COVARIANCE AND PRECISION MATRIX ESTIMATION FOR HIGH-DIMENSIONAL TIME SERIES

BY XIAOHUI CHEN, MENGYU XU AND WEI BIAO WU

University of Illinois at Urbana-Champaign, University of Chicago and University of Chicago

We consider estimation of covariance matrices and their inverses (a.k.a. precision matrices) for high-dimensional stationary and locally stationary time series. In the latter case the covariance matrices evolve smoothly in time, thus forming a covariance matrix function. Using the functional dependence measure of Wu [*Proc. Natl. Acad. Sci. USA* **102** (2005) 14150–14154 (electronic)], we obtain the rate of convergence for the thresholded estimate and illustrate how the dependence affects the rate of convergence. Asymptotic properties are also obtained for the precision matrix estimate which is based on the graphical Lasso principle. Our theory substantially generalizes earlier ones by allowing dependence, by allowing nonstationarity and by relaxing the associated moment conditions.

1. Introduction. Estimation of covariance matrices and their inverses (a.k.a. precision matrices) is of fundamental importance in almost every aspect of statistics, ranging from the principal component analysis [Johnstone and Lu (2009)], graphical modeling [Meinshausen and Bühlmann (2006), Ravikumar et al. (2011), Yuan (2010)], classification based on the linear or quadratic discriminant analysis [Bickel and Levina (2004)], and real-world applications such as portfolio selection [Ledoit and Wolf (2003), Talih (2003)] and wireless communication [Abrahamsson, Selen and Stoica (2007), Guerci (1999), Li, Stocia and Wang (2003), Ward (1994)]. Suppose we have n temporally observed p-dimensional vectors $(\mathbf{z}_i)_{i=1}^n$, with $\mathbf{z}_i$ having mean zero and covariance matrix $\Sigma_i = \mathbb{E}(\mathbf{z}_i \mathbf{z}_i^\top)$ whose dimension is $p \times p$. Our goal is to estimate the covariance matrices Σ_i and their inverses $\Omega_i = \Sigma_i^{-1}$ based on the data matrix $Z_{p \times n} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$. In the classical situation where p is fixed, $n \to \infty$ and $\mathbf{z}_i$ are mean zero independent and identically distributed (i.i.d.) random vectors, it is well known that the sample covariance matrix

$$(1) \qquad \hat{\Sigma}_n = n^{-1} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top$$

is a consistent and well behaved estimator of Σ, and $\hat{\Omega}_n = \hat{\Sigma}_n^{-1}$ is a natural and good estimator of Ω. See Anderson (1958) for a detailed account. However, when the dimensionality p grows with n, random matrix theory asserts that $\hat{\Sigma}_n$ is no longer a consistent estimate of Σ in the sense that its eigenvalues do not converge to those of Σ; see, for example, the Marčenko–Pastur law [Marčenko and Pastur (1967)] or the Tracy–Widom law [Johnstone (2001)]. Moreover, it is clear that $\hat{\Omega}_n$ is not defined when $\hat{\Sigma}_n$ is not invertible in the high-dimensional case with $p > n$.

During the last decade, various special cases of the above covariance matrix estimation problem have been studied. In most of the previous papers it is assumed that the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d. and thus the covariance matrix $\Sigma_i \equiv \Sigma$ is time-invariant. See, for example, Bickel and Levina (2008a, 2008b), Cai, Zhang and Zhou (2010), Cai and Zhou (2012, 2013), where consistency and rates of convergence are established for various regularized (banded, tapered or thresholded) estimates of covariance matrices and their inverses. As an alternative regularized estimate for sparse precision matrix, one can adopt the Lasso-type entry-wise 1-norm penalized likelihood approach; see Banerjee, El Ghaoui and d'Aspremont (2008), Fan, Feng and Wu (2009), Friedman, Hastie and Tibshirani (2008), Ravikumar et al. (2011), Rothman et al. (2008). Other estimates include the Cholesky decomposition based method [Huang et al. (2006), Wu and Pourahmadi (2003)], neighborhood selection for sparse graphical models [Liu and Luo (2012), Meinshausen and Bühlmann (2006), Yuan (2010)], regularized likelihood approach [Fan, Feng and Wu (2009), Lam and Fan (2009)] and the sparse matrix transform [Cao, Bachega and Bouman (2011)]. Xiao and Wu (2012) considered covariance matrix estimation for univariate stationary processes.

The assumption that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d. is quite restrictive for situations that involve temporally observed data. In Zhou, Lafferty and Wasserman (2010) and Kolar and Xing (2011) the authors considered time-varying Gaussian graphical models where the sampling distribution can change smoothly over time. However, they assume that the underlying random vectors are independent. Using nonparametric smoothing techniques, they estimate the time-vary covariance matrices in terms of covariance matrix functions. Their asymptotic theory critically depends on the *independence* assumption.

The importance of estimating covariance matrices for dependent and nonstationary processes has been increasingly seen across a wide variety of research areas. In modeling spatial–temporal data, Wikle and Hooten (2010) proposed quadratic nonlinear dynamic models to accommodate the interactions between the processes which are useful for characterizing dynamic processes in geophysics [Kondrashov et al. (2005)]. Zheng, Chen and Blasch (2007) considered non-Gaussian clutter and noise processes in space–time adaptive processing, where the space–time covariance matrix is important for detecting airborne moving targets in the nonstationary clutter environment [Guerci (1999), Ward (1994)]. In finance, Jacquier, Polson and Rossi (2004) considered multivariate stochastic volatility

models parametrized by time-varying covariance matrices with heavy tails and correlated errors. Talih (2003) investigated the Markowitz portfolio selection problem for optimal returns of a large number of stocks with hidden and heterogeneous Gaussian graphical model structures. In essence, those real-world problems pose a number of challenges: (i) nonlinear dynamics of data generating systems, (ii) temporally dependent and nonstationary observations, (iii) high-dimensionality of the parameter space and (iv) non-Gaussian distributions. Therefore, the combination of more flexible nonlinear and nonstationary components in the models and regularized covariance matrix estimation are essential to perform related statistical inference.

In contrast to the longstanding progresses and extensive research that have been made in terms of heuristics and methodology, theoretical work on estimation of covariance matrices based on high-dimensional time series data is largely untouched. In this paper we shall substantially relax the i.i.d. assumption by establishing an asymptotic theory that can have a wide range of applicability. We shall deal with the estimation of covariance and precision matrices for high-dimensional stationary processes in Sections 2 and 3, respectively. Section 2 provides a rate of convergence for the thresholded estimator, and Section 3 concerns the graphical Lasso estimator for precision matrices. For locally stationary processes, an important class of nonstationary processes, we shall study in Section 4 the estimation of time-varying covariance and precision matrices. This generalization allows us to consider time-varying covariance and precision matrix estimation under temporal dependence; hence our results significantly extend previous ones by Zhou, Lafferty and Wasserman (2010) and Kolar and Xing (2011). Furthermore, by assuming a mild moment condition on the underlying processes, we can relax the multivariate Gaussian assumption that was imposed in Zhou, Lafferty and Wasserman (2010) and Kolar and Xing (2011) [and also by Bickel and Levina (2008a, 2008b) in the i.i.d. setting]. Specifically, we shall show that, thresholding on the kernel smoothed sample covariance matrices, estimators based on the localized graphical Lasso procedure are consistent estimators for time-varying covariance and precision matrices.

To deal with temporal dependence, we shall use the functional dependence measure of Wu (2005). With the latter, we are able to obtain explicit rates of convergence for the thresholded covariance matrix estimates and illustrate how the dependence affects the rates. In particular, we show that, based on the moment condition of the underlying process, there exists a threshold value. If the dependence of the process does not exceed that threshold, then the rates of convergence will be the same as those obtained under independence. On the other hand, if the dependence is stronger, then the rates of convergence will depend on the dependence. This phase transition phenomenon is of independent interest.

We now introduce some notation. We shall use $C, C_1, C_2, \ldots$ to denote positive constants whose values may differ from place to place. Those constants are independent of the sample size n and the dimension p. For some quantities a and b,

which may depend on n and p, we write $a \lesssim b$ if $a \leq Cb$ holds for some constant C that is independent of n and p and $a \asymp b$ if there exists a constant $0 < C < \infty$ such that $C \leq \liminf b/a \leq \limsup b/a \leq C^{-1}$. We use $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. For a vector $\mathbf{x} \in \mathbb{R}^p$, we write $|\mathbf{x}| = (\sum_{j=1}^p x_j^2)^{1/2}$ and for a matrix Σ, $|\Sigma|_1 = \sum_{j,k} |\sigma_{jk}|$, $|\Sigma|_\infty = \max_{j,k} |\sigma_{jk}|$, $|\Sigma|_F = (\sum_{j,k} \sigma_{jk}^2)^{1/2}$ and $\rho(\Sigma) = \max\{|\Sigma \mathbf{x}| : |\mathbf{x}| = 1\}$. For a random vector $\mathbf{z} \in \mathbb{R}^p$, write $\mathbf{z} \in \mathcal{L}^a$, $a > 0$, if $\|\mathbf{z}\|_a =: [\mathbb{E}(|\mathbf{z}|^a)]^{1/a} < \infty$.

2. Covariance matrix estimation for high-dimensional stationary processes. In this section we shall assume that $(\mathbf{z}_i)$ is a p-dimensional stationary process of the form

$$(2) \qquad \mathbf{z}_i = \mathbf{g}(\mathcal{F}_i),$$

where $\mathbf{g}(\mathcal{F}_i) = (g_1(\mathcal{F}_i), \ldots, g_p(\mathcal{F}_i))^\top$ is an $\mathbb{R}^p$-valued measurable function, $\mathcal{F}_i = (\ldots, \mathbf{e}_{i-1}, \mathbf{e}_i)$ is a shift process and $\mathbf{e}_i$ are i.i.d. random vectors. Following Wu (2005), we can view $\mathcal{F}_i$ and $\mathbf{z}_i$ as the input and the output of a physical system, respectively, and $\mathbf{g}(\cdot)$ is the transform representing the underlying physical mechanism. The framework (2) is quite general. Some examples are presented in Wu (2011). It can also be conveniently extended to locally stationary processes; see Section 4.

Write $\mathbf{z}_i = (Z_{1i}, \ldots, Z_{pi})^\top$ and $Z_{p \times n} = (\mathbf{z}_i)_{i=1}^n$, the data matrix observed at time points $i = 1, \ldots, n$. Here we shall consider estimation of the $p \times p$ covariance matrix $\Sigma = \text{cov}(\mathbf{z}_i)$ based on the realization $\mathbf{z}_1, \ldots, \mathbf{z}_n$, while Section 3 concerns estimation of its inverse. We consider Frobenius and spectral norm convergence of the *thresholded estimator*

$$(3) \qquad T_u(\hat{\Sigma}_n) = (\hat{\sigma}_{jk} \mathbb{I}(|\hat{\sigma}_{jk}| \geq u))_{1 \leq j,k \leq p},$$

where $\hat{\Sigma}_n = (\hat{\sigma}_{jk})$ is the sample covariance matrix defined in (1); see Bickel and Levina (2008a). It was shown in the latter paper that, with a properly chosen u, $T_u(\hat{\Sigma}_n)$ is a consistent estimator when $\Sigma_0 \in \mathcal{G}_r(\tilde{M})$ [see (45)] and $(\mathbf{z}_i)$ are i.i.d. sub-Gaussian. Our rates of convergence depend on the dependence of the process and the moment conditions, which can be quite mild. Our main theoretical result is given in Section 2.1. To obtain a consistent estimate for Σ, we need to impose regularization conditions. In particular, we shall assume that Σ is weakly dependent in that most of its entries are small, by providing a bound on the tail empirical process of covariances. Some examples are provided in Section 2.3 with applications to spatial–temporal processes.

