ replaced by $\mathbf{R} \in \mathcal{R}(q, c_p)$. Let $\mathfrak{D}$ denote the distribution of $\mathbf{X}_{p\times n}$. If $\sqrt{(\log p)/n} = o(1)$,

$$p \ge n^{c_1} \quad \text{and} \quad c_p \le c_2 n^{(1-q)/2}(\log p)^{-(3-q)/2} \tag{15}$$

with some constants $c_1 > 1$ and $c_2 > 0$, then for any estimator $\widetilde{\mathbf{\Sigma}}$ we have

$$\inf_{\widetilde{\mathbf{\Sigma}}} \sup_{\mathfrak{D}\in\mathcal{P}_1} E_{\mathbf{X}_{p\times n}|\mathfrak{D}}(\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2^2) \asymp c_p^2\left(\frac{\log p}{n}\right)^{1-q},$$

$$\inf_{\widetilde{\mathbf{\Sigma}}} \sup_{\mathfrak{D}\in\mathcal{P}_1} \frac{1}{p} E_{\mathbf{X}_{p\times n}|\mathfrak{D}}(\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2) \asymp c_p\left(\frac{\log p}{n}\right)^{1-q/2}.$$

Additionally if $c_p > 1$, then for any estimator $\widetilde{\mathbf{R}}$ we have

$$\inf_{\widetilde{\mathbf{R}}} \sup_{\mathfrak{D}\in\mathcal{P}_2} E_{\mathbf{X}_{p\times n}|\mathfrak{D}}(\|\widetilde{\mathbf{R}} - \mathbf{R}\|_2^2) \asymp c_p^2\left(\frac{\log p}{n}\right)^{1-q},$$

$$\inf_{\widetilde{\mathbf{R}}} \sup_{\mathfrak{D}\in\mathcal{P}_2} \frac{1}{p} E_{\mathbf{X}_{p\times n}|\mathfrak{D}}(\|\widetilde{\mathbf{R}} - \mathbf{R}\|_F^2) \asymp c_p\left(\frac{\log p}{n}\right)^{1-q/2}.$$

For sufficiently large positive constants M_1 and M_2, with $\tau_1 = M_1 u_1$ and $\tau_2 = M_2 u_1$, the generalized thresholding estimators $S_{\tau_1}(\widehat{\mathbf{\Sigma}})$ and $S_{\tau_2}(\widehat{\mathbf{R}})$ attain the above minimax optimal rates, respectively.

The assumption (15) follows [22] who studied the minimax optimal rates of the covariance matrix estimation for i.i.d. data. From Remarks 1 and 2, we see that suitable (g_F, g_2, κ) in $\mathcal{P}_1$ and $\mathcal{P}_2$ allow $\mathbf{X}_{p\times n}$ to have short memory (3), or to follow the PDD model (4) with $\alpha \in (1/2, 1]$ (thus with some time series to be long-memory) under some additional conditions for n and p discussed in Remark 2.

4. Estimation of precision matrix for sub-Gaussian data.

We consider both the CLIME and the SPICE methods for the estimation of $\mathbf{\Omega}$, which were originally developed for i.i.d. observations.

4.1. *CLIME estimation.*

Following [18], we consider the following set of precision matrices:

$$\mathcal{G}_1(q, c_p, M_p, v_0) = \left\{\mathbf{\Omega} \succ 0 : \max_{1\le i\le p}\sum_{j=1}^{p}|\omega_{ij}|^q \le c_p, \|\mathbf{\Omega}\|_1 \le M_p, \max_{1\le i\le p}\sigma_{ii} \le v_0\right\},$$

where constant $0 \le q < 1$, and (c_p, M_p) are allowed to depend on p. Though the condition $\max_i \sigma_{ii} \le v_0$ is not explicitly provided by [18] in their original set definition, it is implied by their moment conditions [see their (C1) and (C2)]. Note that the above $\mathcal{G}_1$ contains ℓ_q-ball sparse matrices such as those with exponentially decaying entries from the diagonal, for example, AR(1) matrices. For an invertible band matrix $\mathbf{\Sigma}$, its inverse matrix $\mathbf{\Omega}$ generally has exponentially decaying entries from the diagonal [30].

Let $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda_1} := (\hat{\theta}_{ij}^{(\varepsilon,\lambda_1)})_{p\times p}$ be a solution of the following optimization:

$$\min |\mathbf{\Theta}|_1 \quad \text{subject to} \quad |\widetilde{\mathbf{\Sigma}}_\varepsilon \mathbf{\Theta} - \mathbf{I}_{p\times p}|_\infty \le \lambda_1, \qquad \mathbf{\Theta} \in \mathbb{R}^{p\times p}, \tag{16}$$

where $\widetilde{\mathbf{\Sigma}}_\varepsilon = \widehat{\mathbf{\Sigma}} + \varepsilon \mathbf{I}_{p\times p}$, $\widehat{\mathbf{\Sigma}}$ is given in (11), $\varepsilon \ge 0$ is a perturbation parameter introduced for the same reasons given in [18] and can be set to be $n^{-1/2}$ in practice (see Remark 4 below), and λ_1 is a tuning parameter. The CLIME estimator $\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} := (\hat{\omega}_{ij}^{(\varepsilon,\lambda_1)})_{p\times p}$ is then obtained by symmetrizing $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda_1}$ with

$$\hat{\omega}_{ij}^{(\varepsilon,\lambda_1)} = \hat{\omega}_{ji}^{(\varepsilon,\lambda_1)} = \hat{\theta}_{ij}^{(\varepsilon,\lambda_1)} \mathbb{1}\big(|\hat{\theta}_{ij}^{(\varepsilon,\lambda_1)}| \le |\hat{\theta}_{ji}^{(\varepsilon,\lambda_1)}|\big) + \hat{\theta}_{ji}^{(\varepsilon,\lambda_1)} \mathbb{1}\big(|\hat{\theta}_{ij}^{(\varepsilon,\lambda_1)}| > |\hat{\theta}_{ji}^{(\varepsilon,\lambda_1)}|\big).$$

For $1 \le i \le p$, let $\hat{\boldsymbol{\beta}}_i^{(\varepsilon,\lambda_1)}$ be a solution of the following convex optimization problem:

$$\min |\boldsymbol{\beta}_i|_1 \quad \text{subject to} \quad |\widetilde{\mathbf{\Sigma}}_\varepsilon \boldsymbol{\beta}_i - \boldsymbol{e}_i|_\infty \le \lambda_1, \tag{17}$$

where $\boldsymbol{\beta}_i$ is a real vector and $\boldsymbol{e}_i$ is the vector with 1 in the ith coordinate and 0 in all other coordinates. Cai et al. [18] showed that solving the optimization problem (16) is equivalent to solving the p optimization problems given in (17), that is, $\widehat{\mathbf{\Theta}}_{\varepsilon,\lambda_1} = (\hat{\boldsymbol{\beta}}_1^{(\varepsilon,\lambda_1)}, \dots, \hat{\boldsymbol{\beta}}_p^{(\varepsilon,\lambda_1)})$. This equivalence is useful for both numerical implementation and theoretical analysis. The following theorem gives the convergence results of CLIME under temporal dependence.

THEOREM 4. *Suppose that $\mathbf{X}_{p\times n}$ is generated from (7) with all e_i satisfying condition (C1) with the same K. Uniformly on $\mathbf{\Omega} \in \mathcal{G}_1(q, c_p, M_p, v_0)$ and $\{\mathbf{R}_{[k]}\}_{k=1}^p$ subject to (2), for sufficiently large constant $M > 0$ depending only on v_0 and K, if $\lambda_1 = MM_pu_1$, $0 \le \varepsilon \le u_1$ and $u_1 = o(1)$ with u_1 defined in (12), then*

$$|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}|_\infty = O_P(M_p^2 u_1),$$

$$\|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}\|_2 = O_P\big(c_p (M_p^2 u_1)^{1-q}\big),$$

$$\frac{1}{p}\|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}\|_F^2 = O_P\big(c_p (M_p^2 u_1)^{2-q}\big).$$

Moreover, if $p \ge n^c$ and $\min\{p^{-C}, u_1\} \le \varepsilon \le u_1$ for some positive constants c and C, then with sufficiently large M additionally depending on c, C and q, we have

$$E\big(|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}|_\infty^2\big) = O\big((M_p^2 u_1)^2\big),$$

$$E\big(\|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}\|_2^2\big) = O\big(c_p^2 (M_p^2 u_1)^{2-2q}\big), \tag{18}$$

$$\frac{1}{p}E\big(\|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}\|_F^2\big) = O\big(c_p (M_p^2 u_1)^{2-q}\big). \tag{19}$$

REMARK 3. If $(\log p)/n = o(1)$ and $u_1 = O(\sqrt{(\log p)/n})$, then the above convergence rates are the same as those for i.i.d. data given in [18]. Additionally, if M_p is a constant, then the mean-square convergence rates of CLIME in (18) and (19) attain the minimax optimal convergence rates for i.i.d. data shown in [19] under slightly different assumptions. From Remarks 1 and 2, we see that $u_1 = O(\sqrt{(\log p)/n})$ when $u_1 = o(1)$ can be achieved for the short-memory case (3) and also for the long-memory cases satisfying (4) with $\alpha \in (1/2, 1]$ under some additional conditions for n and p.

