

## COVARIANCE AND PRECISION MATRIX ESTIMATION FOR HIGH-DIMENSIONAL TIME SERIES

BY XIAOHUI CHEN<sup>\*</sup>, MENGYU XU<sup>†</sup> AND WEI BIAO WU<sup>†</sup>

*University of Illinois at Urbana-Champaign<sup>\*</sup> and University of Chicago<sup>†</sup>*

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 (2005), 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 non-stationarity 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 modelling (Meinshausen and Bühlmann (2006); Ravikumar et al. (2008); 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 (Guerci (1999); Ward (1994); Li, Stocia and Wang (2003); Abrahamsson, Selen and Stoica (2007)). 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  $\Sigma_i$  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,  $n \rightarrow \infty$  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

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

---

*AMS 2000 subject classifications:* Primary 62H12; secondary 62M10

*Keywords and phrases:* High-dimensional inference, sparsity, covariance matrix, precision matrix, thresholding, Lasso, dependence, functional dependence measure, consistency, Nagaev inequality, non-stationary time series, spatial-temporal processes

is a consistent and well-behaved estimator of  $\Sigma$  and  $\hat{\Omega}_n = \hat{\Sigma}_n^{-1}$  is a natural and good estimator of  $\Omega$ . 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  $\Sigma$  in the sense that its eigenvalues do not converge to those of  $\Sigma$ ; see for example the Marchenko-Pastur law ([Marchenko 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  $\mathbf{z}_1, \dots, \mathbf{z}_n$  are i.i.d. and thus the covariance matrix  $\Sigma_i \equiv \Sigma$  is time-invariant. See for example, [Bickel and Levina \(2008b\)](#); [Cai, Zhang and Zhou \(2010\)](#); [Bickel and Levina \(2008a\)](#); [Cai and Zhou \(2012, 2011\)](#), 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 [Rothman et al. \(2008\)](#); [Friedman, Hastie and Tibshirani \(2008\)](#); [Banerjee, El Ghaoui and d'Aspremont \(2008\)](#); [Ravikumar et al. \(2008\)](#); [Fan, Feng and Wu \(2009\)](#). Other estimates include the Cholesky decomposition based method ([Wu and Pourahmadi \(2003\)](#); [Huang et al. \(2006\)](#)), neighborhood selection for sparse graphical models ([Liu and Luo \(2012\)](#); [Yuan \(2010\)](#); [Meinshausen and Bühlmann \(2006\)](#)), regularized likelihood approach ([Lam and Fan \(2009\)](#); [Fan, Feng and Wu \(2009\)](#)) and the sparse matrix transform ([Cao, Bacheга and Bouman \(2011\)](#)). [Xiao and Wu \(2012\)](#) considered covariance matrix estimation for univariate stationary processes.

The assumption that  $\mathbf{z}_1, \dots, \mathbf{z}_n$  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 non-stationary processes has been increasingly seen across a wide variety of research areas. In modelling 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 non-stationary clutter environment (Ward (1994); Guerci (1999)). 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) non-linear dynamics of data generating systems; (ii) temporally dependent and non-stationary observations; (iii) high-dimensionality of the parameter space; and (iv) non-Gaussian distributions. Therefore, combination of more flexible non-linear and non-stationary 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 non-stationary 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 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,b) 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  $C, C_1, C_2, \dots$  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  $\Sigma$ ,  $|\Sigma|_1 = \sum_{j,k} |\sigma_{jk}|$ ,  $|\Sigma|_\infty = \max_{j,k} |\sigma_{jk}|$ ,  $|\Sigma|_F = (\sum_{j,k} \sigma_{jk}^2)^{1/2}$  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

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

where  $\mathbf{g}(\mathcal{F}_i) = (g_1(\mathcal{F}_i), \dots, g_p(\mathcal{F}_i))^\top$  is an  $\mathbb{R}^p$ -valued measurable function,  $\mathcal{F}_i = (\dots, \mathbf{e}_{i-1}, \mathbf{e}_i)$  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  $\mathbf{z}_i = (Z_{1i}, \dots, Z_{pi})^\top$  and  $Z_{p \times n} = (\mathbf{z}_i)_{i=1}^n$ , the data matrix observed at time points  $i = 1, \dots, n$ . Here we shall consider estimation of the  $p \times p$  covariance matrix  $\Sigma = \text{cov}(\mathbf{z}_i)$  based on the realization  $\mathbf{z}_1, \dots, \mathbf{z}_n$ , while Section 3 concerns estimation of its inverse. We consider Frobenius and spectral norm convergence of the *thresholded estimator*