2.1. *Asymptotic results.* To establish a convergence theory for covariance matrix estimates, we shall use the functional dependence measure of Wu (2005). Recall that $Z_{ji} = g_j(\mathcal{F}_i)$, $1 \leq j \leq p$, where $g_j(\cdot)$ is the jth coordinate projection

of the $\mathbb{R}^p$-valued measurable function $\mathbf{g}$. For $w > 0$, the functional dependence measure of Z_{ji} is defined by

$$(4) \qquad \theta_{i,w,j} = \|Z_{ji} - Z'_{ji}\|_w = (\mathbb{E}|Z_{ji} - Z'_{ji}|^w)^{1/w},$$

where $Z'_{ji} = g_j(\mathcal{F}'_i)$, $\mathcal{F}'_i = (\ldots, \mathbf{e}_{-1}, \mathbf{e}'_0, \mathbf{e}_1, \ldots, \mathbf{e}_i)$ and $\mathbf{e}'_0$ is such that $\mathbf{e}'_0, \mathbf{e}_l$, $l \in \mathbb{Z}$, are i.i.d. In other words, Z'_{ji} is a coupled version of Z_{ji} with $\mathbf{e}_0$ in the latter replaced by an i.i.d. copy $\mathbf{e}'_0$. In Wu (2011) functional dependence measures were computed for some commonly used linear and nonlinear stationary processes. We shall assume that the short-range dependence (SRD) condition holds,

$$(5) \qquad \Theta_{m,w} = \max_{1 \leq j \leq p} \sum_{l=m}^{\infty} \theta_{l,w,j} < \infty.$$

If (5) fails, the process $(Z_{ji})_{i \in \mathbb{Z}}$ may exhibit long-range dependence, and the asymptotic behavior can be quite different. A nonlinear process satisfying (5) is given in Example 2.1, while Example 2.2 concerns linear processes. Theorems 2.1 and 2.3 provide rates of convergence under the normalized Frobenius norm and the spectral norm for the thresholded estimate $T_u(\hat{\Sigma}_n)$, respectively. The constants C therein are independent of n, u and p.

THEOREM 2.1. *Assume that there exist $q > 2$, $\alpha > 0$, $\mu < \infty$ and a positive constant $C_0 < \infty$ such that $\max_{j \leq p} \|Z_{ji}\|_{2q} \leq \mu$ and $\Theta_{m,2q} \leq C_0 m^{-\alpha}$ for all $m \geq 1$. Let $\tilde{\alpha} = \alpha \wedge (1/2 - 1/q)$ and $\tilde{\beta} = (3 + 2\tilde{\alpha}q)/(1 + q)$. Define*

$$(6) \qquad H(u) = \begin{cases} u^{2-q} n^{1-q}, & \text{if } \alpha > 1/2 - 1/q; \\ u^{2-q} n^{1-q} (\log n)^{1+q}, & \text{if } \alpha = 1/2 - 1/q; \\ u^{2-q} n^{-q(\alpha+1/2)}, & \text{if } \alpha < 1/2 - 1/q, \end{cases}$$

$$(7) \qquad G(u) = \begin{cases} (n^{-1} + u^2)e^{-nu^2}, & \text{if } \alpha > 1/2 - 1/q; \\ (n^{-1}(\log n)^2 + u^2)e^{-n(\log n)^{-2}u^2}, & \text{if } \alpha = 1/2 - 1/q; \\ (n^{-\tilde{\beta}} + u^2)e^{-n^{\tilde{\beta}}u^2}, & \text{if } \alpha < 1/2 - 1/q \end{cases}$$

and

$$(8) \qquad D(u) = \frac{1}{p^2} \sum_{j,k=1}^{p} (u^2 \wedge \sigma_{jk}^2).$$

Then there exists a constant C, independent of u, n and p, such that

$$(9) \qquad \frac{\mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2}{p^2} \lesssim D(u) + \min\left(\frac{1}{n}, \frac{u^{2-q}}{n^{q/2}}, H(u) + G(Cu)\right).$$

REMARK 1. If $\alpha > 1/2 - 1/q$, elementary calculations indicate that $H(u) + G(Cu) \lesssim u^{2-q} n^{-q/2}$. Hence the right-hand side of (9) is $\asymp D(u) + \min(n^{-1}, H(u) + G(Cu))$. The term $u^{2-q} n^{-q/2}$ is needed if $\alpha \leq 1/2 - 1/q$.

By Theorem 2.1, if $u = O(n^{-1/2})$, then $p^{-2}\mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2 = O(n^{-1})$. Better convergence rates can be achieved if $D(n^{-1/2}) = o(n^{-1})$ by choosing a larger threshold; see cases (i)–(iii) in Corollary 2.2 below.

COROLLARY 2.2.   *Assume that the conditions of Theorem 2.1 hold. Let $\Upsilon = \inf_{u>0} p^{-2}\mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$; let $\tilde{G}(u) = \min(G(u), u^{2-q}n^{-q/2})$ if $\alpha \leq 1/2 - 1/q$ and $\tilde{G}(u) = G(u)$ if $\alpha > 1/2 - 1/q$.*

Let $u_\diamond \geq n^{-1/2}$ be the unique solution to the equation $H(u) = G(u)$. (i) If $\bar{D} =: p^{-2}\sum_{j,k=1}^p \sigma_{jk}^2 = O(H(1))$, then there is a fixed constant $c > 0$ such that $\Upsilon \lesssim H(u) \asymp H(1)$ for all $u \in [c, \mu]$. (ii) If $H(1) = o(\bar{D})$ and $D(u_\diamond) \leq H(u_\diamond)$, let $u_\dagger$ solve $D(u_\dagger) = H(u_\dagger)$, then $\Upsilon \lesssim D(u_\dagger)$. (iii) If $H(1) = o(\bar{D})$, $D(u_\diamond) > H(u_\diamond)$ and $D(n^{-1/2}) = o(n^{-1})$, let $u_\circ$ be the solution to the equation $D(u) = \tilde{G}(u)$ over the interval $u \in [n^{-1/2}, u_\diamond]$, then $\Upsilon \lesssim D(u_\circ)$. (iv) If $n^{-1} = O(D(n^{-1/2}))$, then the right-hand side of (9) is $\asymp n^{-1}$ for all $u \leq n^{-1/2}$ and $\Upsilon \lesssim n^{-1}$.

Theorem 2.1 and Corollary 2.2 describe how the Frobenius rate of convergence depends on the sample size n, the dimension p, the smallness measure quantified by the function $D(u)$ and the heaviness of tails (moment conditions) and strength of dependence which are characterized by q and α, respectively. It suggests the interesting dichotomy phenomenon: under the weaker dependence condition $\alpha > 1/2 - 1/q$, the thresholded estimate $T_u(\hat{\Sigma}_n)$ has the same convergence rates as those obtained under independence. However, the convergence becomes slower under stronger temporal dependence with $\alpha < 1/2 - 1/q$. The phase transition occurring at $\alpha = 1/2 - 1/q$. The theorem also provides information about the optimal threshold u, as revealed in its proof. The optimal threshold balances the bias or the smallness function $D(u)$, the tail function $H(u)$ and the variance component which roughly corresponds to the Gaussian-type function $G(u)$. Under different conditions, the optimal threshold assumes different forms; see Corollaries 2.4 and 2.5.

PROOF OF THEOREM 2.1.   We first assume $\alpha > 1/2 - 1/q$. Note that

$$(10) \quad \mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2 = \sum_{j,k=1}^p \mathbb{E}[\hat{\sigma}_{jk}\mathbb{I}(|\hat{\sigma}_{jk}| \geq u) - \sigma_{jk}]^2$$

$$\leq 2\sum_{j,k=1}^p \mathbb{E}(W_{jk}^2) + 2B(u/2),$$

where $W_{jk} = \hat{\sigma}_{jk}\mathbb{I}(|\hat{\sigma}_{jk}| \geq u) - \sigma_{jk}\mathbb{I}(|\sigma_{jk}| \geq u/2)$ and

$$(11) \quad B(u) = \sum_{j,k=1}^p \sigma_{jk}^2\mathbb{I}(|\sigma_{jk}| < u).$$

Let events $A_{jk}^1 = \{|\hat{\sigma}_{jk}| \geq u, |\sigma_{jk}| \geq u/2\}$, $A_{jk}^2 = \{|\hat{\sigma}_{jk}| < u, |\sigma_{jk}| \geq u/2\}$ and $A_{jk}^3 = \{|\hat{\sigma}_{jk}| \geq u, |\sigma_{jk}| < u/2\}$, $1 \leq j, k \leq p$. Observe that

$$W_{jk} = W_{jk}\mathbb{I}(A_{jk}^1) + W_{jk}\mathbb{I}(A_{jk}^2) + W_{jk}\mathbb{I}(A_{jk}^3).$$

We shall consider these three terms separately. Write $\xi_{jk} = \hat{\sigma}_{jk} - \sigma_{jk}$.

Case I: on the event A_{jk}^1, since the functional dependence measure for the product process $Z_{ji}Z_{ki}$, $i \in \mathbb{Z}$, satisfies

$$(12) \quad \begin{aligned} \|Z_{ji}Z_{ki} - Z'_{ji}Z'_{ki}\|_q &\leq \|Z_{ji}Z_{ki} - Z'_{ji}Z_{ki}\|_q + \|Z'_{ji}Z_{ki} - Z'_{ji}Z'_{ki}\|_q \\ &\leq \mu(\theta_{i,2q,j} + \theta_{i,2q,k}), \end{aligned}$$

it follows from the moment inequality Theorem 2.1 in Wu (2007) that

$$(13) \quad \|\xi_{jk}\|_q \leq c_q n^{-1/2}\mu\Theta_{0,2q},$$

where c_q is a constant only depending on q. Let $C_1 = c_q^2\mu^2\Theta_{0,2q}^2$. Then

$$(14) \quad \mathbb{E}\{W_{jk}^2\mathbb{I}(A_{jk}^1)\} \leq \mathbb{E}\xi_{jk}^2\mathbb{I}(|\sigma_{jk}| \geq u/2) \leq C_1\frac{\mathbb{I}(|\sigma_{jk}| \geq u/2)}{n}.$$

Case II: on the event A_{jk}^2, we observe that

$$(15) \quad \begin{aligned} \mathbb{E}\{W_{jk}^2\mathbb{I}(A_{jk}^2)\} &= \mathbb{E}[\sigma_{jk}^2\mathbb{I}(|\sigma_{jk}| \geq u/2, |\hat{\sigma}_{jk}| < u)] \\ &\leq 2\mathbb{E}[\xi_{jk}^2\mathbb{I}(|\sigma_{jk}| \geq u/2, |\hat{\sigma}_{jk}| < u)] \\ &\quad + 2\mathbb{E}[\hat{\sigma}_{jk}^2\mathbb{I}(|\sigma_{jk}| \geq u/2, |\hat{\sigma}_{jk}| < u)] \\ &\leq 2(C_1 n^{-1} + u^2)\mathbb{I}(|\sigma_{jk}| \geq u/2). \end{aligned}$$

Case III: on the event A_{jk}^3, let

$$(16) \quad \begin{aligned} \Delta_{jk} &= \mathbb{E}[\xi_{jk}^2\mathbb{I}(|\hat{\sigma}_{jk}| \geq u, |\sigma_{jk}| < u/2)] \\ &= \mathbb{E}[\xi_{jk}^2\mathbb{I}(|\hat{\sigma}_{jk}| \geq u, |\sigma_{jk}| < u/2, |\xi_{jk}| > u/2)] \\ &\leq \mathbb{E}[\xi_{jk}^2\mathbb{I}(|\xi_{jk}| > u/2)]. \end{aligned}$$

Then

$$(17) \quad \begin{aligned} \mathbb{E}\{W_{jk}^2\mathbb{I}(A_{jk}^3)\} &= \mathbb{E}[\hat{\sigma}_{jk}^2\mathbb{I}(|\hat{\sigma}_{jk}| \geq u, |\sigma_{jk}| < u/2)] \\ &\leq 2\Delta_{jk} + 2\sigma_{jk}^2\mathbb{I}(|\sigma_{jk}| < u/2). \end{aligned}$$

Since the functional dependence measure for the product process $(Z_{ji}Z_{ki})_i$ satisfies (12), under the decay condition $\Theta_{m,2q} \leq Cm^{-\alpha}$, $\alpha > 1/2 - 1/q$, we have by Theorem 2(ii) in Liu, Xiao and Wu (2013) that

$$(18) \quad \mathbb{P}(|\xi_{jk}| > v) \leq \frac{C_2 n}{(nv)^q} + C_3 e^{-C_4 nv^2}$$

holds for all $v > 0$. Using integration by parts, we obtain

$$
\mathbb{E}[\xi_{jk}^2 \mathbb{I}(|\xi_{jk}| > v)] = v^2 \mathbb{P}(|\xi_{jk}| > v) + \int_{v^2}^{\infty} \mathbb{P}(|\xi_{jk}| > \sqrt{w}) \, dw
$$

$$
\leq v^2 \left[\frac{C_2 n}{(nv)^q} + C_3 e^{-C_4 n v^2} \right]
$$

(19)

$$
+ \int_{v^2}^{\infty} \left[\frac{C_2 n}{(n\sqrt{w})^q} + C_3 e^{-C_4 n w} \right] dw
$$

$$
= C_5 n^{1-q} v^{2-q} + C_3 ((C_4 n)^{-1} + v^2) e^{-C_4 n v^2},
$$

where $C_5 = C_2 q / (q - 2)$. By (13), we also have

$$
(20) \qquad \mathbb{E}[\xi_{jk}^2 \mathbb{I}(|\xi_{jk}| > v)] \leq \min\left(\|\xi_{jk}\|_2^2, \frac{\|\xi_{jk}\|_q^q}{v^{q-2}} \right) \lesssim \min\left(\frac{1}{n}, \frac{v^{2-q}}{n^{q/2}} \right).
$$

Combining cases I, II and III, by (11) and (14)–(20), we have

$$
\frac{\mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2}{p^2} \lesssim \frac{B(u/2)}{p^2} + \frac{1 + n u^2}{n p^2} \sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \geq u/2)
$$

(21)

$$
+ \min\left(\frac{1}{n}, \frac{u^{2-q}}{n^{q/2}}, H(u) + G(Cu) \right) =: M_0(u),
$$

where $C = C_4^{1/2}/2$, and the constant of $\lesssim$ is independent of p, u and n. If $u \geq n^{-1/2}$, then (9) clearly follows from the inequality $p^{-2} \sum_{j,k} \mathbb{I}(|\sigma_{jk}| \geq v) \leq v^{-2} D(v)$. If $u < n^{-1/2}$, we also have (9) since in this case $M_0(u) \asymp n^{-1}$ and the right-hand side of (9) has the same order of magnitude n^{-1}.