REMARK 4. As discussed in [18], the perturbation parameter $\varepsilon > 0$ is used for a proper initialization of $\{\boldsymbol{\beta}_i\}$ in the numerical implementation of (17), and also ensures the existence of $E(\|\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1} - \mathbf{\Omega}\|_2^2)$. Since $g_F \geq 1$, from (12) we have that $u_1 \geq \sqrt{(\log p)/n} \geq n^{-1/2}$. Hence, when $p \geq n^c$, letting $C = 1/(2c)$, we have $p^{-C} \leq n^{-1/2} \leq u_1$. Thus, we can simply let $\varepsilon = n^{-1/2}$ in practice, which is also the default setting of the R package `flare` [49] that implements the CLIME algorithm. The same choice of ε is given in (10) of [18] for i.i.d. observations.

To better recover the sparsity structure of $\mathbf{\Omega}$, [18] introduced additional thresholding on $\widehat{\mathbf{\Omega}}_{\varepsilon,\lambda_1}$. Similarly, we may define a hard-thresholded CLIME estimator $\widetilde{\mathbf{\Omega}}_{\varepsilon,\lambda_1,\xi} = (\tilde{\omega}_{ij}^{(\varepsilon,\lambda_1,\xi)})_{p\times p}$ by $\tilde{\omega}_{ij}^{(\varepsilon,\lambda_1,\xi)} = \hat{\omega}_{ij}^{(\varepsilon,\lambda_1)}\mathbb{1}(|\hat{\omega}_{ij}^{(\varepsilon,\lambda_1)}| > \xi)$ with a tuning parameter $\xi \geq 4M_p\lambda_1$. Although such an estimator enjoys nice theoretical properties given below, how to practically select ξ remains unknown.

THEOREM 5 (Sparsistency and sign-consistency). *Under the conditions for the convergence in probability given in Theorem 4, we have* $\tilde{\omega}_{ij}^{(\varepsilon,\lambda_1,\xi)} = 0$ *for all* (i, j) *where* $\omega_{ij} = 0$ *with probability tending to* 1. *If further assume all nonzero entries of* $\mathbf{\Omega}$ *satisfy* $|\omega_{ij}| > \xi + 4M_p\lambda_1$, *then we have* $\operatorname{sign}(\tilde{\omega}_{ij}^{(\varepsilon,\lambda_1,\xi)}) = \operatorname{sign}(\omega_{ij})$ *for all* (i, j) *where* $\omega_{ij} \neq 0$ *with probability tending to* 1.

4.2. *SPICE estimation.* For i.i.d. observations, [60] proposed the SPICE method for estimating the following precision matrix $\mathbf{\Omega}$:

$$\mathcal{G}_2(s_p, v_0) = \Bigg\{\mathbf{\Omega} : \sum_{1\leq i\neq j\leq p} \mathbb{1}(\omega_{ij} \neq 0) \leq s_p, 0 < v_0^{-1} \leq \varphi_{\min}(\mathbf{\Omega}) \leq \varphi_{\max}(\mathbf{\Omega}) \leq v_0\Bigg\},$$

where s_p determines the sparsity of $\mathbf{\Omega}$ and can depend on p, and v_0 is a constant. Two types of SPICE estimators were proposed:

$$\widetilde{\mathbf{\Omega}}_{\lambda_2} = \mathop{\arg\min}_{\mathbf{\Theta}\succ 0, \mathbf{\Theta}=\mathbf{\Theta}^\top} \{\operatorname{tr}(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}) - \log\det(\mathbf{\Theta}) + \lambda_2|\mathbf{\Theta}|_{1,\mathrm{off}}\}, \tag{20}$$

and

$$\begin{aligned} \widehat{\mathbf{\Omega}}_{\lambda_2} &:= (\hat{\omega}_{ij}^{(\lambda_2)})_{p\times p} = \widehat{\mathbf{W}}^{-1}\widehat{\mathbf{K}}_{\lambda_2}\widehat{\mathbf{W}}^{-1} \\ &\quad \text{with } \widehat{\mathbf{K}}_{\lambda_2} = \mathop{\arg\min}_{\mathbf{\Theta}\succ 0, \mathbf{\Theta}=\mathbf{\Theta}^\top} \{\operatorname{tr}(\mathbf{\Theta}\widehat{\mathbf{R}}) - \log\det(\mathbf{\Theta}) + \lambda_2|\mathbf{\Theta}|_{1,\mathrm{off}}\}, \end{aligned} \tag{21}$$

where $\lambda_2 > 0$ is a tuning parameter, and $\widehat{\mathbf{W}} = \operatorname{diag}\{\sqrt{\hat{\sigma}_{11}}, \dots, \sqrt{\hat{\sigma}_{pp}}\}$. We can see that $\widehat{\mathbf{K}}_{\lambda_2}$ is the SPICE estimator of $\mathbf{K} := \mathbf{R}^{-1}$. The SPICE estimator (20) is a slight modification of the graphical Lasso (GLasso) estimator of [35]. GLasso uses $|\boldsymbol{\Omega}|_1$ rather than $|\boldsymbol{\Omega}|_{1,\mathrm{off}}$ in the penalty, but the SPICE estimators (20) and (21) are more amenable to theoretical analysis [48, 59, 60], and numerically they give similar results for i.i.d. data [60]. It is worth noting that for i.i.d. data, (20) requires $\sqrt{(p + s_p)(\log p)/n} = o(1)$ but (21) relaxes it to $\sqrt{(1 + s_p)(\log p)/n} = o(1)$. Similar requirements also hold for temporally dependent observations. Hence, in this article, we only consider the SPICE estimator given in (21).

THEOREM 6. *Suppose that $\mathbf{X}_{p\times n}$ is generated from (7) with all e_i satisfying condition (C1) with the same K. Uniformly on $\boldsymbol{\Omega} \in \mathcal{G}_2(s_p, v_0)$ and $\{\mathbf{R}_{[k]}\}_{k=1}^p$ subject to (2), for sufficiently large constant $M > 0$ depending only on v_0 and K, if $\lambda_2 = Mu_1$ and $u_1 = o(1/\sqrt{1 + s_p})$ with u_1 defined in (12), then we have*

$$\|\widehat{\mathbf{K}}_{\lambda_2} - \mathbf{K}\|_F = O_P(u_1\sqrt{s_p}),$$

$$\|\widehat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_2 = O_P(u_1\sqrt{1 + s_p}),$$

$$\frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{\Omega}}_{\lambda_2} - \boldsymbol{\Omega}\|_F = O_P(u_1\sqrt{1 + s_p/p}).$$

Again by Remarks 1 and 2, $u_1 = O(\sqrt{(\log p)/n})$ is achievable for the short-memory case (3) and also for some long-memory cases, thus for such temporally dependent data Theorem 6 gives the same convergence rates as those given in [60] for i.i.d. observations. The condition $u_1 = o(1/\sqrt{1 + s_p})$ implies $s_p = o(u_1^{-2}) = o(n/\log p)$, meaning that $\boldsymbol{\Omega}$ needs to be very sparse. Such a condition easily fails for many simple band matrices when $p \geq n$.

Under the irrepresentability condition, however, the sparsity requirement can be relaxed [59]. In particular, define $\boldsymbol{\Gamma} = \mathbf{R} \otimes \mathbf{R}$. By (i, j)th row of $\boldsymbol{\Gamma}$ we refer to its $[i + (j - 1)p]$th row, and by (k, ℓ)th column to its $[k + (\ell - 1)p]$th column. For any two subsets T and T' of $\{1, \dots, p\}\times\{1, \dots, p\}$, denote $\boldsymbol{\Gamma}_{TT'}$ be the $\operatorname{card}(T)\times\operatorname{card}(T')$ matrix with rows and columns of $\boldsymbol{\Gamma}$ indexed by T and T', respectively, where $\operatorname{card}(T)$ denotes the cardinality of set T. Let S be the set of nonzero entries of $\boldsymbol{\Omega}$ and S^c be the complement of S in $\{1, \dots, p\}\times\{1, \dots, p\}$. Define $\kappa_{\mathbf{R}} = \|\mathbf{R}\|_1$ and $\kappa_{\boldsymbol{\Gamma}} = \|\boldsymbol{\Gamma}_{SS}^{-1}\|_1$. Assume the following irrepresentability condition of [59]:

$$\max_{e\in S^c}|\boldsymbol{\Gamma}_{eS}\boldsymbol{\Gamma}_{SS}^{-1}|_1 \leq 1 - \beta \tag{22}$$

for some $\beta \in (0, 1]$. Define d to be the maximum number of nonzeros per row in $\boldsymbol{\Omega}$. Then we have the following result.

