Dynamic Covariance Models*

Ziqi Chen† and Chenlei Leng‡

Abstract

An important problem in contemporary statistics is to understand the relationship among a large number of variables based on a dataset, usually with p, the number of the variables, much larger than n, the sample size. Recent efforts have focused on modeling static covariance matrices where pairwise covariances are considered invariant. In many real systems, however, these pairwise relations often change. To characterize the changing correlations in a high dimensional system, we study a class of dynamic covariance models (DCMs) assumed to be sparse, and investigate for the first time a unified theory for understanding their non-asymptotic error rates and model selection properties. In particular, in the challenging high dimension regime, we highlight a new uniform consistency theory in which the sample size can be seen as $n^{4/5}$ when the bandwidth parameter is chosen as $h \propto n^{-1/5}$ for accounting for the dynamics. We show that this result holds uniformly over a range of the variable used for modeling the dynamics. The convergence rate bears the mark of the familiar bias-variance trade-off in the kernel smoothing literature. We illustrate the results with simulations and the analysis of a neuroimaging dataset.

Key Words: *Covariance model; Dynamic covariance; Functional connectivity; High Dimensionality; Marginal independence; Rate of convergence; Sparsity; Uniform consistency.*

*We thank the Joint Editor, an Associate Editor and two reviewers for their constructive comments. We also acknowledge the Neuro Bureau and the ADHD-200 consortium for making the fMRI dataset used in this paper freely available.

†Chen is with School of Mathematics and Statistics, Central South University (Email: chenzq453@gmail.com). Chen's research is supported in part by National Nature Science Foundation of China(no. 11401593), Specialized Research Fund for the Doctoral Program of Higher Education of China (no. 20130162120086), China Postdoctoral Science Foundation (no. 2013M531796), and China Postdoctoral Science Foundation (no. 2014T70778).

‡Corresponding author. Leng is with Department of Statistics, University of Warwick (Email: C.Leng@warwick.ac.uk).

1 Introduction

A common feature of contemporary data sets is their complexity in terms of dimensionality. To understand the relationship between the large number of variables in a complex dataset, a number of approaches are proposed to study a covariance matrix, or its inverse, sometimes with additional structures, built on the premise that many variables are either marginally independent or conditionally independent (Yuan and Lin, 2006; Bickel and Levina, 2008a,b; Rothman et al., 2009; Yuan, 2010; Cai and Liu, 2011; Cai et al., 2011; Chandrasekara et al., 2012; Xue et al., 2012; Cui et al., 2015).

The existing literature on estimating a sparse covariance matrix (or a sparse precision matrix) in high dimensions has an implicit assumption that this matrix is static, treating its entries as constant. Under this assumption, many methods were proposed to estimate the static covariance matrix consistently. Bickel and Levina (2008a) considered the problem of estimating this matrix by banding the sample covariance matrix, if a natural ordering of the variables or a notion of distance between the variables exists. Bickel and Levina (2008b), Rothman et al. (2009), Cai and Liu (2011) proposed to threshold a sample covariance matrix when an ordering of the variables is not available. Xue et al. (2012) developed a positive definite l_1-penalized covariance estimator. Yuan and Lin (2006) proposed a penalized likelihood based method for estimating sparse inverse covariance matrices. Yuan (2010) and Cai et al. (2011) estimated the static precision matrix under a sparsity assumption via linear programming. Guo et al. (2011) and Danaher et al. (2014) studied the estimation of multiple graphical models when several datasets are available.

The aforementioned papers all treat the matrix of interest as static. Thus, the resulting covariance matrix remains a constant matrix throughout the entire course of data collection. In reality, however, the assumption that the covariance matrix is constant may not be true. To illustrate this point, we considered the fMRI data collected by New York University Child Study Center. This particular experiment we considered has 172 scans from one subject, each recording the BOLD signals of 351 regions of interests (ROIs) in the brain. The detail of this dataset is discussed in Section 4.2. If we simply divide the dataset into two equal parts, with the first 86 scans in population one and the remaining 86 scans in population two, a formal test of the equality of the two 351×351 covariance matrices gives a p-value less than 0.001 (Li and Chen,

2012). This result implies that it may be more appropriate to use different covariance matrices for the scans at different time. Indeed, it is reasonable to expect that in a complex data situation, the covariance matrix often behaves dynamically in response to the changes in the underlying data collection process. This phenomenon is severely overlooked when the dimensionality is high. Other important examples where the underlying processes are fundamentally varying include genomics, computational social sciences, and financial time series (Kolar et al., 2010).