The other cases with $0 < \alpha < 1/2 - 1/q$ and $\alpha = 1/2 - 1/q$ can be similarly handled. The key difference is that, instead of (18), we shall now use the following versions of Nagaev inequalities which can allow stronger dependence:

$$
\mathbb{P}(|\xi_{jk}| > v) \leq \begin{cases} \dfrac{C_2 n^{q(1/2-\alpha)}}{(nv)^q} + C_3 e^{-C_4 n^{\tilde{\beta}} v^2}, & \text{if } \alpha < 1/2 - 1/q; \\[3mm] \dfrac{C_2 n (\log n)^{1+q}}{(nv)^q} + C_3 e^{-C_4 n (\log n)^{-2} v^2}, & \text{if } \alpha = 1/2 - 1/q. \end{cases}
$$

See also Liu, Xiao and Wu (2013). □

PROOF OF COROLLARY 2.2.   Let $M_1(u)$ be the term on the right-hand side of (9). We now minimize $M_1(u)$ over $u > 0$. Let

$$
M_2(v) = D(v) + \min\left(\frac{1}{n}, \frac{v^{2-q}}{n^{q/2}}, \max(H(v), G(v)) \right).
$$

Then $\inf_{u>0} M_1(u) \asymp \inf_{v>0} M_2(v)$. Clearly, $\inf_{v \leq n^{-1/2}} M_2(v) \asymp n^{-1}$. Let $v \geq n^{-1/2}$. If $\alpha > 1/2 - 1/q$, then for some constant c_q, we have $v^{2-q} n^{-q/2} \geq c_q v^2 e^{-nv^2} \geq c_q G(v)/2$. Also we have $v^{2-q} n^{-q/2} \geq H(v)$. Hence

$$(22) \qquad \inf_{v \geq n^{-1/2}} M_2(v) \asymp \inf_{v \geq n^{-1/2}} \max[D(v), H(v), G(v)].$$

Note that the equation $H(u) = G(u)$ has a unique solution $u_\diamond$ on $(n^{-1/2}, \infty)$, and the function $\max[H(u), G(u)]$ is decreasing over $u \geq n^{-1/2}$. A plot of the function in (22) is given in Figure 2(a). Let $u_\natural$ be the minimizer of the right-hand side of (22). For (i), assume $\bar{D} \leq C_0 n^{1-q}$ for some $C_0 > 0$. Then $u_\natural$ satisfies $D(u) = H(u)$, which implies $u \geq C_0^{1/(2-q)}$, and hence (i) follows. Note that (ii) follows in view of $u_\dagger = u_\natural \geq u_\diamond$ and $u_\dagger \to 0$. Similarly we have (iii) since $u_\natural = u_\circ$. The last case (iv) is straightforward since $M_2(u) \asymp n^{-1}$ for all $u \leq n^{-1/2}$.

If $0 < \alpha \leq 1/2 - 1/q$, assume $v \geq n^{-1/2}$, and then (22) still holds with $G(v)$ therein replaced by $\tilde{G}(v)$. A plot for this case is given in Figure 2(b). Note that $\tilde{G}(v) = G(v)$ if $v \geq u_\diamond$. Then we can similarly have (i)–(iv). $\square$

REMARK 2. From the proof of Corollary 2.2, if $0 < \alpha \leq 1/2 - 1/q$, in case (iii), we can actually have the following dichotomy: let $u_\triangle$ be the solution to the equation $G(u) = u^{2-q} n^{-q/2}$. Then the minimizer $u_\natural \in [n^{-1/2}, u_\triangle]$ if $D(u_\triangle) \geq \tilde{G}(u_\triangle)$ and $u_\natural \in [u_\triangle, u_\diamond]$ if $D(u_\triangle) \leq \tilde{G}(u_\triangle)$. For $\alpha > 1/2 - 1/q$, (22) indicates that $v^{2-q} n^{-q/2}$ is not needed; see also Remark 1.

Using the argument for Theorem 2.1, we can similarly establish a spectral norm convergence rate. Bickel and Levina (2008a) considered the special setting with i.i.d. vectors. Our Theorem 2.3 is a significant improvement by relaxing the independence assumption, by obtaining a sharper rate and by presenting a moment bound. As in Theorem 2.1, we also have the phase transition at $\alpha = 1/2 - 1/q$. Note that Bickel and Levina (2008a) only provides a probabilistic bound.

THEOREM 2.3. *Let the moment and the dependence conditions in Theorem 2.1 be satisfied. Let $L_\alpha = n^{1/q-1}, n^{1/q-1}(\log n)^{1+1/q}, n^{-\alpha-1/2}$ and $J_\alpha = n^{-1/2}, n^{-1/2} \log n, n^{-\tilde{\beta}/2}$, for $\alpha > 1/2 - 1/q, \alpha = 1/2 - 1/q$, and $\alpha < 1/2 - 1/q$, respectively. Define*

$$(23) \quad D_*(u) = \max_{1 \leq k \leq p} \sum_{j=1}^{p} (|\sigma_{jk}| \wedge u), \qquad N_*(u) = \max_{1 \leq k \leq p} \sum_{j=1}^{p} \mathbb{I}(|\sigma_{jk}| \geq u),$$

and $M_*(u) = L_\alpha p^{1/q} N_*^{1+1/q}(u) + J_\alpha (\log p)^{1/2} N_*(u)$. Then there exists a constant C, independent of u, n and p, such that

$$(24) \quad \begin{aligned} \|\rho(T_u(\hat{\Sigma}_n) - \Sigma)\|_2 &\lesssim D_*(u) + M_*(u/2) \\ &\quad + p \min\left(\frac{1}{\sqrt{n}}, \frac{u^{1-q/2}}{n^{q/4}}, (H(u) + G(Cu))^{1/2}\right), \end{aligned}$$

where $H(\cdot)$ and $G(\cdot)$ are given in (6) *and* (7), *respectively.*

PROOF.   We shall only deal with the weaker dependent case with $\alpha > 1/2 - 1/q$. The other cases similarly follow. Recall the proof of Theorem 2.1 for W_{jk}, ξ_{jk} and A_{jk}^l, $l = 1, 2, 3$. Let matrices $V_l = (W_{jk}\mathbb{I}(A_{jk}^l))_{j,k \leq p}$. Similar to (11), let $B_*(u) = \max_{1 \leq k \leq p} \sum_{j=1}^{p} |\sigma_{jk}|\mathbb{I}(|\sigma_{jk}| < u)$. Then

$$(25) \qquad \big|\rho\big(T_u(\hat{\Sigma}_n) - \Sigma\big)\big| \leq B_*(u/2) + \sum_{l=1}^{3}\big|\rho(V_l)\big|.$$

Let $N_k(u) = \{j : |\sigma_{jk}| \geq u/2\}$ and $z_u = C_1 M_*(u/2)$, where $C_1 > 0$ is a large constant. Since $\rho(V_1) \leq \max_{k \leq p} \sum_{j \in N_k(u)} |\hat{\sigma}_{jk} - \sigma_{jk}| =: Q$, by (18),

$$
\begin{aligned}
\frac{\|\rho(V_1)\|_2^2}{2} &\leq \int_0^{\infty} z\mathbb{P}(Q \geq z)\,dz \\
(26) \qquad\qquad &\lesssim \frac{z_u^2}{2} + \int_{z_u}^{\infty} zpS_u\left[\frac{n}{(nz/S_u)^q} + e^{-C_4 nz^2 S_u^{-2}}\right]dz \\
&\lesssim M_*^2(u/2),
\end{aligned}
$$

where $S_u = N_*(u/2)$. Similar to (15), since $\sigma_{jk} \leq |\hat{\sigma}_{jk} - \sigma_{jk}| + u$ on A_{jk}^2,

$$(27) \qquad \big|\rho(V_2)\big| \leq Q + uS_u \leq Q + 2D_*(u).$$

Using the idea of (17), we have

$$
\begin{aligned}
\rho^2(V_3) &\leq \sum_{j,k}\big|W_{jk}\mathbb{I}(A_{jk}^3)\big|^2 \\
(28) \qquad\qquad &\leq 2\sum_{j,k}\xi_{jk}^2\mathbb{I}(|\xi_{jk}| > u/2) + 2B_*^2(u/2).
\end{aligned}
$$

By (16)–(20) and (25)–(28), we have (24) since $B_*(u/2) \leq B_*(u) \leq D_*(u)$.   □

The bounds in Theorems 2.1 and 2.3 depend on the smallness measures, the moment order q, the dependence parameter α, the dimension p and the sample size n. The problem of selecting optimal thresholds is highly nontrivial. Our numeric experiments show that the cross-validation based method has a reasonably good performance. However, we are unable to provide a theoretical justification of the latter method, and pose it as an open problem.

EXAMPLE 2.1 (Stationary Markov chains).   We consider the nonlinear process $(\mathbf{z}_i)$ defined by the iterated random function

$$(29) \qquad \mathbf{z}_i = \mathbf{g}(\mathbf{z}_{i-1}, \mathbf{e}_i),$$

where $\mathbf{e}_i$'s are i.i.d. innovations, and $\mathbf{g}(\cdot, \cdot)$ is an $\mathbb{R}^p$-valued and jointly measurable function, which satisfies the following two conditions: (i) there exists some $\mathbf{x}_0$ such that $\|\mathbf{g}(\mathbf{x}_0, \mathbf{e}_0)\|_{2q} < \infty$ and (ii)

$$(30) \qquad L = \sup_{\mathbf{x} \neq \mathbf{x}'} \frac{\|\mathbf{g}(\mathbf{x}, \mathbf{e}_0) - \mathbf{g}(\mathbf{x}', \mathbf{e}_0)\|_{2q}}{|\mathbf{x} - \mathbf{x}'|} < 1.$$

Then, it can be shown that $\mathbf{z}_i$ defined in (29) has a stationary ergodic distribution $\mathbf{z}_0 \in \mathcal{L}^{2q}$ and, in addition, $(\mathbf{z}_i)$ has the *geometric moment contraction* (GMC) property; see Wu and Shao (2004) for details. Therefore, we have $\Theta_{m,2q} = O(L^m)$ and Theorems 2.1 and 2.3 with $\alpha > 1/2 - 1/q$ and $\tilde{\beta} = 1$ can be applied.