THEOREM 7. *Let $r = (0.5 + 2.5(1 + 8/\beta)\kappa_{\mathbf{\Gamma}})Mu_1v_0$, where u_1 is defined in* (12). *Suppose that $\mathbf{X}_{p\times n}$ is generated from* (7) *with all e_i satisfying condition* (C1) *with the same K. Uniformly on $\mathbf{\Omega} \in \mathcal{G}_2(s_p, v_0)$ and $\{\mathbf{R}_{[k]}\}_{k=1}^p$ subject to* (2), *for sufficiently large constant $M > 0$ depending on v_0 and K, if $\lambda_2 = 8Mu_1/\beta \le [6(1 + \beta/8)d\max\{\kappa_{\mathbf{R}}\kappa_{\mathbf{\Gamma}}, \kappa_{\mathbf{R}}^3\kappa_{\mathbf{\Gamma}}^2\}]^{-1}$ and $u_1 = o(\min\{1, [(1 + 8/\beta)\kappa_{\mathbf{\Gamma}}]^{-1}\})$, then with probability tending to* 1 *we have*

$$|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}|_\infty \le r,$$

$$\|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 \le r\min\{d, \sqrt{p + s_p}\},$$

$$\frac{1}{\sqrt{p}}\|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_F \le r\sqrt{1 + s_p/p},$$

and $\hat{\omega}_{ij}^{(\lambda_2)} = 0$ for all (i, j) with $\omega_{ij} = 0$. If we further assume all nonzero entries of $\mathbf{\Omega}$ satisfy $|\omega_{ij}| > r$, then with probability tending to 1, $\operatorname{sign}(\hat{\omega}_{ij}^{(\lambda_2)}) = \operatorname{sign}(\omega_{ij})$ *for all (i, j) where $\omega_{ij} \ne 0$.*

Consider the case when β remains constant and $\max\{\kappa_{\mathbf{R}}, \kappa_{\mathbf{\Gamma}}\}$ has a constant upper bound. Then the conditions in Theorem 7 about λ_2 and u_1 reduce to $\lambda_2 = M'u_1$ and $u_1 = o(1)$ with a constant $M' = 8M/\beta$, and meanwhile we have $\|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = O_P(u_1 d)$. Then the desired result of $\|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = o_P(1)$ is achieved under a relaxed sparsity condition $d = o(u_1^{-1})$. If $d^2 > 1 + s_p$, then $s_p = o(u_1^{-2})$ and the condition of Theorem 6 satisfies. Hence $\|\widehat{\mathbf{\Omega}}_{\lambda_2} - \mathbf{\Omega}\|_2 = O_P(u_1\sqrt{\min\{d^2, 1 + s_p\}}) = o_P(1)$, which is the better rate between those from Theorems 6 and 7.

5. Extension to heavy tail data. In this section, we generalize the theoretical results for sub-Gaussian data to the cases when all the basis random variables $\{e_i\}$ have the generalized subexponential tails under condition (C2) or the polynomial-type tails under condition (C3). Define

$$u_2 = \max\{(\log p)^{1+2/\vartheta} g_2/n, (\log p)^{1+2/\vartheta}(g_F/n)^{1/2}\}, \tag{23}$$

and

$$u_3 = \max\{p^{(2+2C)/k} g_2/n, p^{(4+2C)/k}(g_F/n)^{1/2}\} \tag{24}$$

with an arbitrary constant $C > 0$. The quantities u_2 and u_3 will substitute u_1 in characterizing the matrix estimation convergence rates under the tail conditions (C2) and (C3), respectively. The first term in either u_2 or u_3 can be dropped if $\boldsymbol{\mu}_p$ is known thus no need to be estimated.

THEOREM 8 (Generalized subexponential tails). *Theorems* 1, 2, 4–7 *and Corollary* 1 *hold with condition* (C1), *parameter K, and u_1 replaced by condition* (C2), *parameters $\{K, \vartheta\}$, and u_2, respectively.*

THEOREM 9 (Polynomial-type tails). *Theorems 1, 2, 4–7 and Corollary 1, except those mean-square convergence results therein, hold with condition (C1), parameter K, and u_1 replaced by condition (C3), parameters $\{k, \eta_k\}$, and u_3, respectively.*

For data with polynomial-type tails, the mean-square convergence results may require higher order moment conditions, thus are not pursued here.

A referee pointed out potential connections to the recent work of [25] and [72]. Chen et al. [25] considered the estimation of $\mathbf{\Omega} \boldsymbol{b}$ with $\boldsymbol{b} \in \mathbb{R}^p$ for high-dimensional mean-zero stationary processes given in (8) which satisfy PDD given in (4). Under the assumption that $\boldsymbol{\mu}_p$ is known, our exploration shows that their concentration inequalities for the true $\boldsymbol{\mu}_p$ centered sample covariance matrix (see page 3 of their Supplementary Material) can be used to obtain the same convergence rates for our concerned estimators to their sub-Gaussian time series. Their inequalities also can be applied to their time series with the generalized subexponential tails, but result in slower convergence rates. If applied to their time series data with polynomial-type tails, however, their inequalities seem to yield faster convergence rates than ours. When $\boldsymbol{\mu}_p$ is unknown, the concentration inequalities in [72] for the sample mean may be useful in deriving the matrix convergence rates for time series considered in [25]. We leave the details to interested readers. Note that it is not clear if the concentration inequalities in [25] and [72] are applicable to the matrix estimation under the more general temporal dependence that we consider here in this article. Also note that the concentration inequalities in [72] can handle the sample mean for nonlinear time series. It is of great interest to develop useful concentration inequalities for the sample covariance matrix for nonlinear, particularly long-memory, time series, which is beyond the scope of this article.

6. Numerical results.

6.1. *Gap-block cross-validation.*

For tuning parameter selection, we propose a gap-block cross-validation method that includes the following steps:

1. Split the data $\mathbf{X}_{p\times n}$ into $H_1 \geq 4$ approximately equal-sized nonoverlapping blocks $\mathbf{X}_i^*$, $i = 1, \dots, H_1$, such that $\mathbf{X}_{p\times n} = (\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_{H_1}^*)$. For each i, set aside block $\mathbf{X}_i^*$ that will be used as the validation data, and use the remaining data after further dropping the neighboring blocks at both sides of $\mathbf{X}_i^*$ as the training data that are denoted by $\mathbf{X}_i^{**}$.
2. Randomly sample H_2 blocks $\mathbf{X}_{H_1+1}^*, \dots, \mathbf{X}_{H_1+H_2}^*$ from $\mathbf{X}_{p\times n}$, where $\mathbf{X}_{H_1+j}^*$ consists of $\lceil n/H_1 \rceil$ consecutive columns of $\mathbf{X}_{p\times n}$ for each $j = 1, \dots, H_2$. Note that these sampled blocks can overlap. For each $i = H_1 + 1, \dots, H_1 + H_2$, set aside block $\mathbf{X}_i^*$ as the validation data, and use the remaining data by further excluding the $\lceil n/H_1 \rceil$ columns at both sides of $\mathbf{X}_i^*$ from $\mathbf{X}_{p\times n}$ as the training data that are denoted by $\mathbf{X}_i^{**}$.

3. Let $H = H_1 + H_2$. Select the optimal values of tuning parameters $\tau_1, \tau_2, \lambda_1$ and λ_2 among their corresponding prespecified candidate values $\{\tau_{1j}\}_{j=1}^J$, $\{\tau_{2j}\}_{j=1}^J$, $\{\lambda_{1j}\}_{j=1}^J$ and $\{\lambda_{2j}\}_{j=1}^J$, and denote them respectively by

$$\hat{\tau}_1 = \arg\min_{1\le j\le J} \frac{1}{H}\sum_{i=1}^{H} \| S_{\tau_{1j}}(\widehat{\boldsymbol{\Sigma}}_i^{**}) - \widehat{\boldsymbol{\Sigma}}_i^{*} \|_F^2,$$

$$\hat{\tau}_2 = \arg\min_{1\le j\le J} \frac{1}{H}\sum_{i=1}^{H} \| S_{\tau_{2j}}(\widehat{\mathbf{R}}_i^{**}) - \widehat{\mathbf{R}}_i^{*} \|_F^2,$$

$$\hat{\lambda}_1 = \arg\min_{1\le j\le J} \frac{1}{H}\sum_{i=1}^{H} \bigl[\operatorname{tr}\bigl(\widehat{\boldsymbol{\Omega}}_{\varepsilon,\lambda_{1j},i}^{**} \widehat{\boldsymbol{\Sigma}}_i^{*}\bigr) - \log\det\bigl(\widehat{\boldsymbol{\Omega}}_{\varepsilon,\lambda_{1j},i}^{**}\bigr)\bigr],$$

$$\hat{\lambda}_2 = \arg\min_{1\le j\le J} \frac{1}{H}\sum_{i=1}^{H} \bigl[\operatorname{tr}\bigl(\widehat{\boldsymbol{\Omega}}_{\lambda_{2j},i}^{**} \widehat{\boldsymbol{\Sigma}}_i^{*}\bigr) - \log\det\bigl(\widehat{\boldsymbol{\Omega}}_{\lambda_{2j},i}^{**}\bigr)\bigr],$$

where $\widehat{\boldsymbol{\Sigma}}_i^{*}$ and $\widehat{\mathbf{R}}_i^{*}$ are obtained from $\mathbf{X}_i^{*}$, $\widehat{\boldsymbol{\Sigma}}_i^{**}$ and $\widehat{\mathbf{R}}_i^{**}$ are obtained from $\mathbf{X}_i^{**}$, and $\widehat{\boldsymbol{\Omega}}_{\varepsilon,\lambda_{1j},i}^{**}$ and $\widehat{\boldsymbol{\Omega}}_{\lambda_{2j},i}^{**}$ are the CLIME and SPICE estimators, respectively, obtained from $\mathbf{X}_i^{**}$.