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

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  $\Sigma$ , we need to impose regularization conditions. In particular, we shall assume that  $\Sigma$  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  $j$ -th 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

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

where  $Z'_{ji} = g_j(\mathcal{F}'_i)$ ,  $\mathcal{F}'_i = (\dots, \mathbf{e}_{-1}, \mathbf{e}'_0, \mathbf{e}_1, \dots, \mathbf{e}_i)$  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 non-linear stationary processes. We shall assume that the short-range dependence (SRD) condition holds:

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

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 non-linear 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*

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

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

and

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

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

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

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  $\alpha \leq 1/2 - 1/q$ .  $\square$

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  $u_\diamond$  be the solution to the equation  $D(u) = \tilde{G}(u)$  over the interval  $u \in [n^{-1/2}, u_\diamond]$ , then  $\Upsilon \lesssim D(u_\diamond)$ . (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  $\alpha$ , 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 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

$$\begin{aligned} \mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2 &= \sum_{j,k=1}^p \mathbb{E}[\hat{\sigma}_{jk}\mathbb{I}(|\hat{\sigma}_{jk}| \geq u) - \sigma_{jk}]^2 \\ (10) \qquad \qquad \qquad &\leq 2 \sum_{j,k=1}^p \mathbb{E}(W_{jk}^2) + 2B(u/2), \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

Then

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

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

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

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

$$\begin{aligned}
\mathbb{E}[\xi_{jk}^2 \mathbb{I}(|\xi_{jk}| > v)] &= v^2 \mathbb{P}(|\xi_{jk}| > v) + \int_{v^2}^{\infty} \mathbb{P}(|\xi_{jk}| > \sqrt{w}) dw \\
&\leq v^2 \left[ \frac{C_2 n}{(nv)^q} + C_3 e^{-C_4 n v^2} \right] \\
&\quad + \int_{v^2}^{\infty} \left[ \frac{C_2 n}{(n\sqrt{w})^q} + C_3 e^{-C_4 n w} \right] dw \\
(19) \quad &= C_5 n^{1-q} v^{2-q} + C_3 ((C_4 n)^{-1} + v^2) e^{-C_4 n v^2},
\end{aligned}$$

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

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

Combining Cases I, II and III, by (11), (14)-(20), we have

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

$$(21) \quad + \min \left( \frac{1}{n}, \frac{u^{2-q}}{n^{q/2}}, H(u) + G(Cu) \right) =: M_0(u),$$

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 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:

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

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

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

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  $u_\dagger \rightarrow 0$ . Similarly we have (iii) since  $u_\natural = u_\diamond$ . 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  $v \geq n^{-1/2}$ , 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). □

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  $u_\Delta$  be the solution to the equation  $G(u) = u^{2-q}n^{-q/2}$ . Then the minimizer  $u_{\natural} \in [n^{-1/2}, u_\Delta]$  if  $D(u_\Delta) \geq \tilde{G}(u_\Delta)$  and  $u_{\natural} \in [u_\Delta, u_\diamond]$  if  $D(u_\Delta) \leq \tilde{G}(u_\Delta)$ . For  $\alpha > 1/2 - 1/q$ , (22) indicates that  $v^{2-q}n^{-q/2}$  is not needed; see also Remark 1.  $\square$

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*

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

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

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

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  $W_{jk}$ ,  $\xi_{jk}$  and  $A_{jk}^l, l = 1, 2, 3$ . Let matrices  $V_l = (W_{jk} \mathbb{I}(A_{jk}^l))_{j,k \leq p}$ . Similarly as (11), let  $B_*(u) = \max_{1 \leq k \leq p} \sum_{j=1}^p |\sigma_{jk}| \mathbb{I}(|\sigma_{jk}| < u)$ . Then

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

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),

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

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

$$(27) \quad |\rho(V_2)| \leq Q + u S_u \leq Q + 2D_*(u).$$

Using the idea of (17), we have

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

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

The bounds in Theorems 2.1 and 2.3 depend on the smallness measures, the moment order  $q$ , the dependence parameter  $\alpha$ , 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 non-linear process  $(\mathbf{z}_i)$  defined by the iterated random function

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

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)

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

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

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

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  $\mathbf{e}_i = (e_{1i}, \dots, e_{pi})^\top$ , 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 Rothman et al. (2008); Lam and Fan (2009); 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  $\beta = 1$ .