EXAMPLE 2.2 (Stationary linear processes). An important special class of (2) is the vector linear process

$$(31) \qquad \mathbf{z}_i = \sum_{m=0}^{\infty} A_m \mathbf{e}_{i-m},$$

where $A_m, m \geq 0$, are $p \times p$ matrices, and $\mathbf{e}_i$ are i.i.d. mean zero random vectors with finite covariance matrix $\Sigma_{\mathbf{e}} = \mathbb{E}(\mathbf{e}_i \mathbf{e}_i^\top)$. Then $\mathbf{z}_i$ exists almost surely with covariance matrix $\Sigma = \sum_{m=0}^{\infty} A_m \Sigma_{\mathbf{e}} A_m^\top$ if the latter converges. Assume that the innovation vector $\mathbf{e}_i = (e_{1i}, \ldots, e_{pi})^\top$, where e_{ji} are i.i.d. with mean zero, variance 1 and $e_{ji} \in \mathcal{L}^{2q}$, $q > 2$, and the coefficient matrices $A_i = (a_{i,jk})_{1 \leq j,k \leq p}$ satisfy $\max_{j \leq p} \sum_{k=1}^{p} a_{i,jk}^2 = O(i^{-2-2\gamma})$, $\gamma > 0$. By Rosenthal's inequality, the functional dependence measure $\theta_{i,2q,j}^2 \leq c_q \sum_{k=1}^{p} a_{i,jk}^2 = O(i^{-2-2\gamma})$, and hence by (5) $\Theta_{m,2q} = O(m^{-\gamma})$. By Theorem 2.1, the normalized Frobenius norm of the thresholded estimator has a convergence rate established in (9) with $\alpha = \gamma$, $\tilde{\alpha} = \gamma \wedge (1/2 - 1/q)$ and $\tilde{\beta}$. Note that our moment condition relaxes the commonly assumed sub-Gaussian condition in previous literature [Lam and Fan (2009), Rothman et al. (2008), Zhou, Lafferty and Wasserman (2010)]. For the vector AR(1) process $\mathbf{z}_i = A\mathbf{z}_{i-1} + \mathbf{e}_i$, where A is a real matrix with spectral norm $\rho(A) < 1$, it is of form (31) with $A_m = A^m$, and the functional dependence measure $\theta_{i,2q,j} = O(\rho(A)^i)$. The rates of convergence established in (9) hold with $\alpha > 1/2 - 1/q$ and $\tilde{\beta} = 1$.

2.2. *Positive-definitization.* The thresholded estimate $T_u(\hat{\Sigma}_n)$ may not be positive definite. Here we shall propose a simple modification that is positive definite and has the same rate of convergence. Let $T_u(\hat{\Sigma}_n) = \mathbf{Q}\hat{\Lambda}\mathbf{Q}^\top = \sum_{j=1}^{p} \hat{\lambda}_j \mathbf{q}_j \mathbf{q}_j^\top$ be its eigen-decomposition, where $\mathbf{Q}$ is an orthonormal matrix and $\hat{\Lambda}$ is a diagonal matrix. For $v > 0$, consider

$$(32) \qquad \tilde{S}_v = \sum_{j=1}^{p} (\hat{\lambda}_j \vee v) \mathbf{q}_j \mathbf{q}_j^\top,$$

where $0 < v \le \sqrt{p}\varpi$ and ϖ^2 is the rate of convergence in (9). Let $\mu_1, \dots, \mu_p$ be the diagonal elements of $\mathbf{Q}^\top \Sigma \mathbf{Q}$. Then we have by Theorem 2.1 that $\sum_{j=1}^p (\hat{\lambda}_j - \mu_j)^2 \le p^2 \varpi^2$, and consequently

$$
\begin{aligned}
|\tilde{S}_v - \Sigma|_F^2 &\le 2|\tilde{S}_v - T_u(\hat{\Sigma}_n)|_F^2 + 2|T_u(\hat{\Sigma}_n) - \Sigma|_F^2 \\
&\le 2 \sum_{j=1}^p (\hat{\lambda}_j - (\hat{\lambda}_j \vee v))^2 + 2\varpi^2 p^2 \\
&\le 2 \sum_{j=1}^p (2\hat{\lambda}_j^2 \mathbf{1}_{\hat{\lambda}_j \le 0} + 2v^2) + 2\varpi^2 p^2.
\end{aligned}
$$

If $\hat{\lambda}_j \le 0$, since $\mu_i \ge 0$, we have $|\hat{\lambda}_j| \le |\hat{\lambda}_j - \mu_i|$. Then $|\tilde{S}_v - \Sigma|_F^2 \le 4v^2 p + 6\varpi^2 p^2 \le 10\varpi^2 p^2$. Note that the eigenvalues of $\tilde{S}_v$ are bounded below by v, and thus it is positive definite. In practice we suggest using $v = (p^{-1} \sum_{j,k=1}^p u^2 \times \mathbb{I}(|\hat{\sigma}_{jk}| \ge u))^{1/2}$. The same positive-definization procedure also applies to the spectral norm and its rate can be similarly preserved.

2.3. *Classes of covariance matrices.* In this section we shall compute the smallness measure $D(u)$ for certain class of covariance matrices, so that Theorem 2.1 is applicable. We consider some widely used spatial processes. Let the vectors $\mathbf{z}_i = (Z_{1i}, \dots, Z_{pi})^\top$, $1 \le i \le n$, be observed at sites $s_1^\circ, \dots, s_p^\circ \in \mathbb{R}^2$. Assume that the covariance function between Z_{ji} and Z_{ki} satisfies

(33) $$\sigma_{jk} = \operatorname{cov}(Z_{ji}, Z_{ki}) = f(d(s_j^\circ, s_k^\circ)),$$

where $d(s_j^\circ, s_k^\circ)$ is a distance between sites s_j° and s_k°, and f is a real-valued function with $f(0) = 1$ and $f(\infty) = 0$. For example, we can choose $d(s, s') = |s - s'|$ as the Euclidean distance between sites s and s'. Assume that, as $m \to \infty$,

(34) $$f(m) = O(m^{-K}),$$

where the index $K > 0$ characterizes the spatial dependence, or

(35) $$f(m) \le \exp(-C(m/\tau)^\theta), \qquad 0 < \theta \le 2,$$

where τ is the characteristic length-scale, and

(36) $$\frac{1}{p} \sum_{j,k=1}^p \mathbb{I}(d(s_j^\circ, s_k^\circ) \le m) = O(m^\chi).$$

Condition (36) outlines the geometry of the sites $(s_j^\circ)_{j=1}^p$, and χ can be roughly interpreted as the correlation dimension. It holds with $\chi = 2$ if s_j° are $\mathbb{Z}^2$ points in a disk or a square, and $\chi = 1$ if $s_j^\circ = (j, 0)$, $j = 1, \dots, p$. The rational quadratic

covariance function [Rasmussen and Williams (2006)] is an example of (34), and it is widely used in spatial statistics,

$$(37) \qquad f(m) = \left(1 + \frac{m^2}{K\tau^2}\right)^{-K/2},$$

where K is the smoothness parameter and $\tau > 0$ is the length scale parameter. We now provide a bound for $D(u)$. By (34) and (36), as $u \downarrow 0$, the covariance tail empirical process function

$$(38) \qquad F(u) =: \frac{1}{p^2} \sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \geq u) \leq p^{-1} \min(p, Cu^{-\chi/K})$$

for some constant $C > 0$ independent of n, u and p. If $K > \chi/2$, then

$$D(u) = u^2 F(u) + \frac{1}{p^2} \sum_{l=0}^{\infty} \sum_{j,k=1}^{p} \sigma_{jk}^2 \mathbb{I}(u2^{-l-1} \leq |\sigma_{jk}| < u2^{-l})$$

$$(39) \qquad \leq u^2 F(u) + \sum_{l=0}^{\infty} (2^{-l}u)^2 F(2^{-l-1}u)$$

$$\leq u^2 p^{-1} \min(p, Cu^{-\chi/K}) = u^2 \min(1, Cp^{-1}u^{-\chi/K}).$$

In the strong spatial dependence case with $K < \chi/2$, we have

$$(40) \qquad D(u) \leq \min(Cp^{-K}, u^2).$$

To this end, it suffices to prove this relation with $u^2 > p^{-K}$. Let $u_0 = p^{-K/\chi}$. Then

$$\bar{D} \leq \sum_{l=0}^{\infty} (2^{1+l}u_0)^2 F(2^l u_0)$$

$$\leq \sum_{l=0}^{\infty} (2^{1+l}u_0)^2 Cp^{-1}(2^{1+l}u_0)^{-\chi/K} \leq Cp^{-2K/\chi}.$$

Class (35) allows the γ-exponential covariance function with $f(m) = \exp(-(m/\tau)^{\gamma})$, and some Matérn covariance functions [Stein (1999)] that are widely used in spatial statistics. With (36), following the argument in (39), we can similarly have

$$(41) \qquad D(u) \leq \min(u^2, Cp^{-1}\tau^{\chi} u^2 (\log(2 + u^{-1}))^{\chi/\theta}).$$

Corollary 2.4 of Theorem 2.1 concerns covariance matrices satisfying (38). Slightly more generally, we introduce a decay condition on the tail empirical process of covariances. Note that (38) is a special case of (42) with $M = Cp$ and $r = \chi/K$. For (37) with possibly large length scale parameter τ, we can let $M = C\tau^2 p$. Similarly, Corollary 2.5 can be applied to f satisfying (35) and the class $\mathcal{L}_r(M)$ defined in (43), with $M = p\tau^{\chi}$ and $r = \chi/\theta$.

DEFINITION 2.1.   For $M > 0$, let $\mathcal{H}_r(M)$, $0 \leq r < 2$, be the collection of $p \times p$ covariance matrices $\Sigma = (\sigma_{jk})$ such that $\sup_{j \leq p} \sigma_{jj} \leq 1$ and, for all $0 < u \leq 1$,

$$\sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \geq u) \leq M u^{-r}, \tag{42}$$

and $\mathcal{L}_r(M)$, $r > 0$, be the collection of $\Sigma = (\sigma_{jk})$ with $\sup_{j \leq p} \sigma_{jj} \leq 1$ and

$$\sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \geq u) \leq M \log^r(2 + u^{-1}). \tag{43}$$

COROLLARY 2.4.   *Assume* (42). *Let conditions in Theorem* 2.1 *be satisfied and* $\alpha > 1/2 - 1/q$. *Let* $\Upsilon = p^{-2} \sup_{\Sigma \in \mathcal{H}_r(M)} \inf_{u > 0} \mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$. (i) *If* $n^{q-1} = O(p^2/M)$, *then for* $u \asymp 1$, $\Upsilon = O(H(u)) = O(n^{1-q})$. (ii) *If* $p^2/M = o(n^{q-1})$ *and* $n^{(r+q)/2-1}(\log n)^{(q-r)/2} \leq p^2/M$, *let* $u'_{\dagger} = (n^{1-q} p^2/M)^{1/(q-r)}$, *then* $\Upsilon = O(u'^{2-q}_{\dagger} n^{1-q})$. (iii) *If* $p^2/M = o(n^{q-1})$ *and*

$$\frac{n^{1-q/2}}{(\log n)^{(q-r)/2}} \leq \frac{M}{p^2} n^{r/2} \leq 1, \tag{44}$$

then the equation $u^{2-r} M/p^2 = u^2 e^{-nu^2}$ *has solution* $u'_{\circ} \asymp [n^{-1} \log(2 + p^2 M^{-1} n^{-r/2})]^{1/2}$ *and* $\Upsilon = O(u'^{2-r}_{\circ} M/p^2)$. (iv) *If* $n^{r/2} \geq p^2/M$, *then the right-hand side of* (9) *is* $\asymp n^{-1}$ *for* $u = O(n^{-1/2})$ *and* $\Upsilon = O(n^{-1})$.

In particular, if $p^2/M \asymp n^{\phi}$, $\phi > 0$, *then we have* (i), (ii), (iii) *or* (iv) *if* $\phi > q - 1$, $q - 1 > \phi > (q + r - 2)/2$, $(q + r - 2)/2 > \phi > r/2$ *or* $r/2 > \phi$ *holds, respectively.*

PROOF.   Similar to (39), we have $D(u) \leq \min(u^2, Cu^{2-r} M/p^2)$. Note that the solution $u_{\diamond} \geq n^{-1/2}$ to the equation $H(u) = G(u)$ satisfies $u_{\diamond} \sim ((q/2 - 1)n^{-1} \log n)^{1/2}$. Then by Corollary 2.2, (i)–(iv) follow from elementary but tedious manipulations. Details are omitted.   □

By taking into consideration of M in the tail empirical process condition (42), we can view p^2/M as the *effective dimension*. Corollary 2.4 describes the choice of the optimal threshold u at different regions of the effective dimension p^2/M and the sample size n. Case (i) [resp., (iv)] corresponds to the overly large (resp., small) dimension case. The most interesting cases are (ii) and (iii). For the former, the tail function $H(\cdot)$ determines the rate of convergence with a larger threshold $u_{\dagger}$, while for the latter with moderately large dimension the Gaussian-type function $G(\cdot)$ leads to the optimal threshold $u_{\circ} < u_{\dagger}$.

COROLLARY 2.5.   *Assume* (43). *Let conditions in Theorem* 2.1 *be satisfied with* $\alpha > 1/2 - 1/q$ *and* $\Upsilon = p^{-2} \sup_{\Sigma \in \mathcal{L}_r(M)} \inf_{u > 0} \mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$.