In the above cross-validation (CV), due to lack of independent observations, we use gap blocks, each of size $\approx \lceil n/H_1 \rceil$, to separate training and validation datasets so that they are nearly uncorrelated. The idea of using gap blocks has been employed by the hv-block CV of [58] for linear models with dependent data. Similar to the k-fold CV for i.i.d. data, Step 1 guarantees all observations are used for both training and validation, but is limited due to the constrain of keeping the temporal ordering of the observations. Step 2 allows more data splits. This is particularly useful when Step 1 only allows a small number of data splits due to large-size of the gap block and/or limited sample size n. Step 2 is inspired by the commonly used repeated random subsampling CV for i.i.d. observations [67]. The above loss functions for selecting tuning parameters are widely used in the literature [10, 18, 19, 61]. The theoretical justification for the gap-block CV remains open. In our numerical examples, we simply set $H_1 = H_2 = 10$, mimicking the 10-fold CV recommended by [32, 40]. Our simulation studies show that the method performs well for temporally dependent data.

6.2. *Simulation studies.* We evaluate the numerical performance of the hard and soft thresholding estimators for large correlation matrix and the CLIME and SPICE estimators for large precision matrix. We generate Gaussian data with zero mean and covariance matrix $\boldsymbol{\Sigma}$ or precision matrix $\boldsymbol{\Omega}$ from one of the following four models:

Model 1. $\sigma_{ij} = 0.6^{|i-j|}$;

Model 2. $\sigma_{ii} = 1$, $\sigma_{i,i+1} = \sigma_{i+1,i} = 0.6$, $\sigma_{i,i+2} = \sigma_{i+2,i} = 0.3$, and $\sigma_{ij} = 0$ for $|i - j| \geq 3$;

Model 3. $\omega_{ij} = 0.6^{|i-j|}$;

Model 4. $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$, and $\omega_{ij} = 0$ for $|i - j| \geq 3$.

Similar models have been considered in [10, 18, 19, 60, 61]. For the temporal dependence, we set $\mathrm{Corr}(X_{ki}, X_{\ell j}) = \Lambda_{k\ell}^{ij}\rho_{k\ell}$ with

$$\Lambda_{k\ell}^{ij} = \bigl(|i - j| + 1\bigr)^{-\alpha}, \qquad 1 \leq i, j \leq n, \tag{25}$$

so that $\rho_{[k]}^{ij} \sim |i - j|^{-\alpha}$. It is computationally expensive to simulate data $\boldsymbol{X}_{pn}$ directly from a multivariate Gaussian random number generator because of the large dimension of its covariance matrix $\mathrm{Cov}(\boldsymbol{X}_{pn})$. Instead, we simulate data using the method of [13], which approximately satisfy (25) (see details in the Supplementary Material [65]).

Simulations are conducted with sample size $n = 200$, variable dimension p ranging from 100 to 400, and 100 replications under each setting, for which α varies from 0.1 to 2. The i.i.d. case is also considered, for which an ordinary 10-fold CV is implemented. For each simulated data set, we choose the optimal tuning parameter from a set of 50 specified values (see Section S.4.1 in the Supplementary Material [65]). The CLIME and SPICE are computed by the R packages `flare` [49] and `QUIC` [44], respectively. For CLIME, we use the default perturbation of `flare` with $\varepsilon = n^{-1/2}$.

The estimation performance is measured by both the spectral norm and the Frobenius norm. True-positive rate (TPR) and false-positive rate (FPR) are used for evaluating sparsity recovering for the correlation matrix:

$$\mathrm{TPR} = \frac{\#\{(i, j) : s_\tau(\hat{\rho}_{ij}) \neq 0 \text{ and } \rho_{ij} \neq 0, i \neq j\}}{\#\{(i, j) : \rho_{ij} \neq 0, i \neq j\}},$$

$$\mathrm{FPR} = \frac{\#\{(i, j) : s_\tau(\hat{\rho}_{ij}) \neq 0 \text{ and } \rho_{ij} = 0, i \neq j\}}{\#\{(i, j) : \rho_{ij} = 0, i \neq j\}}.$$

Similar quantities are also reported for the precision matrix. The TPR and FPR are not provided for Models 1 and 3.

Simulation results are summarized in Tables 1–3. In all setups, the sample correlation matrix and the inverse of sample covariance matrix (whenever possible) perform the worst. It is not surprising that the performance of all the regularized estimators generally is better for weaker temporal dependence or smaller p. The soft thresholding method performs slightly better than the hard thresholding method in terms of matrix losses for small α and slightly worse for large α, and always has higher TPRs but bigger FPRs. The CLIME estimator performs similarly as the SPICE estimator in matrix norms, but generally yields lower FPRs.

TABLE 1

Comparison of average (SD) matrix losses for correlation matrix estimation

		$\widehat{\mathbf{R}}$	**Hard**	**Soft**	$\widehat{\mathbf{R}}$	**Hard**	**Soft**
p	α	**Spectral norm**			**Frobenius norm**		
Model 1							
100	0.1	13.7 (1.68)	2.8 (0.09)	2.6 (0.07)	22.6 (1.08)	9.9 (0.28)	8.7 (0.24)
	0.25	10.5 (1.59)	2.4 (0.15)	2.4 (0.08)	17.4 (0.95)	8.1 (0.42)	7.5 (0.26)
	0.5	7.8 (1.14)	2.0 (0.15)	2.2 (0.08)	14.3 (0.69)	6.8 (0.33)	6.6 (0.23)
	1	4.2 (0.45)	1.5 (0.10)	1.7 (0.08)	9.9 (0.29)	5.2 (0.23)	5.1 (0.20)
	2	2.6 (0.24)	1.1 (0.09)	1.4 (0.08)	7.5 (0.17)	3.9 (0.15)	4.0 (0.19)
	i.i.d.	2.4 (0.18)	1.0 (0.08)	1.3 (0.08)	7.0 (0.15)	3.5 (0.13)	3.7 (0.15)
200	0.1	27.2 (2.69)	2.9 (0.05)	2.8 (0.04)	45.6 (1.54)	14.5 (0.25)	13.1 (0.22)
	0.25	20.6 (2.54)	2.5 (0.14)	2.5 (0.06)	35.0 (1.39)	12.2 (0.56)	11.4 (0.29)
	0.5	15.2 (1.77)	2.2 (0.12)	2.3 (0.06)	28.7 (0.99)	10.2 (0.40)	10.1 (0.25)
	1	7.8 (0.64)	1.6 (0.08)	1.9 (0.06)	20.1 (0.35)	7.9 (0.24)	7.9 (0.21)
	2	4.3 (0.24)	1.3 (0.08)	1.6 (0.06)	15.1 (0.15)	5.9 (0.19)	6.3 (0.18)
	i.i.d.	3.8 (0.22)	1.1 (0.07)	1.5 (0.06)	14.1 (0.15)	5.3 (0.14)	5.8 (0.17)
300	0.1	40.6 (3.39)	3.0 (0.03)	2.8 (0.03)	68.5 (1.88)	18.0 (0.21)	16.5 (0.24)
	0.25	30.9 (3.23)	2.6 (0.11)	2.6 (0.04)	52.6 (1.75)	15.4 (0.63)	14.5 (0.30)
	0.5	22.5 (2.16)	2.3 (0.12)	2.4 (0.04)	43.2 (1.16)	12.8 (0.47)	12.9 (0.27)
	1	11.2 (0.79)	1.7 (0.05)	2.0 (0.05)	30.2 (0.42)	9.9 (0.21)	10.1 (0.25)
	2	5.8 (0.27)	1.3 (0.08)	1.7 (0.05)	22.8 (0.16)	7.5 (0.25)	8.2 (0.19)
	i.i.d.	5.0 (0.20)	1.2 (0.08)	1.6 (0.05)	21.2 (0.15)	6.7 (0.12)	7.5 (0.17)
400	0.1	54.2 (4.01)	3.0 (0.02)	2.9 (0.02)	91.7 (2.17)	20.9 (0.17)	19.4 (0.22)
	0.25	41.0 (3.88)	2.7 (0.09)	2.7 (0.04)	70.1 (2.09)	18.4 (0.61)	17.1 (0.29)
	0.5	29.8 (2.62)	2.3 (0.12)	2.5 (0.04)	57.7 (1.38)	15.2 (0.59)	15.3 (0.30)
	1	14.6 (0.91)	1.7 (0.05)	2.1 (0.04)	40.3 (0.48)	11.6 (0.22)	12.1 (0.20)
	2	7.2 (0.26)	1.4 (0.07)	1.8 (0.04)	30.4 (0.16)	9.0 (0.27)	9.8 (0.23)
	i.i.d.	6.0 (0.21)	1.2 (0.08)	1.6 (0.05)	28.2 (0.15)	7.9 (0.14)	8.9 (0.17)
Model 2							
100	0.1	13.8 (1.71)	1.8 (0.04)	1.6 (0.04)	22.6 (1.05)	8.7 (0.29)	7.7 (0.22)
	0.25	10.5 (1.61)	1.5 (0.18)	1.4 (0.09)	17.5 (0.94)	6.7 (0.48)	6.5 (0.24)
	0.5	7.8 (1.10)	1.2 (0.17)	1.3 (0.07)	14.3 (0.66)	5.2 (0.34)	5.6 (0.21)
	1	4.2 (0.40)	0.7 (0.09)	1.0 (0.05)	10.0 (0.27)	4.0 (0.17)	4.1 (0.16)
	2	2.5 (0.18)	0.6 (0.05)	0.8 (0.04)	7.5 (0.14)	2.6 (0.25)	3.2 (0.13)
	i.i.d.	2.3 (0.15)	0.5 (0.07)	0.7 (0.04)	7.0 (0.13)	2.0 (0.23)	2.8 (0.12)
200	0.1	27.2 (2.62)	1.8 (0.02)	1.7 (0.03)	45.6 (1.51)	12.9 (0.28)	11.6 (0.21)
	0.25	20.6 (2.29)	1.6 (0.15)	1.5 (0.07)	35.0 (1.29)	10.3 (0.56)	9.9 (0.27)
	0.5	15.1 (1.58)	1.3 (0.14)	1.4 (0.05)	28.8 (0.88)	7.9 (0.43)	8.6 (0.21)
	1	7.7 (0.57)	0.8 (0.10)	1.1 (0.04)	20.1 (0.34)	5.8 (0.15)	6.5 (0.20)
	2	4.2 (0.18)	0.6 (0.05)	0.9 (0.04)	15.2 (0.14)	4.2 (0.30)	5.0 (0.13)
	i.i.d.	3.6 (0.16)	0.6 (0.06)	0.8 (0.04)	14.1 (0.14)	3.2 (0.23)	4.4 (0.12)