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

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

where  $0 < v \leq \sqrt{p} \varpi$  and  $\varpi^2$  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  $\sum_{j=1}^p (\hat{\lambda}_j - \mu_j)^2 \leq p^2 \varpi^2$ , and consequently

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

$$\leq 2 \sum_{j=1}^p (2\hat{\lambda}_j^2 \mathbf{1}_{\hat{\lambda}_j \leq 0} + 2v^2) + 2\varpi^2 p^2.$$

If  $\hat{\lambda}_j \leq 0$ , since  $\mu_i \geq 0$ , we have  $|\hat{\lambda}_j| \leq |\hat{\lambda}_j - \mu_i|$ . Then  $|\tilde{S}_v - \Sigma|_F^2 \leq 4v^2 p + 6\varpi^2 p^2 \leq 10\varpi^2 p^2$ . Note that the eigenvalues of  $\tilde{S}_v$  are bounded below by  $v$ , and thus it is positive definite. In practice we suggest using  $v = (p^{-1} \sum_{j,k=1}^p u^2 \mathbb{I}(|\hat{\sigma}_{jk}| \geq u))^{1/2}$ . The same positive-definitization 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  $\mathbf{z}_i = (Z_{1i}, \dots, Z_{pi})^\top$ ,  $1 \leq i \leq n$ , 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

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

where  $d(s_j^\circ, s_k^\circ)$  is a distance between sites  $s_j^\circ$  and  $s_k^\circ$  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  $m \rightarrow \infty$ ,

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

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

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

where  $\tau$  is the characteristic length-scale, and

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

Condition (36) outlines the geometry of the sites  $(s_j^\circ)_{j=1}^p$ , and  $\chi$  can be roughly interpreted as the correlation dimension. It holds with  $\chi = 2$  if  $s_j^\circ$  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:

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

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

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

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

$$(39) \quad \begin{aligned} D(u) &= u^2 F(u) + \frac{1}{p^2} \sum_{l=0}^{\infty} \sum_{j,k=1}^p \sigma_{jk}^2 \mathbb{I}(u2^{-l-1} \leq |\sigma_{jk}| < u2^{-l}) \\ &\leq u^2 F(u) + \sum_{l=0}^{\infty} (2^{-l}u)^2 F(2^{-l-1}u) \\ &\leq u^2 p^{-1} \min(p, Cu^{-\chi/K}) = u^2 \min(1, Cp^{-1}u^{-\chi/K}). \end{aligned}$$

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

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

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

$$\begin{aligned} \bar{D} &\leq \sum_{l=0}^{\infty} (2^{1+l}u_0)^2 F(2^l u_0) \\ &\leq \sum_{l=0}^{\infty} (2^{1+l}u_0)^2 Cp^{-1} (2^{1+l}u_0)^{-\chi/K} \leq Cp^{-2K/\chi}. \end{aligned}$$

Class (35) allows the  $\gamma$ -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

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

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  $\tau$ , 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  $0 < u \leq 1$ ,

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

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

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

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  $\Upsilon = O(u'^2_{\dagger}n^{1-q})$ . (iii) If  $p^2/M = o(n^{q-1})$  and

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

then the equation  $u^{2-r}M/p^2 = u^2e^{-nu^2}$  has solution  $u'_\circ \asymp [n^{-1} \log(2 + p^2M^{-1}n^{-r/2})]^{1/2}$  and  $\Upsilon = O(u'^2_{\circ}M/p^2)$ . (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. Similarly as (39), we have  $D(u) \leq \min(u^2, Cu^{2-r}M/p^2)$ . Note that the solution  $u_\circ \geq n^{-1/2}$  to the equation  $H(u) = G(u)$  satisfies  $u_\circ \sim ((q/2-1)n^{-1} \log n)^{1/2}$ . Then by Corollary 2.2, (i)-(iv) follow from elementary but tedious manipulations. Details are omitted.  $\square$

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  $u_\circ < u'_\dagger$ .