(i) *If $n^{q-1} = O(p^2/M)$, then for $u \asymp 1$, $\Upsilon = O(H(u)) = O(n^{1-q})$. (ii) If $p^2/M = o(n^{q-1})$ and $n^{q/2-1}(\log n)^{r+q/2} \leq p^2/M$, let $\varepsilon_\dagger = n^{1-q}p^2/M$ and $u'_\dagger = \varepsilon_\dagger^{1/q}(\log(2 + \varepsilon_\dagger^{-1}))^{-r/q}$. Then $\Upsilon = O(u'^{2-q}_\dagger n^{1-q})$. (iii) If $n^{q/2-1}(\log n)^{r+q/2} > p^2/M \geq (\log n)^r$, let $\eta = (\log n)^{-r}p^2/M$. If $\eta \geq 2^{-r}$ let $u'_\circ = (n^{-1}\log \eta)^{1/2}$. Then $\Upsilon = O(n^{-1}\eta^{-1}\log \eta)$. (iv) If η in (iii) is less than 2^{-r}, then the right-hand side of (9) is $\asymp n^{-1}$ for $u = O(n^{-1/2})$ and $\Upsilon = O(n^{-1})$.*

PROOF.   We have $D(u) = u^2 \min(1, p^{-2}M \log^r(2 + u^{-1}))$. We shall again apply Corollary 2.2. Case (i) is straightforward. For (ii), we note that the equation $u^q \log^r(2 + u^{-1}) = \varepsilon$ has solution $u_\dagger \asymp \varepsilon_\dagger^{1/q}(\log(2 + \varepsilon_\dagger^{-1}))^{-r/q}$. Under (iii), the equation $u^2 p^{-2}M \log^r(2 + u^{-1}) = G(u)$ has solution $\asymp u'_\circ$.   □

Corollaries 2.4 and 2.5 deal with the weaker dependence case with $\alpha > 1/2 - 1/q$. By Corollary 2.2, similar versions can be obtained for $\alpha \leq 1/2 - 1/q$. Details are omitted.

As a numeric example, we use the rational quadratic covariances (37) to illustrate the rates of convergence given in Theorem 2.1 and Corollary 2.2. We choose $n = 100$, $p = 200$, $K = 4$, the moment $q = 4$ and consider the weaker ($\alpha > 1/4$) and stronger ($\alpha = 1/8$) temporal dependence cases. We first generate p random sites uniformly distributed on the $p^{1/2} \times p^{1/2}$ square; see Figure 1(a). Figure 1(b), 1(c) and 1(d) show three 200×200 rational quadratic covariance matrices (37) respectively with length scale parameters $\tau = p^{1/3}$, $p^{1/6}$ and $p^{1/9}$, which correspond to different levels of spatial dependence. Next, we calculate the terms in Corollary 2.2 for the thresholded estimator. The results are shown in Figure 2. In the plots, $u_\diamond$ is the solution of $G(u) = H(u)$. Note that, $u_\natural$, the minimizer of $\max[D(u), H(u), G(u)]$ over $u \geq n^{-1/2}$, can be either $u_\dagger$ or $u_\circ$. We observe that when the spatial dependence decreases, that is, the covariance matrix Σ has more small entries [e.g., Figure 1(d)], a larger threshold is needed to yield the optimal rate of convergence. When the temporal dependence increases (i.e., $\alpha = 1/8$), a larger threshold is needed and the rate of convergence is slower than the one in the weaker dependence case (i.e., $\alpha > 1/4$).

2.4. *Comparison with earlier results.* We now compare (42) with the commonly used sparsity condition defined in terms of the *strong ℓ^q-ball* [Bickel and Levina (2008a), Cai, Liu and Luo (2011), Cai and Zhou (2012)]

$$(45) \quad \mathcal{G}_r(\tilde{M}) = \left\{ \Sigma \,\middle|\, \max_{j \leq p} \sigma_{jj} \leq 1; \; \max_{1 \leq k \leq p} \sum_{j=1}^{p} |\sigma_{jk}|^r \leq \tilde{M} \right\}, \qquad 0 \leq r < 1.$$

When $r = 0$, (45) becomes $\max_{1 \leq k \leq p} \sum_{j=1}^{p} \mathbb{I}(\sigma_{jk} \neq 0) \leq \tilde{M}$, a sparsity condition in the rigid sense. We observe that condition (42) defines a broader class of sparse
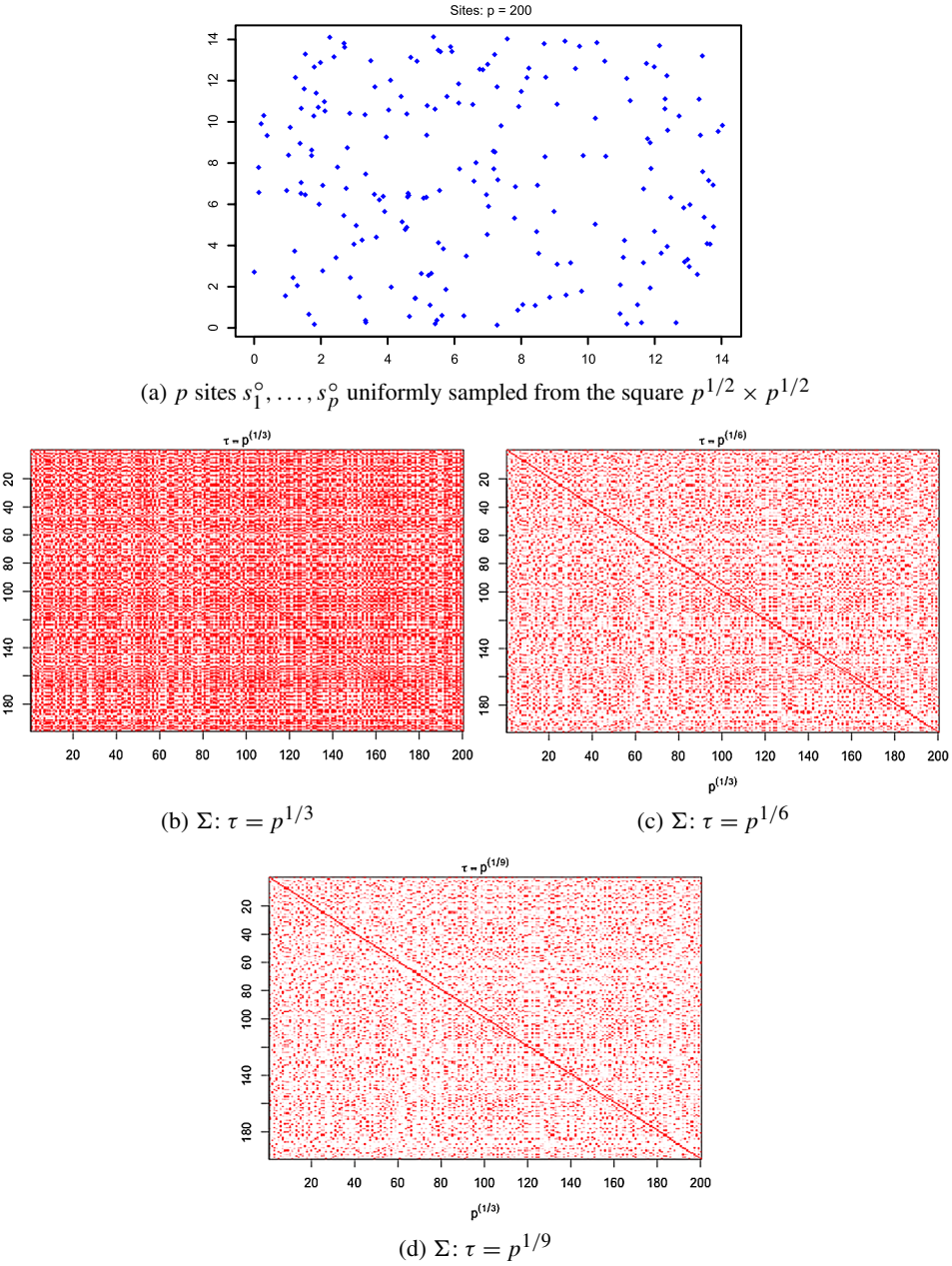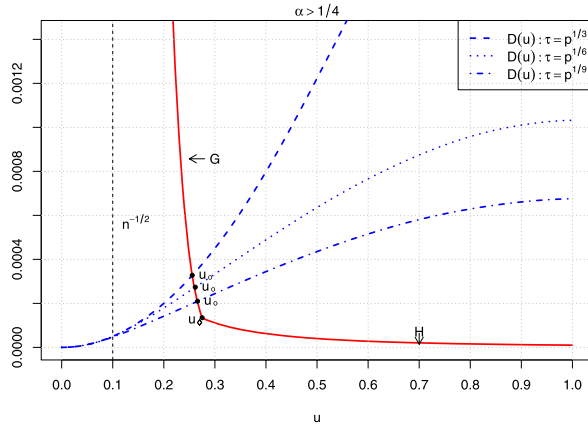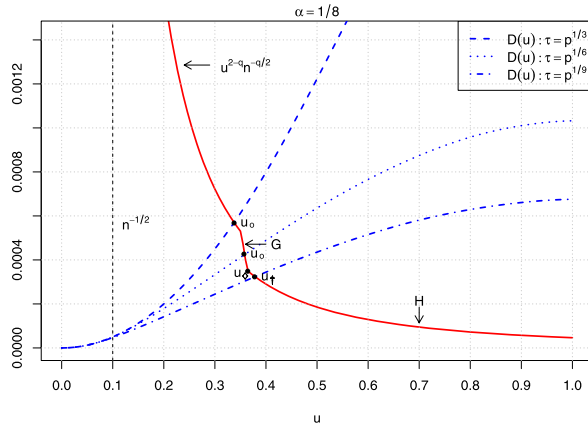
(a) p sites $s_1^\circ, \ldots, s_p^\circ$ uniformly sampled from the square $p^{1/2} \times p^{1/2}$

(b) Σ: $\tau = p^{1/3}$

(c) Σ: $\tau = p^{1/6}$

(d) Σ: $\tau = p^{1/9}$

FIG. 1. *Rational quadratic covariance matrix Σ for the uniform random sites model on the* $[0, p^{1/2}]^2$ *square with three different scale length parameters*: $\tau = p^{1/3}$, $p^{1/6}$ *and* $p^{1/9}$.

(a) Weaker temporal dependence with $\alpha > 1/4$

(b) Stronger temporal dependence with $\alpha = 1/8$

FIG. 2. *Rates of convergence for the thresholded estimator in the weaker ($\alpha > 1/4$) and stronger ($\alpha = 1/8$) temporal dependence cases.*

covariance matrices in the sense that $\mathcal{G}_r(M/p) \subset \mathcal{H}_r(M)$, which follows from

$$\sum_{j,k} \mathbb{I}(|\sigma_{jk}| \geq u) \leq p \max_k \sum_j \frac{|\sigma_{jk}|^r}{u^r} \leq M u^{-r}.$$

Hence Corollary 2.4 generalizes the consistency result of $T_u(\hat{\Sigma}_n)$ in Bickel and Levina (2008a) to the non-Gaussian time series. Note that our convergence is in $\mathcal{L}^2$ norm, while the error bounds in previous work [see, e.g., Bickel and Levina (2008a, 2008b)] are of probabilistic nature; namely in the form $|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$ is bounded with large probability under the strong ℓ^q-ball conditions.

The reverse inclusion $\mathcal{H}_r(M) \subset \mathcal{G}_r(M/p)$ may be false since the class $\mathcal{G}_r$ specifies the uniform size of sums in matrix columns, whereas (42) can be viewed as an

overall smallness measure over all entries of the matrix. As an example, consider the covariance matrix

$$
(46) \qquad \Sigma_{p \times p} = \begin{pmatrix} 1 & \varepsilon & \varepsilon & \cdots & \varepsilon \\ \varepsilon & 1 & 0 & \cdots & 0 \\ \varepsilon & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon & 0 & 0 & \cdots & 1 \end{pmatrix}
$$

where $0 < \varepsilon \le (p-1)^{-1/2}$ so that Σ is positive-definite. Then for any threshold level $u \in (\varepsilon, 1)$, $\sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \ge u) = p$ and for any $u \in (0, \varepsilon]$, $\sum_{j,k=1}^{p} \mathbb{I}(|\sigma_{jk}| \ge u) = 3p - 2$. In both cases, we may choose $M = O(p)$. On the other hand, $\max_k \sum_j |\sigma_{jk}|^r = 1 + (p-1)\varepsilon^r$. So $\Sigma \notin \mathcal{G}_r(M/p)$ for any $\varepsilon \ge (p-1)^{\eta/r - 1/r}$ with $\eta \in (0, 1 - r/2)$.