TABLE 1
(*Continued*)

		$\widehat{\mathbf{R}}$	**Hard**	**Soft**	$\widehat{\mathbf{R}}$	**Hard**	**Soft**
p	α	**Spectral norm**			**Frobenius norm**		
300	0.1	40.8 (3.54)	1.8 (0.05)	1.7 (0.02)	68.7 (1.84)	16.0 (0.27)	14.6 (0.24)
	0.25	30.8 (2.95)	1.7 (0.17)	1.6 (0.13)	52.6 (1.62)	13.2 (0.69)	12.5 (0.28)
	0.5	22.4 (2.04)	1.4 (0.12)	1.4 (0.09)	43.3 (1.10)	10.1 (0.57)	10.9 (0.25)
	1	11.1 (0.73)	0.9 (0.08)	1.1 (0.03)	30.3 (0.41)	7.3 (0.16)	8.3 (0.20)
	2	5.6 (0.22)	0.6 (0.04)	0.9 (0.04)	22.8 (0.14)	5.5 (0.29)	6.5 (0.18)
	i.i.d.	4.7 (0.15)	0.6 (0.05)	0.8 (0.03)	21.2 (0.13)	4.1 (0.21)	5.7 (0.12)
400	0.1	54.0 (3.61)	1.8 (0.04)	1.7 (0.02)	91.7 (1.97)	18.6 (0.16)	17.2 (0.15)
	0.25	41.1 (3.58)	1.7 (0.09)	1.7 (0.12)	70.2 (1.89)	15.8 (0.63)	14.9 (0.33)
	0.5	29.7 (2.53)	1.5 (0.17)	1.5 (0.08)	57.7 (1.29)	12.1 (0.62)	13.0 (0.24)
	1	14.5 (0.86)	0.9 (0.08)	1.1 (0.03)	40.4 (0.46)	8.6 (0.16)	10.0 (0.23)
	2	7.0 (0.26)	0.7 (0.04)	0.9 (0.03)	30.4 (0.14)	6.6 (0.26)	7.7 (0.15)
	i.i.d.	5.7 (0.18)	0.6 (0.05)	0.9 (0.03)	28.3 (0.13)	4.9 (0.21)	6.8 (0.12)

We notice that the SPICE algorithm in the R package `QUIC` is much faster than the CLIME algorithm in the R package `flare` by using a single computer core. However, the column-by-column estimating nature of CLIME can speed up using parallel computing on multiple cores.

6.3. *rfMRI data analysis.* Here, we analyze a rfMRI data set for the estimation of brain functional connectivity. The preprocessed rfMRI data of a healthy young woman are provided by the WU-Minn Human Connectome Project (www.humanconnectome.org). The original data consist of 1200 temporal brain images and each image contains 229,404 brain voxels with size $2 \times 2 \times 2$ mm^3. We discard the first 10 images due to concerns of early nonsteady magnetization. For the ease of implementation, we use a grid-based method [66] to reduce the image dimension to 907 functional brain nodes that are placed in a regular three-dimensional grid spaced at 12-mm intervals throughout the brain. Each node consists of a 3-mm voxel-center-to-voxel-center radius pseudosphere, which encompasses 19 voxels, and the time series at the node is a spatially averaged time series of these 19 voxels. The temporal dependence of the 907 time series is approximated by the PDD model (4) with $C_0 = 1$ and $\alpha = 0.30$ (see Figure 1).

The functional connectivity between two brain nodes can be evaluated by either correlation or partial correlation; here, we follow the convention by simply calling them the marginal connectivity and the direct connectivity, respectively. For the marginal connectivity, we only apply the hard thresholding method for estimating the correlation matrix, which usually yields less number of false discoveries than the soft thresholding. We find that 1.47% of all the pairs of nodes are connected