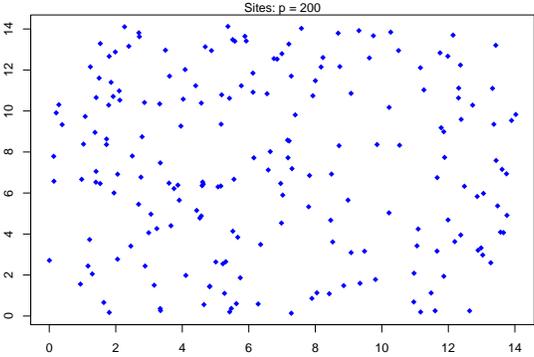
**COROLLARY 2.5.** *Assume (43). Let conditions in Theorem 2.1 be satisfied with  $\alpha > 1/2 - 1/q$  and  $\Upsilon = p^{-2} \sup_{\Sigma \in \mathcal{L}_r(M)} \inf_{u > 0} \mathbb{E}|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$ . (i) If  $n^{q-1} = O(p^2/M)$ , then for  $u \asymp 1$ ,  $\Upsilon = O(H(u)) = O(n^{1-q})$ . (ii) If  $p^2/M = o(n^{q-1})$  and  $n^{q/2-1}(\log n)^{r+q/2} \leq p^2/M$ , let  $\epsilon_{\dagger} = n^{1-q}p^2/M$  and  $u'_{\dagger} = \epsilon_{\dagger}^{1/q}(\log(2 + \epsilon_{\dagger}^{-1}))^{-r/q}$ . Then  $\Upsilon = O(u'_{\dagger}{}^{2-q}n^{1-q})$ . (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  $\eta$  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  $u^q \log^r(2 + u^{-1}) = \epsilon$  has solution  $u_{\dagger} \asymp \epsilon_{\dagger}^{1/q}(\log(2 + \epsilon_{\dagger}^{-1}))^{-r/q}$ . Under (iii), the equation  $u^2 p^{-2}M \log^r(2 + u^{-1}) = G(u)$  has solution  $\asymp u'_{\circ}$ .  $\square$

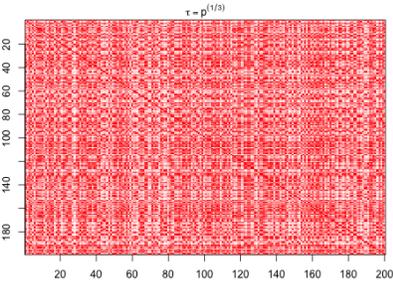
Corollaries 2.4 and 2.5 deal with the weaker dependence case with  $\alpha > 1/2 - 1/q$ . By Corollary 2.2, similar versions can be obtained for  $\alpha \leq 1/2 - 1/q$ . Details are omitted.

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). Figures 1(b), 1(c) and 1(d) show three  $200 \times 200$  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,  $u_{\circ}$  is the solution of  $G(u) = H(u)$ . Note that,  $u_{\dagger}$ , 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, i.e. the covariance matrix  $\Sigma$  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.  $\alpha > 1/4$ ).

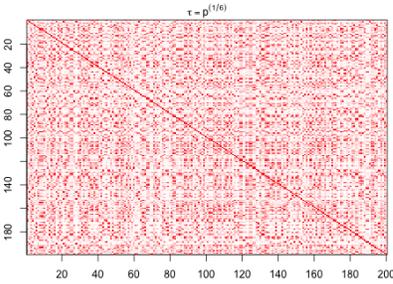
**2.4. Comparison with earlier results.** We now compare (42) with the commonly used sparsity condition defined in terms of the *strong  $\ell^q$ -ball*



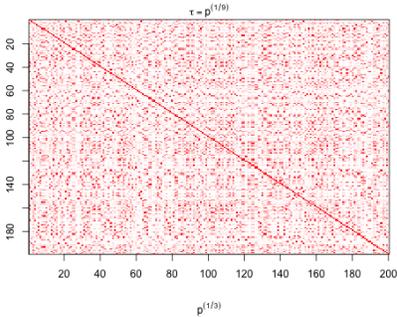
(a)  $p$  sites  $s_1^\circ, \dots, s_p^\circ$  uniformly sampled from the square  $p^{1/2} \times p^{1/2}$ .



(b)  $\Sigma: \tau = p^{1/3}$ .



(c)  $\Sigma: \tau = p^{1/6}$ .



(d)  $\Sigma: \tau = p^{1/9}$ .

FIG 1. Rational quadratic covariance matrix  $\Sigma$  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  $p^{1/9}$ .