With the strong ℓ^q-ball and sub-Gaussian conditions, Cai and Zhou (2012) showed that the minimax rate under the Bregman divergence is $O(n^{-1} + \tilde{M}(\log p/n)^{1-r/2})$. Observing that the upper bounds in Corollary 2.4 is established under the larger parameter space $\mathcal{H}_r(M) \supset \mathcal{G}_r(\tilde{M})$ where $M = p\tilde{M}$ and milder polynomial moments conditions, the lower bound of Cai and Zhou (2012) automatically becomes a lower bound in our setup. Therefore, in the moderately high-dimensional situation with weaker temporal dependence, we can conclude that the Frobenius norm bound in Corollary 2.4(iii) is minimax rate optimal.

COROLLARY 2.6. *Let $\alpha > 1/2 - 1/q$. Under the conditions in Corollary* 2.4(iii) *and in addition assume $p^2 M^{-1} n^{-r/2} \ge p^\varepsilon$ for some $\varepsilon > 0$. Then*

$$
(47) \qquad \inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{H}_r(M)} p^{-1} \mathbb{E} |\hat{\Sigma} - \Sigma|_F^2 \asymp \frac{M}{p} \left(\frac{\log p}{n} \right)^{1-r/2},
$$

where the inf is taken over all possible estimators based on the data $Z_{p \times n}$.

We next compare our Theorem 2.3 with the result in Section 2.3 of Bickel and Levina (2008a), where the special class (45) is considered. Assuming $\max_j \|Z_{ji}\|_{2q} \le \mu$, they obtained the *probabilistic bound*

$$
(48) \qquad \rho(T_{u_{\mathrm{BL}}}(\hat{\Sigma}_n) - \Sigma) = O_p(\tilde{M} u_{\mathrm{BL}}^{1-r}) \qquad \text{where } u_{\mathrm{BL}} = C p^{2/q} n^{-1/2},
$$

and $C > 0$ is a sufficiently large constant. As a natural requirement for consistency, we assume $u_{\mathrm{BL}} \to 0$, namely $p = o(n^{q/4})$. Since $\Sigma \in \mathcal{G}_r(\tilde{M})$, we have $D_*(u) \le \tilde{M} u^{1-r} =: \bar{D}_*(u)$ and $N_*(u) \le \min(p, \tilde{M} u^{-r}) =: \bar{N}_*(u)$. Consider the weaker dependence case with $\alpha > 1/2 - 1/q$. Note that in (24) $D_*(\cdot)$ is nondecreasing, while all other three functions are nonincreasing. Let u_1, u_2, u_3 be the solutions to the equations $\bar{N}_*^{1+1/q}(u) p^{1/q} n^{1/q-1} = \bar{D}_*(u)$, $\bar{N}_*(u)(n^{-1} \log p)^{1/2} = \bar{D}_*(u)$, and $H_*(u) = p u^{1-q/2} n^{(1-q)/2} = \bar{D}_*(u)$, respectively; let $u_4 = \max(u_1, u_2, u_3, (n^{-1} \log p)^{1/2})$. For a sufficiently large constant

$C_2 > 0$, $G_*(C_2 u_4) = o(D_*(u_4))$ and hence the right-hand side of (24) is of order $D_*(u_4) = O(\tilde{M} u_4^{1-r})$ if $u = C_2 u_4$. Let $u_1' = (\tilde{M} p n^{1-q})^{1/(q+r)}$ and $u_1'' = (p^{1+2/q} \tilde{M}^{-1} n^{1/q-1})^{1/(1-r)}$. Note that $u_1 = u_1'$ if $p \geq M(u_1')^{-r}$ and $u_1 = u_1''$ if $p \leq M(u_1')^{-r}$. In both cases we have by elementary calculations that $u_1 = o(u_{\text{BL}})$. Similarly, we have $u_2 = o(u_{\text{BL}})$ and $u_3 = o(u_{\text{BL}})$. Hence $u_4 = o(u_{\text{BL}})$ and our rate of convergence $D_*(u_4)$ is sharper.

Based on Theorem 2.3 and the above discussion, we have:

COROLLARY 2.7. *Let the conditions in Theorem* 2.1 *be satisfied and* $\alpha > 1/2 - 1/q$. *Let* $\Lambda = \sup_{\Sigma \in \mathcal{G}_r(\tilde{M})} \inf_{u>0} \|\rho(T_u(\hat{\Sigma}_n) - \Sigma)\|_2$. *Assume* $\tilde{M} \asymp p^\theta$, $0 \leq \theta \leq 1$ *and* $p \asymp n^\tau$, $\tau > 0$. *Let* $\phi_1' = (\tau\theta + \tau + 1 - q)/(q+r)$, $\phi_1'' = (\tau(1 - \theta + 2/q) - 1 + 1/q)/(1-r)$, $\phi_1 = \min(\phi_1', \phi_1'')$, $\phi_3 = (2\tau - 2\tau\theta + 1 - q)/(q - 2r)$ *and* $\phi = \max(\phi_1, \phi_3)$. (i) *If* $\phi > -1/2$, *then* $\Lambda = O(n^{\phi(1-r)+\theta\tau})$. (ii) *If* $\phi \leq -1/2$, *then* $\Lambda = O(n^{\theta\tau}(n^{-1} \log p)^{(1-r)/2})$.

3. Precision matrix estimation for high-dimensional stationary processes.
As a straightforward estimate for precision matrices, one can invert the regularized covariance matrix estimates. However, this inversion procedure may cause the precision matrix estimate to lose sparsity. Sparsity of the precision matrix $\Omega = \Sigma^{-1}$ has important statistical meaning because a zero entry in $\Omega = (\omega_{jk})_{1 \leq j,k \leq p}$ reflects the conditional independence when $\mathbf{z}_i$ are multivariate Gaussian. In the graphical model representation, $\omega_{ij} = 0$ indicates that there is a missing edge between node i and node j. Performance bounds for estimating Ω under dependence is useful for statistical learning problems. For direct estimation of precision matrices that can preserve sparsity, one can adopt entry-wise 1-norm penalized likelihood approaches; see Banerjee, El Ghaoui and d'Aspremont (2008), Fan, Feng and Wu (2009), Friedman, Hastie and Tibshirani (2008), Ravikumar et al. (2011), Rothman et al. (2008), which we refer them as Lasso-type precision matrix estimators. Friedman, Hastie and Tibshirani (2008) proposed a graphical Lasso model and developed a computationally efficient and scalable algorithm for estimating large precision matrices. This 1-norm penalized multivariate Gaussian likelihood approach was also considered by Banerjee, El Ghaoui and d'Aspremont (2008). Consistency of the graphical Lasso were studied in Ravikumar et al. (2011), Rothman et al. (2008).

The precision matrix estimation procedure considered here is the graphical Lasso model [Friedman, Hastie and Tibshirani (2008)] which minimizes the objective function

$$(49) \qquad \hat{\Omega}_n(\lambda) = \arg\min_{\Psi \succ 0} \{\text{tr}(\Psi \hat{\Sigma}_n) - \log\det(\Psi) + \lambda|\Psi|_1\},$$

where λ is the penalty to be determined later. In (49) $\Psi \succ 0$ means that Ψ is positive-definite. Here we assume the maximum eigenvalue

$$(50) \qquad \rho(\Omega) \leq \varepsilon_0^{-1} \qquad \text{for some } \varepsilon_0 > 0,$$

or equivalently the minimum eigenvalue of Σ is larger than ε_0. Note that we do not assume the minimum eigenvalue of Ω is uniformly bounded below from zero. To introduce an asymptotic theory for the estimate $\hat{\Omega}_n$, we recall (6) and (7) of Theorem 2.1 for the definition of the functions $H(\cdot)$ and $G(\cdot)$ and also $\tilde{\alpha}$ and $\tilde{\beta}$. An analogue of the function $D(\cdot)$ in this context is

$$(51) \qquad D^*(u) = \frac{1}{p^2} \sum_{j,k=1}^{p} u(u \wedge |\omega_{jk}|).$$

Recall Corollary 2.2 for $\tilde{G}(\cdot)$.

It is interesting and surprising to note that the structure of Theorem 3.1 is very similar to that in Theorem 2.1. However, the main idea for the proof of Theorem 3.1 seems quite different, and our key argument here is based on convex minimization. It is also interesting to note that our rate of convergence is expressed in terms of the $\mathcal{L}^2$ norm; see (52), while in the previous literature probabilistic bounds are obtained; see Lam and Fan (2009), Ravikumar et al. (2011), Rothman et al. (2008). The constant C in Theorem 3.1 can be the same as the one in Theorem 2.1.

THEOREM 3.1. *Let the moment and the dependence conditions in Theorem 2.1 be satisfied and $\lambda = 4u$. Then*

$$(52) \qquad \frac{1}{p^2} \mathbb{E}|\hat{\Omega}_n(\lambda) - \Omega|_F^2 \lesssim D^*(u) + \min\left(\frac{1}{n}, \frac{u^{2-q}}{n^{q/2}}, H(u) + G(Cu)\right),$$

where C is independent of u, n and p. Let $u_\flat$ be the solution to the equation

$$(53) \qquad D^*(u_\flat) = \min(n^{-1}, \max(\tilde{G}(u_\flat), H(u_\flat))).$$

Then $\inf_{\lambda > 0} p^{-2} \mathbb{E}|\hat{\Omega}_n(\lambda) - \Omega|_F^2 \lesssim D^*(u_\flat)$.

REMARK 3. As an immediate consequence of Theorem 3.1, if the entries ω_{jk} of the inverse matrix Ω satisfy (42) with $0 \leq r < 1$, then we have by the argument in (39) that $D^*(u) \leq Cu^{2-r}M/p^2$. Similarly, if ω_{jk} satisfy (43), then $D^*(u) \leq Cu^2 M \log^r(2+u^{-1})$. Therefore Corollaries 2.4 and 2.5 are still valid in the context of precision matrix estimation.

PROOF OF THEOREM 3.1. Using $\Psi = \Omega + \Delta$, we see that $\hat{\Delta}_n = \hat{\Omega}_n(\lambda) - \Omega$ minimizes

$$G(\Delta) = \text{tr}(\Delta \hat{\Sigma}_n) - \log \det(\Psi) + \lambda|\Psi|_1 + \log \det(\Omega) - \lambda|\Omega|_1.$$

Hence $G(\hat{\Delta}_n) \leq G(0) = 0$. Let $\Omega_v = \Omega + v\Delta$. By Taylor's expansion,

$$(54) \qquad \begin{aligned} G(\Delta) = {} & \text{tr}[\Delta(\hat{\Sigma}_n - \Sigma)] + \lambda(|\Omega + \Delta|_1 - |\Omega|_1) \\ & + \text{vec}(\Delta)^\top\left[\int_0^1 (1-v)\Omega_v^{-1} \otimes \Omega_v^{-1}\,dv\right]\text{vec}(\Delta), \end{aligned}$$

where $\otimes$ denotes the Kronecker product. Write $\Xi = \hat{\Sigma}_n - \Sigma = (\xi_{jk})$, $\mathcal{S}_u = \{(j,k) : |\omega_{jk}| \geq u\}$ and $\mathcal{W}_u = \{(j,k) : |\xi_{jk}| \geq u\}$. Let $\mathcal{W}_u^c$ be the complement of $\mathcal{W}_u$. Then

$$(55) \qquad \mathrm{tr}(\Delta \Xi) = \mathrm{tr}(\Delta \Xi_{\mathcal{W}_u}) + \mathrm{tr}(\Delta \Xi_{\mathcal{W}_u^c}) \geq -|\Delta|_F |\Xi_{\mathcal{W}_u}|_F - u|\Delta|_1,$$

where the matrix $\Xi_{\mathcal{W}_u} = (\xi_{jk} \mathbf{1}_{(j,k) \in \mathcal{W}_u})_{1 \leq j, k \leq p}$. Assume $\alpha > 1/2 - 1/q$. By (19),

$$(56) \qquad \mathbb{E}(|\Xi_{\mathcal{W}_u}|_F^2) \lesssim p^2 (n^{1-q} u^{2-q} + (n^{-1} + u^2) e^{-C_4 n u^2}) =: N(u)^2.$$