TABLE 2

Comparison of average (SD) matrix losses for precision matrix estimation

		$\widehat{\Sigma}^{-1}$	**CLIME**	**SPICE**	$\widehat{\Sigma}^{-1}$	**CLIME**	**SPICE**
p	α	**Spectral norm**			**Frobenius norm**		
				Model 3			
100	0.1	381.7 (40.07)	4.9 (0.26)	5.7 (0.53)	850.5 (38.22)	28.8 (1.54)	27.1 (1.46)
	0.25	97.6 (9.23)	1.8 (0.09)	2.2 (0.08)	214.6 (9.38)	9.5 (0.34)	9.3 (0.20)
	0.5	43.3 (4.60)	2.4 (0.09)	2.7 (0.06)	93.9 (4.36)	7.7 (0.15)	8.6 (0.15)
	1	21.8 (2.74)	2.6 (0.06)	2.9 (0.04)	45.4 (2.73)	8.0 (0.19)	9.2 (0.15)
	2	14.1 (1.80)	2.7 (0.05)	2.9 (0.04)	28.9 (1.86)	8.0 (0.20)	9.1 (0.14)
	i.i.d.	12.6 (1.66)	2.5 (0.06)	2.8 (0.04)	25.5 (1.56)	7.4 (0.20)	8.6 (0.15)
200	0.1	N/A	6.2 (0.38)	5.8 (0.48)	N/A	49.6 (2.46)	38.4 (1.48)
	0.25	N/A	2.1 (0.12)	2.4 (0.06)	N/A	14.8 (0.52)	13.7 (0.18)
	0.5	N/A	2.6 (0.07)	2.8 (0.04)	N/A	11.9 (0.18)	12.8 (0.12)
	1	N/A	2.9 (0.05)	3.1 (0.03)	N/A	12.4 (0.23)	13.7 (0.14)
	2	N/A	2.9 (0.04)	3.1 (0.02)	N/A	12.6 (0.21)	13.8 (0.09)
	i.i.d.	N/A	2.7 (0.04)	3.0 (0.02)	N/A	11.6 (0.24)	13.3 (0.14)
300	0.1	N/A	5.3 (0.36)	5.9 (0.45)	N/A	51.2 (2.85)	47.1 (1.48)
	0.25	N/A	2.4 (0.11)	2.4 (0.05)	N/A	18.0 (0.36)	17.1 (0.18)
	0.5	N/A	2.8 (0.07)	2.9 (0.03)	N/A	15.7 (0.27)	15.9 (0.13)
	1	N/A	3.0 (0.04)	3.1 (0.02)	N/A	15.9 (0.28)	17.1 (0.12)
	2	N/A	3.0 (0.03)	3.1 (0.01)	N/A	16.1 (0.22)	17.3 (0.09)
	i.i.d.	N/A	2.8 (0.04)	3.1 (0.02)	N/A	15.0 (0.26)	16.8 (0.11)
400	0.1	N/A	5.8 (0.44)	6.0 (0.37)	N/A	63.9 (4.29)	54.7 (1.60)
	0.25	N/A	2.6 (0.08)	2.5 (0.05)	N/A	20.8 (0.22)	20.0 (0.19)
	0.5	N/A	2.9 (0.06)	2.9 (0.03)	N/A	19.0 (0.31)	18.6 (0.12)
	1	N/A	3.0 (0.04)	3.1 (0.02)	N/A	19.0 (0.32)	19.9 (0.13)
	2	N/A	3.1 (0.03)	3.2 (0.01)	N/A	19.0 (0.24)	20.2 (0.10)
	i.i.d.	N/A	2.9 (0.04)	3.1 (0.01)	N/A	17.9 (0.31)	19.7 (0.10)
				Model 4			
100	0.1	355.4 (37.62)	4.9 (0.40)	5.9 (0.72)	829.5 (35.78)	28.0 (2.05)	26.5 (1.68)
	0.25	91.1 (8.42)	1.9 (0.31)	1.7 (0.19)	209.0 (8.63)	8.2 (1.03)	7.3 (0.30)
	0.5	40.7 (4.29)	1.1 (0.10)	1.4 (0.07)	91.6 (3.96)	4.7 (0.17)	5.8 (0.19)
	1	20.5 (2.44)	1.3 (0.07)	1.6 (0.06)	44.4 (2.44)	5.1 (0.26)	6.2 (0.21)
	2	13.3 (1.62)	1.4 (0.07)	1.6 (0.05)	28.3 (1.70)	5.3 (0.25)	6.3 (0.17)
	i.i.d.	11.8 (1.44)	1.2 (0.06)	1.4 (0.05)	25.0 (1.37)	4.6 (0.24)	5.7 (0.18)
200	0.1	N/A	5.4 (0.50)	5.6 (0.61)	N/A	41.4 (2.89)	33.9 (1.61)
	0.25	N/A	1.8 (0.19)	1.6 (0.14)	N/A	11.5 (0.59)	10.5 (0.18)
	0.5	N/A	1.4 (0.11)	1.7 (0.04)	N/A	8.5 (0.32)	9.6 (0.17)
	1	N/A	1.6 (0.06)	1.8 (0.03)	N/A	9.1 (0.38)	10.5 (0.21)
	2	N/A	1.6 (0.05)	1.8 (0.03)	N/A	9.2 (0.32)	10.8 (0.17)
	i.i.d.	N/A	1.4 (0.06)	1.7 (0.03)	N/A	7.8 (0.34)	9.9 (0.17)

TABLE 2
(*Continued*)

		$\widehat{\Sigma}^{-1}$	CLIME	SPICE	$\widehat{\Sigma}^{-1}$	CLIME	SPICE
p	α		**Spectral norm**			**Frobenius norm**	
300	0.1	N/A	6.0 (0.54)	5.6 (0.67)	N/A	54.7 (4.26)	39.8 (1.58)
	0.25	N/A	1.6 (0.12)	1.6 (0.14)	N/A	14.0 (0.30)	13.2 (0.13)
	0.5	N/A	1.8 (0.07)	1.8 (0.04)	N/A	13.1 (0.51)	12.5 (0.20)
	1	N/A	1.9 (0.06)	1.9 (0.03)	N/A	13.1 (0.53)	13.8 (0.20)
	2	N/A	1.8 (0.05)	2.0 (0.03)	N/A	12.6 (0.39)	14.2 (0.20)
	i.i.d.	N/A	1.5 (0.05)	1.8 (0.02)	N/A	10.5 (0.38)	13.2 (0.19)
400	0.1	N/A	5.1 (0.46)	5.4 (0.62)	N/A	54.4 (4.12)	44.6 (1.43)
	0.25	N/A	1.8 (0.09)	1.7 (0.14)	N/A	17.5 (0.28)	15.5 (0.11)
	0.5	N/A	2.0 (0.06)	1.9 (0.03)	N/A	17.3 (0.55)	14.9 (0.19)
	1	N/A	2.0 (0.06)	2.0 (0.02)	N/A	16.7 (0.59)	16.5 (0.20)
	2	N/A	1.9 (0.05)	2.0 (0.02)	N/A	15.9 (0.50)	17.1 (0.20)
	i.i.d.	N/A	1.7 (0.06)	1.9 (0.02)	N/A	13.5 (0.48)	16.0 (0.20)

with a threshold value of 0.12 to the sample correlations. For the direct connectivity, we calculate the estimated partial correlations $\{-\hat{\omega}_{ij}/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}, i \neq j\}$ from the precision matrix estimator $\widehat{\mathbf{\Omega}} := (\hat{\omega}_{ij})_{p\times p}$. Both CLIME and SPICE yield similar results, hence we only report the result of CLIME. We find that 2.71% of all the pairs of nodes are connected conditional on all other nodes. Most of the nonzero estimated partial correlations have small absolute values, with the medium at 0.01 and the maximum at 0.45. About 0.62% of all the pairs of nodes are connected both marginally and directly.

Define the degree of a node to be the number of its connected nodes, and a hub to be a high-degree node. The marginal connectivity node degrees range from 0 to 164 with the medium at 2, and the direct connectivity node degrees range from 5 to 85 with the medium at 22. The top 10 hubs found by either method are provided in the Supplementary Material [65] with six overlapping hubs. Seven of the top 10 hubs of marginal connectivity are spatially close to those in [16] and [28] obtained from multiple subjects. Note that they arbitrarily used 0.25 as the threshold value for the sample correlations, whereas our threshold value of 0.12 is selected from cross-validation. Some additional results are provided in the Supplemental Material [65].

APPENDIX: TECHNICAL LEMMAS

The keys to the proofs of Theorems 1–9 are the proper concentration inequalities for $|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty$ and $|\widehat{\mathbf{R}} - \mathbf{R}|_\infty$ under temporal dependence. Once these inequalities are established, the rest of the proofs are straightforward extensions of

TABLE 3
Comparison of average (SD) TPR(%)/FPR(%) for Models 2 and 4

p	α	**Model 2, Hard**	**Model 2, Soft**
100	0.1	10.86 (4.35)/0.02 (0.03)	54.19 (4.41)/4.98 (1.26)
	0.25	35.16 (5.43)/0.07 (0.06)	70.72 (3.96)/6.10 (1.16)
	0.5	48.43 (3.76)/0.06 (0.06)	80.43 (3.19)/6.75 (1.19)
	1	60.92 (4.25)/0.02 (0.03)	94.34 (2.12)/7.23 (1.39)
	2	83.93 (4.08)/0.04 (0.05)	99.33 (0.73)/7.47 (1.57)
	i.i.d.	93.42 (2.63)/0.13 (0.09)	99.91 (0.21)/11.42 (1.82)
200	0.1	5.57 (2.93)/0.00 (0.00)	45.91 (3.86)/2.40 (0.55)
	0.25	28.31 (4.75)/0.02 (0.02)	64.71 (3.23)/3.20 (0.69)
	0.5	44.48 (3.02)/0.02 (0.02)	74.38 (2.42)/3.40 (0.59)
	1	57.45 (2.14)/0.01 (0.01)	91.40 (2.11)/3.84 (0.81)
	2	79.04 (3.66)/0.02 (0.01)	98.71 (0.67)/3.73 (0.58)
	i.i.d.	90.74 (2.68)/0.07 (0.05)	99.68 (0.31)/6.64 (0.65)
300	0.1	4.15 (2.50)/0.00 (0.00)	40.61 (3.94)/1.50 (0.43)
	0.25	24.28 (4.85)/0.01 (0.01)	61.27 (2.70)/2.13 (0.42)
	0.5	41.75 (3.51)/0.01 (0.01)	71.65 (2.51)/2.43 (0.47)
	1	55.42 (2.10)/0.00 (0.00)	89.41 (1.80)/2.61 (0.44)
	2	74.39 (3.23)/0.01 (0.01)	98.11 (0.69)/2.49 (0.57)
	i.i.d.	88.97 (2.29)/0.04 (0.02)	99.57 (0.34)/4.77 (0.84)
400	0.1	2.65 (1.29)/0.00 (0.00)	36.80 (2.27)/1.02 (0.23)
	0.25	20.81 (3.74)/0.01 (0.00)	58.30 (2.86)/1.54 (0.35)
	0.5	40.14 (3.58)/0.01 (0.01)	68.74 (2.06)/1.68 (0.35)
	1	53.82 (1.65)/0.00 (0.00)	87.51 (1.87)/1.80 (0.40)
	2	72.19 (2.58)/0.00 (0.00)	97.79 (0.66)/1.97 (0.22)
	i.i.d.	87.51 (1.65)/0.03 (0.01)	99.38 (0.30)/3.93 (0.40)

those in [10, 18, 59–61]. We provide these inequalities in the following lemmas, where Part (i) in Lemma A1 is an extension of the Hoeffding-type inequality [34], Theorem 7.27, and the Hanson–Wright inequality [63], Theorem 1.1, from finite-dimensional to infinite-dimensional sub-Gaussian random vectors. These lemmas can also be applied to the estimation of large band matrix [11] and other high-dimensional time series problems such as linear regression [72] and linear functionals [25].