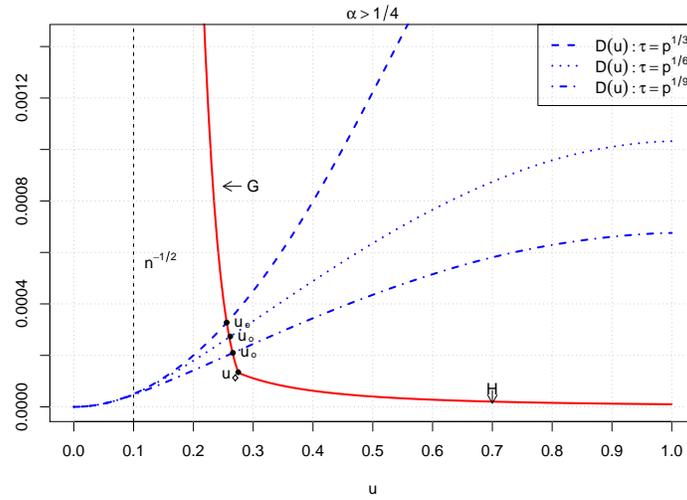
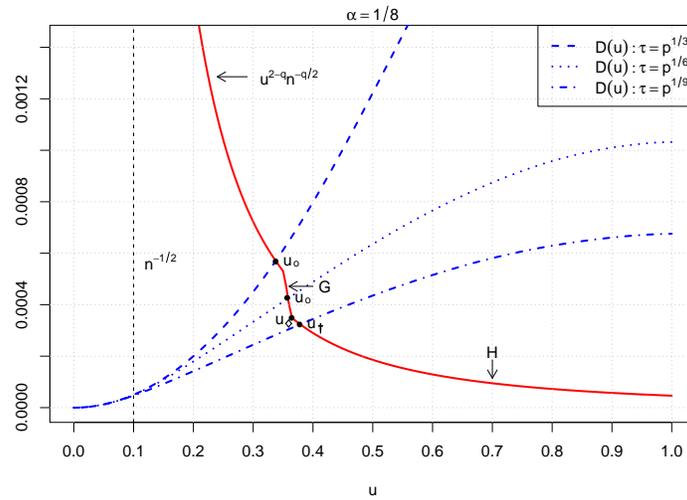
(a) Weaker temporal dependence with  $\alpha > 1/4$ .(b) Stronger temporal dependence with  $\alpha = 1/8$ .

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

(Bickel and Levina (2008a); Cai and Zhou (2012); Cai, Liu and Luo (2011))

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

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

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

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. in Bickel and Levina (2008a,b)) are of probabilistic nature; namely in the form  $|T_u(\hat{\Sigma}_n) - \Sigma|_F^2$  is bounded with large probability under the strong  $\ell^q$ -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

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

where  $0 < \varepsilon \leq (p-1)^{-1/2}$  so that  $\Sigma$  is positive-definite. Then for any threshold level  $u \in (\varepsilon, 1)$ ,  $\sum_{j,k=1}^p \mathbb{I}(|\sigma_{jk}| \geq u) = p$  and for any  $u \in (0, \varepsilon]$ ,  $\sum_{j,k=1}^p \mathbb{I}(|\sigma_{jk}| \geq u) = 3p - 2$ . 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  $\varepsilon \geq (p-1)^{\eta/r-1/r}$  with  $\eta \in (0, 1-r/2)$ .

With the strong  $\ell^q$ -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  $p^2 M^{-1} n^{-r/2} \geq p^\varepsilon$  for some  $\varepsilon > 0$ . Then*

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

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  $\max_j \|Z_{ji}\|_{2q} \leq \mu$ , they obtained the *probabilistic bound*

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

and  $C > 0$  is a sufficiently large constant. As a natural requirement for consistency, we assume  $u_{\text{BL}} \rightarrow 0$ , namely  $p = o(n^{q/4})$ . Since  $\Sigma \in \mathcal{G}_r(\tilde{M})$ , we have  $D_*(u) \leq \tilde{M} u^{1-r} =: \bar{D}_*(u)$  and  $N_*(u) \leq \min(p, \tilde{M} u^{-r}) =: \bar{N}_*(u)$ . Consider the weaker dependence case with  $\alpha > 1/2 - 1/q$ . Note that in (24)  $D_*(\cdot)$  is non-decreasing, while all other three functions are non-increasing. 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  $H_*(u) = pu^{1-q/2} n^{(1-q)/2} = \bar{D}_*(u)$ , 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 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 precision matrix estimate 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  $\Omega$  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 [Friedman, Hastie and Tibshirani \(2008\)](#); [Banerjee, El Ghaoui and d’Aspremont \(2008\)](#); [Ravikumar et al. \(2008\)](#); [Rothman et al. \(2008\)](#); [Fan, Feng and Wu \(2009\)](#), 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 [Rothman et al. \(2008\)](#); [Ravikumar 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