Using the arguments for Theorem 1 in Rothman et al. (2008), we have by (50) that

$$(57) \qquad \mathrm{vec}(\Delta)^\top \left[\int_0^1 (1-v) \Omega_v^{-1} \otimes \Omega_v^{-1} \, dv \right] \mathrm{vec}(\Delta) \geq \frac{1}{4} \varepsilon_0^2 |\Delta|_F^2,$$

and by letting the penalty $\lambda = 4u$ that

$$(58) \qquad \begin{aligned} & \lambda(|\Omega + \Delta|_1 - |\Omega|_1) - u|\Delta|_1 \\ & \qquad \geq \lambda(|\Delta_{\mathcal{S}_u^c}^-|_1 - 2|\Omega_{\mathcal{S}_u^c}|_1 - |\Delta^+|_1 - |\Delta_{\mathcal{S}_u}^-|_1) - u|\Delta|_1 \\ & \qquad \geq 3u|\Delta_{\mathcal{S}_u^c}^-|_1 - 8u|\Omega_{\mathcal{S}_u^c}|_1 - 5u(|\Delta^+|_1 + |\Delta_{\mathcal{S}_u}^-|_1), \end{aligned}$$

where, for a matrix Σ, $\Sigma^+ = \mathrm{diag}(\Sigma)$ and $\Sigma^- = \Sigma - \Sigma^+$. By the Cauchy–Schwarz inequality, $|\Delta^+|_1 + |\Delta_{\mathcal{S}_u}^-|_1 \leq |\Delta|_F \sqrt{s_u}$, where $s_u = \#\mathcal{S}_u$. By (54)–(58),

$$(59) \qquad G(\Delta) \geq \frac{1}{4} \varepsilon_0^2 |\Delta|_F^2 - |\Delta|_F |\Xi_{\mathcal{W}_u}|_F - 8u|\Omega_{\mathcal{S}_u^c}|_1 - 5u|\Delta|_F \sqrt{s_u}.$$

Since $G(\hat{\Delta}_n) \leq 0$, there exists a deterministic constant $C > 0$ such that

$$(60) \qquad |\hat{\Delta}_n|_F^2 \leq C(|\Xi_{\mathcal{W}_u}|_F^2 + u^2 s_u + u|\Omega_{\mathcal{S}_u^c}|_1) \leq C(|\Xi_{\mathcal{W}_u}|_F^2 + p^2 D^*(u)).$$

Then (52) follows from (56) and by choosing u to minimize the right-hand side of (60); see the argument in (22). The case with $\alpha \leq 1/2 - 1/q$ can be similarly handled with special care (20) being taken into (56). $\square$

Ravikumar et al. (2011) studied the graphical Lasso estimator with off-diagonal entries penalized by the 1-norm. For i.i.d. p-variate vectors with polynomial moment condition, they showed that if $p = O((n/d^2)^{q/(2\tau)})$ for some $\tau > 2$, where d is the maximum degree in the Gaussian graphical model, then

$$(61) \qquad \frac{1}{p^2} |\hat{\Omega}_n - \Omega|_F^2 = O_P \left(\frac{s+p}{p^2} \cdot \frac{p^{2\tau/q}}{n} \right),$$

where s is the number of nonzero off-diagonal entries in Ω. For $\Omega \in \mathcal{H}_0(M)$, we can choose $M = s + p$. Note that $d \geq s/p$ and thus $d + 1 \geq M/p$. By Remark 3, Corollary 2.4 holds. Under case (ii) [resp., (iii)], our rate of convergence is $(M/p^2)^{1-2/q} n^{2/q-2}$ [resp., $n^{-1}(\log p) M/p^2$]. Elementary calculations show that both of our rates are of order $o(M p^{-2} n^{-1} p^{2\tau/q})$. Hence our bounds are much better than (61), the one obtained in Ravikumar et al. (2011).

We now compare our results with the CLIME (constrained L_1-minimization for inverse matrix estimation) method, a non-Lasso type estimator proposed in Cai, Liu and Luo (2011), which is to

$$(62) \qquad \text{minimize} \quad |\Theta|_1 \quad \text{subject to } |\hat{\Sigma}_n \Theta - I|_\infty \le \lambda_n.$$

Cai, Liu and Luo (2011) showed that with n i.i.d. p-variate observations, if $p = o(n^{q/2-1})$, then the rate of convergence for the CLIME estimator under the normalized Frobenius norm is $O(\tilde{C}^{4-2r}\tilde{M}(\log p/n)^{1-r/2})$, where $\tilde{C}$ is the upper bound for the matrix L_1-norm on the true precision matrix, and $\tilde{M}$ is in (45). We see that the rates of convergence under the normalized Frobenius norm are the same for both papers. This rate of convergence is in general better than those obtained for the Lasso-type estimators in the polynomial moment case [Ravikumar et al. (2011)].

REMARK 4. Following Rothman et al. (2008), we can consider the slightly modified version of the graphical Lasso: let $V = \text{diag}(\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2})$ and R be the correlation matrix; let $\hat{V}$ and $\hat{R}$ be their sample versions, respectively. Let $K = R^{-1}$. We estimate $\Omega = V^{-1}KV^{-1}$ by $\hat{\Omega}_\lambda = \hat{V}^{-1}\hat{K}_\lambda\hat{V}^{-1}$, where

$$(63) \qquad \hat{K}_\lambda = \arg\min_{\Psi \succ 0}\{\text{tr}(\Psi\hat{R}) - \log\det(\Psi) + \lambda|\Psi^-|_1\}.$$

Let $D^-(u) = p^{-2}\sum_{1 \le j \ne k \le p} u(u \wedge |\omega_{jk}|)$. Using the arguments of Theorem 2 in Rothman et al. (2008), we have the following result on the spectral norm rate of convergence of $\hat{\Omega}_\lambda$: Assuming the moment and dependence conditions in Theorem 3.1 are satisfied and $\varepsilon_0 \le \rho(\Omega) \le \varepsilon_0^{-1}$, and then

$$(64) \qquad \frac{\rho^2(\hat{\Omega}_\lambda - \Omega)}{p^2} \lesssim_\mathbb{P} D^-(\lambda) + \min\left(\frac{1}{n}, \frac{\lambda^{2-q}}{n^{q/2}}, H(\lambda) + G(C\lambda)\right)$$

holds if $\max[p^{1/q}n^{-1+1/q}, (\log p/n)^{1/2}] \lesssim \lambda$. Details of the derivation of (64) is given in the supplementary material [Chen, Xu and Wu (2013)]. If Ω satisfies $|\{(j,k): \omega_{jk} \ne 0, j \ne k\}| \le s$ [Rothman et al. (2008)], we have $\Omega \in \mathcal{H}_0(M)$ with $M = s + p$. Simple calculations show that, if $\alpha > 1/2 - 1/q$ and $s = O(p)$, then for $\lambda_\sharp \asymp \max[(\log p/n)^{1/2}, (s^{-1}p^2n^{1-q})^{1/q}]$, we have by (64) that $\rho(\hat{\Omega}(\lambda_\sharp) - \Omega) = O_\mathbb{P}(\sqrt{s}\lambda_\sharp)$, and it reduces to Theorem 2 in Rothman et al. (2008).

4. Evolutionary covariance matrix estimation for nonstationary high-dimensional processes. The time series processes considered in Sections 2 and 3 are stationary. In many situations the stationarity assumption can be violated, and the graphical structure is time-varying. One may actually be interested in how the covariance matrices and dependence structures vary with respect to time. Zhou, Lafferty and Wasserman (2010) and Kolar and Xing (2011) studied the estimation of covariance matrices for independent, locally stationary Gaussian processes. Both requirements can be quite restrictive in practice.

Here we shall consider nonstationary processes that can be both dependent and non-Gaussian with mild moment conditions, thus having a substantially broader spectrum of applicability. To allow such nonstationary processes, following the framework in Draghicescu, Guillas and Wu (2009), we shall consider locally stationary process

$$
(65) \qquad \mathbf{z}_i = \mathbf{g}(\mathcal{F}_i; i/n), \qquad 1 \le i \le n,
$$

where $\mathbf{g}(\cdot, \cdot) = (g_1(\cdot, \cdot), \ldots, g_p(\cdot, \cdot))^\top$ is a jointly measurable function such that the uniform stochastic Lipschitz continuity holds: there exists $C > 0$ for which

$$
(66) \qquad \max_{j \le p} \| g_j(\mathcal{F}_0; t) - g_j(\mathcal{F}_0; t') \| \le C |t - t'| \qquad \text{for all } t, t' \in [0, 1].
$$

In Examples 4.1–4.3 below we present some popular models of locally stationary processes. Let $\mathbf{z}_i^\diamond(t) = \mathbf{g}(\mathcal{F}_i; t)$. The preceding condition (66) suggests local stationarity in the sense that, for a fixed $t \in (0, 1)$ and bandwidth $b_n \to 0$ with $n b_n \to \infty$,

$$
(67) \qquad \max_{j \le p} \max_{\lfloor n(t-b_n) \rfloor \le i \le \lfloor n(t+b_n) \rfloor} \| \mathbf{z}_{j,i}^\diamond(t) - Z_{j,i} \| \le C b_n = o(1),
$$

indicating that the process $(\mathbf{z}_i)$ over the range $\lfloor n(t - b_n) \rfloor \le i \le \lfloor n(t + b_n) \rfloor$ can be approximated by the *stationary* process $\mathbf{z}_i^\diamond(t)$. The locally stationarity property suggests that the data generating mechanism $\mathbf{g}(\cdot; i/n)$ at time i is close to the one $\mathbf{g}(\cdot; i'/n)$ at time i' if $|i - i'|/n$ is small. Hence the following covariance matrix function is continuous:

$$
(68) \qquad \Sigma(t) = \operatorname{cov}(\mathbf{g}(\mathcal{F}_0; t)) = \mathbb{E}(\mathbf{z}(t)\mathbf{z}(t)^\top), \qquad t \in (0, 1).
$$

The covariance matrix $\Sigma_i = \Sigma(i/n)$ of $\mathbf{z}_i$ can then be estimated by the approximate stationary process $\mathbf{z}_l$, $\lfloor n(t - b_n) \rfloor \le l \le \lfloor n(t + b_n) \rfloor$, by using the Nadaraya–Watson or other smoothing techniques. Recall that in the stationary case the thresholded estimator is defined as $T_u(\hat{\Sigma}_n) = (\hat{\sigma}_{jk} \mathbb{I}(|\hat{\sigma}_{jk}| \ge u))_{jk}$, where $\hat{\Sigma}_n = (\hat{\sigma}_{jk})$ is the sample covariance matrix given in (1). To estimate $\Sigma(t)$, we substitute $\hat{\Sigma}_n$ by the kernel smoothed version

$$
(69) \qquad \hat{\Sigma}_n(t) = \sum_{m=1}^n w_m(t) \mathbf{z}_m \mathbf{z}_m^\top \qquad \text{where } w_m(t) = \frac{K((t - m/n)/b_n)}{\sum_{m=1}^n K((t - m/n)/b_n)}.
$$

Write $\hat{\Sigma}_n(t) = (\hat{\sigma}_{jk}(t))_{jk}$. In (69), $K(\cdot)$ is a symmetric, nonnegative kernel with bounded support in $[-1, 1]$ and $\int_{-1}^1 K(v)\, dv = 1$. As per convention, we assume that the bandwidth b_n satisfies the natural condition: $b_n \to 0$ and $n b_n \to \infty$. The thresholded covariance estimator for nonstationary processes is then defined as

$$
T_u(\hat{\Sigma}_n(t)) = (\hat{\sigma}_{jk}(t) \mathbb{I}(|\hat{\sigma}_{jk}(t)| \ge u))_{1 \le j,k \le p}.
$$

Parallelizing Theorem 2.1, we give a general result for the thresholded estimator for time-varying covariance matrices of the nonstationary, nonlinear high-dimensional time series. As in (4) and (5), we similarly define the functional dependence measure

$$\theta_{i,w,j} = \max_{0 \le t \le 1} \| Z_{ji}(t) - Z'_{ji}(t) \|_w, \tag{70}$$

where $Z'_{ji}(t) = g_j(\mathcal{F}'_i, t)$. We also assume that (5) holds. For presentational simplicity let $\alpha > 1/2 - 1/q$. Let $n_\sharp = nb_n$, $H_\sharp(u) = u^{2-q} n_\sharp^{1-q}$,

$$D(u) = \frac{1}{p^2} \max_{0 \le t \le 1} \sum_{j,k=1}^{p} (u^2 \wedge \sigma_{jk}(t)^2), \qquad G_\sharp(u) = (n_\sharp^{-1} + u^2) e^{-n_\sharp u^2}. \tag{71}$$

Theorem 4.1 provides convergence rates for the thresholded covariance matrix function estimator $T_u(\hat{\Sigma}_n(t))$. Due to the nonstationarity, the bound is worse than the one in Theorem 2.1 since we only use data in the local window $[n(t-b_n), n(t+b_n)]$. Therefore, in the nonstationary case a larger sample size is needed for achieving the same level of estimation accuracy.