LEMMA A1. *Let* $\boldsymbol{e} = (e_1, e_2, \ldots)^\top$ *be an infinite-dimensional random vector with each entry* e_i *satisfying* $E(e_i) = 0$ *and* $E(e_i^2) = 1$. *Let* $\boldsymbol{X} = \mathbf{A}\boldsymbol{e}$ *and* $\boldsymbol{Y} = \mathbf{B}\boldsymbol{e}$ *be two well-defined random vectors with length* n *in the sense of entrywise almost-sure convergence and mean-square convergence*, *where* $\mathbf{A}$ *and* $\mathbf{B}$ *are two deterministic matrices*. *For any* n-*dimensional deterministic vector* $\boldsymbol{b}$ *and all* $u > 0$:

TABLE 3
(*Continued*)

p	α	Model 4, CLIME	Model 4, SPICE
100	0.1	91.28 (2.76)/25.49 (2.37)	82.99 (2.76)/28.97 (1.04)
	0.25	92.65 (2.35)/17.82 (1.84)	90.93 (2.19)/29.68 (1.31)
	0.5	95.30 (1.73)/17.80 (1.47)	96.00 (1.54)/31.58 (1.49)
	1	98.47 (0.90)/14.37 (1.21)	99.24 (0.66)/30.65 (1.49)
	2	99.71 (0.36)/11.99 (1.27)	99.94 (0.17)/27.77 (1.34)
	i.i.d.	99.91 (0.20)/16.21 (1.63)	99.99 (0.07)/31.40 (1.28)
200	0.1	82.24 (2.70)/12.72 (0.64)	76.07 (1.95)/17.78 (0.56)
	0.25	84.83 (2.28)/15.70 (2.62)	84.75 (1.90)/18.87 (0.59)
	0.5	89.55 (2.39)/13.21 (3.00)	91.65 (1.45)/20.07 (0.64)
	1	93.81 (1.52)/7.27 (0.58)	97.12 (0.97)/19.07 (0.85)
	2	97.77 (0.97)/4.86 (0.55)	99.31 (0.42)/16.25 (0.81)
	i.i.d.	99.56 (0.36)/7.24 (0.79)	99.88 (0.18)/19.42 (0.81)
300	0.1	82.60 (3.59)/12.71 (2.55)	71.66 (1.71)/13.05 (0.34)
	0.25	77.62 (2.62)/14.39 (2.62)	81.09 (1.71)/14.06 (0.39)
	0.5	82.23 (2.48)/14.33 (3.57)	88.71 (1.44)/14.98 (0.42)
	1	86.84 (2.58)/4.71 (0.67)	94.87 (1.02)/14.20 (0.54)
	2	94.88 (1.38)/2.84 (0.41)	98.27 (0.68)/11.59 (0.65)
	i.i.d.	98.83 (0.49)/4.89 (0.58)	99.56 (0.29)/14.32 (0.70)
400	0.1	83.04 (2.84)/14.91 (2.84)	68.51 (1.49)/10.36 (0.24)
	0.25	76.76 (3.46)/15.11 (3.40)	78.50 (1.41)/11.41 (0.32)
	0.5	78.58 (2.35)/15.67 (3.64)	86.19 (1.44)/12.20 (0.35)
	1	79.44 (3.05)/4.40 (0.77)	92.85 (1.09)/11.55 (0.41)
	2	90.47 (2.32)/1.92 (0.35)	96.68 (0.85)/8.97 (0.55)
	i.i.d.	97.63 (0.82)/3.50 (0.52)	99.09 (0.39)/11.34 (0.60)

(i) *if all e_i satisfy condition* (C1) *with the same K, then*

$$P\big[|\boldsymbol{b}^\top \boldsymbol{X}| \ge u\big] \le 2\exp\left\{-\frac{Cu^2}{K^2\|\boldsymbol{b}\|_F^2\|\mathbf{A}\mathbf{A}^\top\|_2}\right\} \tag{A.1}$$

and

$$\begin{aligned} P\big[|\boldsymbol{X}^\top \boldsymbol{Y} - E(\boldsymbol{X}^\top \boldsymbol{Y})| \ge u\big] \le 2\exp\Bigg\{-C\min\Bigg(&\frac{u^2}{K^4\|\mathbf{A}\mathbf{A}^\top\|_F\|\mathbf{B}\mathbf{B}^\top\|_F}, \\ &\frac{u}{K^2\sqrt{\|\mathbf{A}\mathbf{A}^\top\|_2\|\mathbf{B}\mathbf{B}^\top\|_2}}\Bigg)\Bigg\} \end{aligned} \tag{A.2}$$

with an absolute constant $C > 0$;

(ii) *if all e_i satisfy condition* (C2) *with the same $K \ge 1$ and ϑ, then*

$$P\big[|\boldsymbol{b}^\top \boldsymbol{X}| \ge u\big] \le 2\exp\left\{-\frac{(u\|\boldsymbol{b}\|_F^{-1}\|\mathbf{A}\mathbf{A}^\top\|_2^{-1/2})^{\frac{1}{1/2+1/\vartheta}}}{CK(2/\vartheta)^{1/\vartheta}}\right\} \tag{A.3}$$

and

$$
\begin{aligned}
P\big[|\boldsymbol{X}^\top\boldsymbol{Y}-E(\boldsymbol{X}^\top\boldsymbol{Y})|\ge u\big]\le{}&2\exp\bigg\{-\frac{(u\|\mathbf{A}\mathbf{A}^\top\|_F^{-1/2}\|\mathbf{B}\mathbf{B}^\top\|_F^{-1/2})^{\frac{1}{1/2+2/\vartheta}}}{CK^2(4/\vartheta)^{4/\vartheta}}\bigg\}\\
&+2\exp\bigg\{-\frac{(u\|\mathbf{A}\mathbf{A}^\top\|_F^{-1/2}\|\mathbf{B}\mathbf{B}^\top\|_F^{-1/2})^{\frac{1}{1+2/\vartheta}}}{CK^2(2/\vartheta)^{2/\vartheta}}\bigg\}
\end{aligned}
\tag{A.4}
$$

with an absolute constant $C>0$;

(iii) *if all e_i satisfy condition (C3) with the same k and η_k, then*

$$
P\big[|\boldsymbol{b}^\top\boldsymbol{X}|\ge u\big]\le\big(C\eta_k k^{1/2}/u\big)^k\|\mathbf{A}\mathbf{A}^\top\|_2^{k/2}\|\boldsymbol{b}\|_F^k
\tag{A.5}
$$

and

$$
\begin{aligned}
P\big[|\boldsymbol{X}^\top\boldsymbol{Y}-E(\boldsymbol{X}^\top\boldsymbol{Y})|\ge u\big]\le{}&\big(C\eta_k^2k^{1/2}/u\big)^{k/2}\|\mathbf{A}\mathbf{A}^\top\|_F^{k/4}\|\mathbf{B}\mathbf{B}^\top\|_F^{k/4},\\
&+\big(C\eta_k^2k/u\big)^k\|\mathbf{A}\mathbf{A}^\top\|_F^{k/2}\|\mathbf{B}\mathbf{B}^\top\|_F^{k/2}
\end{aligned}
\tag{A.6}
$$

with an absolute constant $C>0$.

LEMMA A2. *Let $v_0>0$ be an absolute constant. Suppose that $\mathbf{X}_{p\times n}$ is generated from* (7). *Uniformly on $\boldsymbol{\Sigma}$ satisfying $|\boldsymbol{\Sigma}|_\infty\le v_0$ and $\{\mathbf{R}_{[k]}\}_{k=1}^p$ subject to* (2), *for any absolute constant $C>0$, if any of the following three conditions holds*:

(i) *all e_i satisfy condition (C1) with the same K, $u^*=C_1u_1$ with $C_1>0$ being a sufficiently large constant depending only on (v_0,K,C), and u_1 is given in* (12);

(ii) *all e_i satisfy condition (C2) with the same K and ϑ, $u^*=C_2u_2$ with $C_2>0$ being a sufficiently large constant depending only on (v_0,K,ϑ,C), and u_2 is given in* (23);

(iii) *all e_i satisfy condition (C3) with the same k and η_k, $u^*=C_3u_3$ with $C_3>0$ being a sufficiently large constant depending only on (v_0,k,η_k), and u_3 is given in* (24) *with the same constant C given above*;

then we have

$$
P\big[|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}|_\infty\ge u^*\big]=O(p^{-C}).
$$

LEMMA A3. *Suppose that $\mathbf{X}_{p\times n}$ is generated from* (7). *Uniformly on $\{\mathbf{R}_{[k]}\}_{k=1}^p$ subject to* (2), *for any absolute constant $C>0$, if any of the three conditions given in Lemma A2 holds and the corresponding $u_j=o(1)$, $j\in\{1,2,3\}$, then we have*

$$
P\big[|\widehat{\mathbf{R}}-\mathbf{R}|_\infty\ge u^*\big]=O(p^{-C}).
$$

Acknowledgments. The rfMRI data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657), which is funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University. We are grateful to the Editor Professor Edward George, the Associate Editor and three anonymous referees for their constructive comments that greatly reshaped this article.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation of large covariance and precision matrices from temporally dependent observations" (DOI: 10.1214/18-AOS1716SUPP; .pdf). The Supplementary Material contains technical preparations, detailed proofs of the technical lemmas given in the Appendix and all the theorems in the main text, useful numerical considerations and additional results of the rfMRI data analysis.