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

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

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

or equivalently the minimum eigenvalue of  $\Sigma$  is larger than  $\varepsilon_0$ . Note that we do not assume the minimum eigenvalue of  $\Omega$  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

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

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 as the one 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 Ravikumar et al. (2008); Rothman et al. (2008); Lam and Fan (2009). 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*

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

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

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

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

**REMARK 3.** As an immediate consequence of Theorem 3.1, if the entries  $\omega_{jk}$  of the inverse matrix  $\Omega$  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  $\omega_{jk}$  satisfy (43), then  $D^*(u) \leq Cu^2M \log^r(2 + u^{-1})$ . Therefore Corollaries 2.4 and 2.5 are still valid in the context of precision matrix estimation.  $\square$

**PROOF.** 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,

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

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

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

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),

$$(56) \quad \mathbb{E}(|\Xi_{\mathcal{W}_u}|_F^2) \lesssim p^2(n^{1-q}u^{2-q} + (n^{-1} + u^2)e^{-C_4nu^2}) =: N(u)^2.$$

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

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

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

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

where, for a matrix  $\Sigma$ ,  $\Sigma^+ = \text{diag}(\Sigma)$  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),

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

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

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

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. (2008) 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

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

where  $s$  is the number of non-zero off-diagonal entries in  $\Omega$ . 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  $o(Mp^{-2}n^{-1}p^{2r/q})$ . Hence our bounds are much better than (61), the one obtained in [Ravikumar et al. \(2008\)](#).

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

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

[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. \(2008\)](#)).

REMARK 4. Following [Rothman et al. \(2008\)](#), we can consider the slightly modified version of the graphical Lasso: let  $V = \text{diag}(\sigma_{11}^{1/2}, \dots, \sigma_{pp}^{1/2})$  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

$$(63) \quad \hat{K}_\lambda = \arg \min_{\Psi \succ 0} \{\text{tr}(\Psi\hat{R}) - \log \det(\Psi) + \lambda|\Psi^{-1}|_1\}.$$

Let  $D^-(u) = p^{-2} \sum_{1 \leq j \neq k \leq p} u(u \wedge |\omega_{jk}|)$ . 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  $\varepsilon_0 \leq \rho(\Omega) \leq \varepsilon_0^{-1}$ , then

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

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. If  $\Omega$  satisfies  $|\{(j, k) : \omega_{jk} \neq 0, j \neq k\}| \leq s$  ([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  $\lambda_\# \asymp \max[(\log p/n)^{1/2}, (s^{-1}p^2n^{1-q})^{1/q}]$ , we have by (64) that  $\rho(\hat{\Omega}(\lambda_\#) - \Omega) = O_{\mathbb{P}}(\sqrt{s}\lambda_\#)$  and it reduces to Theorem 2 in [Rothman et al. \(2008\)](#).

#### 4. Evolutionary Covariance Matrix Estimation for Non-Stationary 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 non-stationary processes that can be both dependent and non-Gaussian with mild moment conditions, thus having a substantially broader spectrum of applicability. To allow such non-stationary processes, following the framework in Draghicescu, Guillas and Wu (2009), we shall consider locally stationary process

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

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

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

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  $b_n \rightarrow 0$  with  $nb_n \rightarrow \infty$ ,

$$(67) \quad \max_{j \leq p} \max_{[n(t-b_n)] \leq i \leq [n(t+b_n)]} \|\mathbf{z}_{j,i}^\diamond(t) - Z_{j,i}\| \leq Cb_n = o(1),$$

indicating that the process  $(\mathbf{z}_i)$  over the range  $[n(t-b_n)] \leq i \leq [n(t+b_n)]$  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:

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

The covariance matrix  $\Sigma_i = \Sigma(i/n)$  of  $\mathbf{z}_i$  can then be estimated by the approximate stationary process  $\mathbf{z}_l$ ,  $[n(t-b_n)] \leq l \leq [n(t+b_n)]$ , by using the Nadaraya-Watson or other smoothing techniques. Recall that in the stationary case the thresholded estimator is defined as  $T_u(\hat{\Sigma}_n) = (\hat{\sigma}_{jk} \mathbb{I}(|\hat{\sigma}_{jk}| \geq$