THEOREM 4.1. *Assume* $\max_{1 \le j,k \le p} \sup_{t \in [0,1]} |\sigma''_{jk}(t)| < \infty$ *and* $\alpha > 1/2 - 1/q$. *Under the moment and dependence conditions of Theorem* 2.1, *we have*

$$\frac{\mathbb{E}|T_u(\hat{\Sigma}_n(t)) - \Sigma(t)|_F^2}{p^2} \lesssim D(u) + \min(n_\sharp^{-1}, H_\sharp(u) + G_\sharp(Cu)) + b_n^4 \tag{72}$$

uniformly over $t \in [b_n, 1 - b_n]$, *where* C *is independent of* u, n, b_n *and* p.

PROOF. Let $\Sigma_n^\circ(t) = \mathbb{E}\hat{\Sigma}_n(t) = (\sigma_{jk}^\circ(t))_{jk}$. Under the condition on $\sigma''_{jk}(t)$, we have $\sigma_{jk}^\circ(t) - \sigma_{jk}(t) = O(b_n^2)$ uniformly over j, k and $t \in [b_n, 1 - b_n]$. Hence $|\Sigma_n^\circ(t) - \Sigma(t)|_F^2/p^2 = O(b_n^4)$. It remains to deal with $\mathbb{E}|T_u(\hat{\Sigma}_n(t)) - \Sigma_n^\circ(t)|_F^2$. With a careful check of the proof of Theorem 2.1, if we replace $\hat{\sigma}_{jk}$ and σ_{jk} therein by $\hat{\sigma}_{jk}(t)$ and $\sigma_{jk}^\circ(t)$, respectively, then we can have

$$\frac{\mathbb{E}|T_u(\hat{\Sigma}_n(t)) - \Sigma^\circ(t)|_F^2}{p^2} \lesssim D(u) + \min(n_\sharp^{-1}, H_\sharp(u) + G_\sharp(Cu)) \tag{73}$$

if the following Nagaev inequality holds:

$$\mathbb{P}(|\hat{\sigma}_{jk}(t) - \sigma_{jk}^\circ(t)| > v) \le \frac{C_2 n_\sharp}{(n_\sharp v)^q} + C_3 e^{-C_4 n_\sharp v^2}. \tag{74}$$

The above inequality follows by applying the nonstationary Nagaev inequality in Section 4 in Liu, Xiao and Wu (2013) to the process $X_m = K((t - m/n)/b_n)(Z_{mj}Z_{mk} - \mathbb{E}(Z_{mj}Z_{mk}))$, $\lfloor n(t - b_n) \rfloor \le m \le \lfloor n(t + b_n) \rfloor$. Note that the functional dependence measure of the latter process is bounded by $\mu(\theta_{i,2q,j} + \theta_{i,2q,k}) \sup_u |K(u)|$; see (12) and (70). □

REMARK 5. If in (69) we use the local linear weights [Fan and Gijbels (1996)], then it is easily seen based on the proof of Theorem 4.1 that (72) holds over the whole interval $t \in [0, 1]$, and the boundary effect is removed. This applies to the Theorem 4.2 below as well.

A similar result can be obtained for estimating evolutionary precision matrices of high-dimensional nonstationary processes $\Omega(t) = \Sigma^{-1}(t)$ where $\Sigma(t)$ is given in (68). As in the stationary case, we assume that $\Omega(t)$ satisfies (50) for all $t \in [0, 1]$. The actual estimation procedure of $\Omega(t)$ based on the data $Z_{p \times n}$ is a variant of the graphical Lasso estimator of Ω, which minimizes the following objective function:

$$(75) \qquad \hat{\Omega}_n(t; \lambda) = \underset{\Psi \succ 0}{\arg \min} \{ \operatorname{tr}(\Psi \hat{\Sigma}_n(t)) - \log \det(\Psi) + \lambda |\Psi|_1 \},$$

where $\hat{\Sigma}_n(t)$ is the kernel smoothed sample covariance matrix given in (69). The same minimization program is also used in Kolar and Xing (2011), Zhou, Lafferty and Wasserman (2010). As in (51) and (71), let

$$(76) \qquad D^*(u) = \frac{1}{p^2} \max_{0 \le t \le 1} \sum_{j,k=1}^{p} u(u \wedge |\omega_{jk}(t)|).$$

As in (53), choose $\lambda = 4u_\flat^\sharp$. For the estimator (75), we have the following theorem. We omit the proof since it is similar to the one in Theorems 3.1 and 4.1.

THEOREM 4.2. *Assume* $\max_{1 \le j,k \le p} \sup_{t \in [0,1]} |\omega_{jk}''(t)| < \infty$ *and* $\alpha > 1/2 - 1/q$. *Under the moment and dependence conditions of Theorem* 2.1, *we have*

$$(77) \qquad \frac{\mathbb{E}|\hat{\Omega}_n(t; 4u) - \Omega(t)|_F^2}{p^2} \lesssim D^*(u) + \min(n_\sharp^{-1}, H_\sharp(u) + G_\sharp(Cu)) + b_n^4$$

uniformly over $t \in [b_n, 1 - b_n]$, *where* C *is independent of* u, n, b_n *and* p. *Let* $u_\flat^\sharp \ge n_\sharp^{-1/2}$ *be the solution to the equation* $\max(G_\sharp(u), H_\sharp(u)) = D^*(u)$. *Then* $\inf_{\lambda > 0} p^{-2} \mathbb{E}|\hat{\Omega}_n(t; \lambda) - \Omega(t)|_F^2 \lesssim D^*(u_\flat^\sharp)$.

EXAMPLE 4.1 (Modulated nonstationary process [Adak (1998)]). Let $(\mathbf{y}_i)$ be a stationary p-dimensional process with mean 0 and identity covariance matrix. Then the modulated process

$$(78) \qquad \mathbf{z}_i = \Sigma^{1/2}(i/n)\mathbf{y}_i,$$

has covariance matrix $\Sigma_i = \Sigma(i/n)$. Zhou, Lafferty and Wasserman (2010) considered the special setting in which $\mathbf{y}_i$ are i.i.d. standard Gaussian vectors, and hence $\mathbf{z}_i$ are independent.

EXAMPLE 4.2 (Nonstationary linear process). Consider the nonstationary linear process

$$(79) \qquad \mathbf{z}_i = \sum_{j=0}^{\infty} A_j(i/n)\mathbf{e}_{i-j}, \qquad 1 \le i \le n,$$

where $A_j(\cdot)$ are continuous matrix functions. We can view (79) as a time-varying version of (31), a framework also adopted in Dahlhaus (1997). As in Example 2.2, we assume a uniform version

$$(80) \qquad \max_{k \le p} \sum_{l=1}^{p} \max_{0 \le t \le 1} a_{j,kl}(t)^2 = O(j^{-2-2\gamma}), \qquad \gamma > 0.$$

EXAMPLE 4.3 (Markov chain example revisited: Nonstationary version). We consider a nonstationary nonlinear example adapted from Example 2.1. Let the process $(\mathbf{z}_i)$ be defined by the iterated random function

$$(81) \qquad \mathbf{z}_i = \mathbf{g}_i(\mathbf{z}_{i-1}, \mathbf{e}_i),$$

where $\mathbf{g}_i(\cdot, \cdot)$ is an $\mathbb{R}^p$-valued and jointly measurable function that may change over time. As in Example 2.1, we assume $\mathbf{g}_i$ satisfy: (i) there exists some $\mathbf{x}_0$ such that $\sup_i \|\mathbf{g}_i(\mathbf{x}_0, \mathbf{e}_0)\|_{2q} < \infty$; (ii)

$$L := \sup_i \mathbb{E}|L_i|^q < 1 \qquad \text{where } L_i = \sup_{\mathbf{x} \ne \mathbf{x}'} \frac{\|\mathbf{g}_i(\mathbf{x}, \mathbf{e}_0) - \mathbf{g}_i(\mathbf{x}', \mathbf{e}_0)\|_{2q}}{|\mathbf{x} - \mathbf{x}'|}.$$

Then $(\mathbf{z}_i)$ have the GMC property with $\Theta_{m,2q} = O(L^m)$. Therefore, Theorem 4.1 can be applied with $\alpha > 1/2 - 1/q$ and $\tilde{\beta} = 1$.

SUPPLEMENTARY MATERIAL

Additional proofs (DOI: 10.1214/13-AOS1182SUPP; .pdf). The supplementary file contains the proof of relation (64): spectral norm convergence rate for precision matrix.

REFERENCES

ABRAHAMSSON, R., SELEN, Y. and STOICA, P. (2007). Enhanced covariance matrix estimators in adaptive beamforming. In 2007 *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*) 969–972. Honolulu, HI.

ADAK, S. (1998). Time-dependent spectral analysis of nonstationary time series. *J. Amer. Statist. Assoc.* **93** 1488–1501. MR1666643

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. MR0091588

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

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

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

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

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

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

CAI, T. and ZHOU, H. (2013). Minimax estimation of large covariance matrices under ℓ_1-norm (with discussion). *Statist. Sinica* **22** 1319–1349. MR3027084

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

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

CHEN, X., XU, M. and WU, W. B. (2013). Supplement to "Covariance and precision matrix estimation for high-dimensional time series." DOI:10.1214/13-AOS1182SUPP.

DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916

DRAGHICESCU, D., GUILLAS, S. and WU, W. B. (2009). Quantile curve estimation and visualization for nonstationary time series. *J. Comput. Graph. Statist.* **18** 1–20. MR2511058

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

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

GUERCI, J. R. (1999). Theory and application of covariance matrix tapers for robust adaptive beamforming. *IEEE Trans. Signal Process.* **47** 977–985.

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

JACQUIER, E., POLSON, N. G. and ROSSI, P. E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J. Econometrics* **122** 185–212. MR2083256

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

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

KOLAR, M. and XING, E. (2011). On time varying undirected graphs. In *Proceedings of the* 14*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 2011 (*JMLR*), *Vol.* 15 407–415. Ft. Lauderdale, FL.

KONDRASHOV, D., KRAVTSOV, S., ROBERTSON, A. W. and GHIL, M. (2005). A hierachy of database ENSO models. *Journal of Climate* **18** 4425–4444.

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

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10** 603–621.

LI, J., STOCIA, P. and WANG, Z. (2003). On robust capon beamforming and diagonal loading. *IEEE Trans. Signal Process.* **51** 1702–1715.

LIU, W. and LUO, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Preprint. Available at arXiv:1203.3896.

LIU, W., XIAO, H. and WU, W. B. (2013). Probability and moment inequalities under dependence. *Statist. Sinica*. To appear. DOI: 10.5705/ss.2011.287.

MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. MR0208649

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

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

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

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

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. Springer, New York. MR1697409

TALIH, M. (2003). Markov random fields on time-varying graphs, with an application to portfolio selection. Ph.D. thesis, Yale Univ., ProQuest LLC, Ann Arbor, MI.

WARD, J. (1994). Space time adaptive processing for airborne radar. Technical Report 1015, MIT, Lincoln Lab, Lexington.

WIKLE, C. K. and HOOTEN, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *TEST* **19** 417–451. MR2745992

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

WU, W. B. (2007). Strong invariance principles for dependent random variables. *Ann. Probab.* **35** 2294–2320. MR2353389

WU, W. B. (2011). Asymptotic theory for stationary processes. *Stat. Interface* **4** 207–226. MR2812816

WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760

WU, W. B. and SHAO, X. (2004). Limit theorems for iterated random functions. *J. Appl. Probab.* **41** 425–436. MR2052582

XIAO, H. and WU, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466–493. MR3014314

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

ZHENG, Y. R., CHEN, G. and BLASCH, E. (2007). A normalized fractionally lower-order moment algorithm for space–time adaptive processing. In *IEEE Military Communications Conference, 2007* (*MILCOM* 2007) 1–6. Orlando, FL.

ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2010). Time varying undirected graphs. *Mach. Learn.* **80** 295–319. MR3108169

X. CHEN
DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: xhchen@illinois.edu

M. XU
W. B. WU
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 S. UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: mengyu@galton.uchicago.edu
         wbwu@galton.uchicago.edu