REFERENCES

[1] ATHREYA, K. B. and LAHIRI, S. N. (2006). *Measure Theory and Probability Theory*. Springer, New York. MR2247694

[2] BAI, J. and NG, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *J. Bus. Econom. Statist.* **23** 49–60. MR2108691

[3] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

[4] BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

[5] 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

[6] BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870

[7] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[8] BHATTACHARJEE, M. and BOSE, A. (2014). Consistency of large dimensional sample covariance matrix under weak dependence. *Stat. Methodol.* **20** 11–26. MR3205718

[9] BHATTACHARJEE, M. and BOSE, A. (2014). Estimation of autocovariance matrices for infinite dimensional vector linear process. *J. Time Series Anal.* **35** 262–281. MR3194965

[10] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[11] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[12] BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York. MR1324786

[13] BOCHUD, T. and CHALLET, D. (2007). Optimal approximations of power laws with exponentials: Application to volatility models with long memory. *Quant. Finance* **7** 585–589. MR2374605

[14] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. Update of, and a supplement to, the 1986 original. MR2178042

[15] BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. Springer, New York. MR1093459

[16] BUCKNER, R. L., SEPULCRE, J., TALUKDAR, T., KRIENEN, F. M., LIU, H., HEDDEN, T., ANDREWS-HANNA, J. R., SPERLING, R. A. and JOHNSON, K. A. (2009). Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci*. **29** 1860–1873.

[17] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

[18] 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

[19] 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

[20] CAI, T. T. and YUAN, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

[21] 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

[22] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

[23] CAO, G., BACHEGA, L. R. and BOUMAN, C. A. (2011). The sparse matrix transform for covariance estimation and analysis of high dimensional signals. *IEEE Trans. Image Process.* **20** 625–640. MR2799176

[24] CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455

[25] CHEN, X., XU, M. and WU, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Trans. Signal Process.* **64** 6459–6470. MR3566612

[26] CHLEBUS, E. (2009). An approximate formula for a partial sum of the divergent p-series. *Appl. Math. Lett.* **22** 732–737. MR2514902

[27] CIUCIU, P., ABRY, P. and HE, B. J. (2014). Interplay between functional connectivity and scale-free dynamics in intrinsic fMRI networks. *NeuroImage* **95** 248–263.

[28] COLE, M. W., PATHAK, S. and SCHNEIDER, W. (2010). Identifying the brain's most globally connected regions. *Neuroimage* **49** 3132–3148.

[29] CRAMÉR, H. (1946). *Mathematical Methods of Statistics. Princeton Mathematical Series* **9**. Princeton Univ. Press, Princeton, NJ. MR0016588

[30] DEMKO, S., MOSS, W. F. and SMITH, P. W. (1984). Decay rates for inverses of band matrices. *Math. Comp.* **43** 491–499. MR0758197

[31] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[32] FANG, Y., WANG, B. and FENG, Y. (2016). Tuning-parameter selection in regularized estimations of large covariance matrices. *J. Stat. Comput. Simul.* **86** 494–509. MR3421741

[33] FOMIN, V. (1999). *Optimal Filtering. Vol. II*: *Spatio-Temporal Fields. Mathematics and Its Applications* **481**. Kluwer Academic, Dordrecht. MR1707320

[34] FOUCART, S. and RAUHUT, H. (2013). *A Mathematical Introduction to Compressive Sensing*. Birkhäuser/Springer, New York. MR3100033

[35] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[36] GEWEKE, J. and PORTER-HUDAK, S. (1983). The estimation and application of long memory time series models. *J. Time Series Anal.* **4** 221–238. MR0738585

[37] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[38] GRANGER, C. W. J. and JOYEUX, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Series Anal.* **1** 15–29. MR0605572

[39] HARRISON, L., PENNY, W. D. and FRISTON, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage* **19** 1477–1491.

[40] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

[41] HE, B. J. (2011). Scale-free properties of the functional magnetic resonance imaging signal during rest and task. *J. Neurosci.* **31** 13786–13795.

[42] HINICH, M. J. (1982). Testing for Gaussianity and linearity of a stationary time series. *J. Time Series Anal.* **3** 169–176. MR0695228

[43] HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176. MR0614953

[44] HSIEH, C.-J., SUSTIK, M. A., DHILLON, I. S. and RAVIKUMAR, P. (2014). QUIC: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.* **15** 2911–2947. MR3277149

[45] HU, T.-C., ROSALSKY, A. and VOLODIN, A. (2008). On convergence properties of sums of dependent random variables under second moment and covariance restrictions. *Statist. Probab. Lett.* **78** 1999–2005. MR2458009

[46] 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

[47] JIANG, T. (2004). The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā* **66** 35–48. MR2082906

[48] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[49] LI, X., ZHAO, T., YUAN, X. and LIU, H. (2015). The `flare` package for high dimensional linear regression and precision matrix estimation in R. *J. Mach. Learn. Res.* **16** 553–557. MR3335796

[50] LIU, H., AUE, A. and PAUL, D. (2015). On the Marčenko–Pastur law for linear time series. *Ann. Statist.* **43** 675–712. MR3319140

[51] 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

[52] MANDELBROT, B. B. and VAN NESS, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10** 422–437. MR0242239

[53] MANOLAKIS, D. G., INGLE, V. K. and KOGON, S. M. (2005). *Statistical and Adaptive Signal Processing*: *Spectral Estimation*, *Signal Modeling*, *Adaptive Filtering*, *and Array Processing* **46**. Artech House, Norwood, MA.

[54] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[55] PALMA, W. (2007). *Long-Memory Time Series*: *Theory and Methods*. Wiley-Interscience, Hoboken, NJ. MR2297359

[56] POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M., SCHLAGGAR, B. L. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

[57] PRIESTLEY, M. B. and SUBBA RAO, T. (1969). A test for non-stationarity of time-series. *J. Roy. Statist. Soc. Ser. B* **31** 140–149. MR0269062

[58] RACINE, J. (2000). Consistent cross-validatory model-selection for dependent data: Hv-block cross-validation. *J. Econometrics* **99** 39–61.

[59] 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

[60] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[61] ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

[62] ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. MR2672482

[63] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258

[64] RYALI, S., CHEN, T., SUPEKAR, K. and MENON, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage* **59** 3852–3861.

[65] SHU, H. and NAN, B. (2019). Supplement to "Estimation of large covariance and precision matrices from temporally dependent observations." DOI:10.1214/18-AOS1716SUPP.

[66] SRIPADA, C., ANGSTADT, M., KESSLER, D., PHAN, K. L., LIBERZON, I., EVANS, G. W., WELSH, R. C., KIM, P. and SWAIN, J. E. (2014). Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks. *NeuroImage* **89** 110–121.

[67] SYED, M. N., PRINCIPE, J. C. and PARDALOS, P. M. (2012). Correntropy in data classification. In *Dynamics of Information Systems*: *Mathematical Foundations*. *Springer Proc. Math. Stat.* **20** 81–117. Springer, New York. MR3067311

[68] TAGLIAZUCCHI, E., VON WEGNER, F., MORZELEWSKI, A., BRODBECK, V., JAHNKE, K. and LAUFS, H. (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. *Proc. Natl. Acad. Sci. USA* **110** 15419–15424.

[69] TAQQU, M. S. (2003). Fractional Brownian motion and long-range dependence. In *Theory and Applications of Long-Range Dependence* 5–38. Birkhäuser, Boston, MA. MR1956042

[70] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[71] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[72] WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186

[73] WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215

[74] WU, W. B. and POURAHMADI, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19** 1755–1768. MR2589209

[75] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[76] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[77] ZHOU, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann. Statist.* **42** 532–562. MR3210978

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF MICHIGAN
1415 WASHINGTON HEIGHTS
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: haishu@umich.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, IRVINE
2066 BREN HALL
IRVINE, CALIFORNIA 92697
USA
E-MAIL: nanb@uci.edu