$u))_{jk}$ , 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

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

Write  $\hat{\Sigma}_n(t) = (\hat{\sigma}_{jk}(t))_{jk}$ . In (69),  $K(\cdot)$  is a symmetric, non-negative 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:  $b_n \rightarrow 0$  and  $nb_n \rightarrow \infty$ . The thresholded covariance estimator for non-stationary processes is then defined as

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

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

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

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  $n_{\#} = nb_n$ ,  $H_{\#}(u) = u^{2-q} n_{\#}^{1-q}$ ,

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

Theorem 4.1 provides convergence rates for the thresholded covariance matrix function estimator  $T_u(\hat{\Sigma}_n(t))$ . Due to the non-stationarity, 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 non-stationary case a larger sample size is needed for achieving the same level of estimation accuracy.

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

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

*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  $\sigma_{jk}$  therein by  $\hat{\sigma}_{jk}(t)$  and  $\sigma_{jk}^\circ(t)$ , respectively, then we can have

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

if the following Nagaev inequality holds:

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

The above inequality follows by applying the non-stationary Nagaev inequality in Section 4 in Liu, Xiao and Wu (2013) to the process  $X_m = K((t - m/n)/b_n)(Z_{mj}Z_{mk} - \mathbb{E}(Z_{mj}Z_{mk}))$ ,  $\lfloor n(t - b_n) \rfloor \leq m \leq \lfloor n(t + b_n) \rfloor$ . 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).  $\square$

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.  $\square$

A similar result can be obtained for estimating evolutionary precision matrices of high-dimensional non-stationary 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  $\Omega$ , which minimizes the following objective function

$$(75) \quad \hat{\Omega}_n(t; \lambda) = \arg \min_{\Psi > 0} \{\text{tr}(\Psi \hat{\Sigma}_n(t)) - \log \det(\Psi) + \lambda |\Psi|_1\},$$

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

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

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

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

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

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

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

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

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, hence  $\mathbf{z}_i$  are independent.

**EXAMPLE 4.2.** (Non-stationary Linear Process) Consider the non-stationary linear process

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

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

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

**EXAMPLE 4.3.** (Markov Chain Example Revisited: Non-stationary Version) We consider a non-stationary non-linear example adapted from Example 2.1. Let the process  $(\mathbf{z}_i)$  be defined by the iterated random function

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

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)

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

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  $\beta = 1$ .

**Acknowledgments.** We thank two anonymous referees, an AE and the Editor for their helpful comments that have improved the paper.

## References.

- ABRAHAMSSON, R., SELEN, Y. and STOICA, P. (2007). Enhanced Covariance Matrix Estimators in Adaptive Beamforming. *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 969-972.
- ADAK, S. (1998). Time-Dependent Spectral Analysis of Nonstationary Time Series. *Journal of American Statistical Association* **93** 1488-1501.
- ANDERSON, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley Publications in Statistics. John Wiley & Sons Inc., New York. [MR0091588 \(19,992a\)](#)
- BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research* **9** 485-516.
- BICKEL, P. J. and LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989-1010.
- BICKEL, P. J. and LEVINA, E. (2008a). Covariance Regularization by Thresholding. *The Annals of Statistics* **36** 2577-2604.
- BICKEL, P. J. and LEVINA, E. (2008b). Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics* **36** 199-227.
- CAI, T., LIU, W. and LUO, X. (2011). A Constrained  $\ell_1$  Minimization Approach to Sparse Precision Matrix Estimation. *Journal of American Statistical Association* **106** 594-607.
- CAI, T., ZHANG, C.-H. and ZHOU, H. (2010). Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics* **38** 2118-2144.
- CAI, T. and ZHOU, H. (2011). Minimax Estimation of Large Covariance Matrices under  $\ell_1$ -norm (with discussion). *Statistica Sinica*, to appear.
- CAI, T. and ZHOU, H. (2012). Optimal Rates of Convergence for Sparse Covariance Matrix Estimation. *The Annals of Statistics*, to appear.
- CAO, G., BACHEGA, L. R. and BOUMAN, C. A. (2011). The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals. *IEEE Transactions on Image Processing* **20** 625-640.
- DAHLHAUS, R. (1997). Fitting Time Series Models to Nonstationary Processes. *The Annals of Statistics* **25** 1-37.
- DRAGHICESCU, D., GUILLAS, S. and WU, W. B. (2009). Quantile Curve Estimation and Visualization for Nonstationary Time Series. *Journal of Computational and Graphical Statistics* **18** 1-20.
- FAN, J., FENG, Y. and WU, Y. (2009). Network Exploration via the Adaptive Lasso and SCAD penalties. *The Annals of Applied Statistics* **3** 521-541.
- FAN, J. and GIJBELS, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability **66**. Chapman & Hall, London. [MR1383587 \(97f:62063\)](#)
- 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 Transactions on Signal Processing* **47** 977-985.

- HUANG, J., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance Matrix Selection and Estimation via Penalised Normal Likelihood. *Biometrika* **93** 85-98.
- JACQUIER, E., POLSON, N. and ROSSI, P. E. (2004). Bayesian Analysis of Stochastic Volatility Models with Fat-Tails and Correlated Errors. *Journal of Econometrics* **122** 185-212.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295-327. [MR1863961 \(2002i:62115\)](#)
- JOHNSTONE, I. M. and LU, A. Y. (2009). On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association* **104** 682-693.
- KOLAR, M. and XING, E. (2011). On time varying undirected graphs. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011 (JMLR)* **15** 407-415.
- KONDRASHOV, D., KRAVTSOV, S., ROBERTSON, A. W. and GHIL, M. (2005). A Hierarchy of Data-Based ENSO Models. *Journal of Climate* **18** 4425-4444.
- LAM, C. and FAN, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *The Annals of Statistics* **37** 4254-4278.
- 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 Transactions on Signal Processing* **51** 1702-1715.
- LIU, W. and LUO, X. (2012). High-dimensional Sparse Precision Matrix Estimation via Sparse Column Inverse Operator. *Preprint, arXiv:1203.3896*.
- LIU, W., XIAO, H. and WU, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica*.
- MARCHENKO, V. A. and PASTUR, L. A. (1967). Distribution of Eigenvalues for Some Sets of Random Matrices. *Mat. Sb. (N.S.)* **72** 507-536.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34** 1436-1462.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian processes for machine learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435 \(2010i:68131\)](#)
- RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electronic Journal of Statistics*.
- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics* **2** 494-515.
- STEIN, M. L. (1999). *Interpolation of spatial data. Springer Series in Statistics*. Springer-Verlag, New York. Some theory for Kriging. [MR1697409 \(2000f:62236\)](#)
- TALIH, M. (2003). Markov random fields on time-varying graphs, with an application to portfolio selection. *Thesis (Ph.D.)*. Yale University.
- WARD, J. (1994). Space Time Adaptive Processing for Airborne Radar. *Technical Report 1015, MIT, Lincoln Lab, Lexington* 977-985.
- WIKLE, C. K. and HOOTEN, M. B. (2010). A General Science-based Framework for Dynamical Spatio-Temporal Models. *Test* **19** 417-451.
- 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 \(2008j:60086\)](#)

- WU, W. B. (2011). Asymptotic theory for stationary processes. *Stat. Interface* **4** 207–226. [MR2812816 \(2012g:62401\)](#)
- WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844.
- WU, W. B. and SHAO, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability* **41** 425–436.
- XIAO, H. and WU, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466–493.
- YUAN, M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research* **11** 2261–2286.
- ZHENG, Y. R., CHEN, G. and BLASCH, E. (2007). A normalized Fractionally Lower-Order Moment Algorithm for Space-Time Adaptive Processing. *IEEE Military Communications Conference, 2007. (MILCOM 2007)* 1–6.
- ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2010). Time Varying Undirected Graphs. *Machine Learning* **80** 295–319.

DEPARTMENT OF STATISTICS  
725 S. WRIGHT STREET  
CHAMPAIGN, IL 61820  
E-MAIL: [xhchen@illinois.edu](mailto:xhchen@illinois.edu)

DEPARTMENT OF STATISTICS  
5734 S. UNIVERSITY AVENUE  
CHICAGO, IL 60637  
E-MAIL: [mengyu@galton.uchicago.edu](mailto:mengyu@galton.uchicago.edu)  
E-MAIL: [wbwu@galton.uchicago.edu](mailto:wbwu@galton.uchicago.edu)