The main purpose of this paper is to develop a general class of dynamic covariance models (DCMs) for capturing the dynamic information in large covariance matrices. In particular, we make use of kernel smoothing (Wand and Jones, 1995; Fan and Gijbels, 1996) for estimating the covariance matrix locally, and apply entrywise thresholding afterwards to this locally estimated matrix for achieving consistency uniformly over the variable used for capturing dynamics. The proposed method is simple, intuitive, easy to compute, and as demonstrated by the non-asymptotic analysis, possesses excellent theoretical properties. To the best of our knowledge, this is the first piece of work that demonstrates the power of nowadays classical kernel smoothing in estimating large covariance matrices. Yin et al. (2010) studied this model for fixed dimensional problems without considering sparsity. In high dimensional cases, Ahmed and Xing (2009) and Kolar et al. (2010) considered time-varying networks by presenting pointwise convergence and model selection results. Zhou et al. (2010) estimated the time-varying graphical models. These papers established theoretical results at each time point. However, it is not clear whether the results provided in these papers hold simultaneously over the entire time period under consideration. This concern greatly hinders the use of dynamic information for modeling high dimensional systems. In contrast, we show that the convergence rate of the estimated matrices via the proposed approach holds uniformly over a compact set of the dynamic variable of interest. A detailed analysis reveals the familiar bias-variance trade-off commonly seen in kernel smoothing (Fan and Gijbels, 1996; Yu and Jones, 2004). In particular, by choosing a bandwidth parameter as $h \propto n^{-1/5}$, the effective sample size becomes of order $n^{4/5}$ which is used for estimating a covariance matrix locally. Although this result matches that when the dimension is fixed, we allow the dimensionality p to be exponentially high compared to the sample size n and the conclusion holds uniformly. This is the first rigorous uniform consistency result combining the strength of kernel smoothing and high dimensional covariance matrix modeling.

The rest of the article is organized as follows. In Section 2, we elaborate the proposed DCMs that require simply local smoothing and thresholding. A unified theory demonstrating that DCMs work for high dimensionality uniformly is presented in Section 3. In this section, we also discuss a simple ad-hoc method to obtain a positive definite estimate in case that thresholding gives a non-positive definite matrix. We present finite-sample performance of the DCMs by extensive simulation studies and an analysis of the fMRI data in Section 4. Section 5 gives concluding remarks. All the proofs are relegated to the Appendix and the Supplementary Materials.

2 The Model and Methodology

Let $Y = (Y_1, \cdots, Y_p)^T$ be a p-dimensional random vector and $U = (U_1, \cdots, U_l)^T$ be the associated index random vector. In many cases, a natural choice of U is time for modeling temporal dynamics. We write the conditional mean and the conditional covariance of Y given U as $m(U) = (m_1(U), \cdots, m_p(U))^T$ and $\Sigma(U)$, respectively, where $\Sigma_{jk}(U) = \mathrm{Cov}(Y_j, Y_k|U)$. That is, we allow both the conditional mean and the conditional covariance matrix to vary with U. When U denotes time, our model essentially states that both the mean and the covariance of the response vector are time dependent processes. Previous approaches for analyzing large dimensional data often assume $m(U) = m$ and $\Sigma(U) = \Sigma$, where both m and Σ are independent of U. Suppose that $\{Y_i, U_i\}$ with $Y_i = (Y_{i1}, \cdots, Y_{ip})^T$ is a random sample from the population $\{Y, U\}$, for $i = 1, \cdots, n$ with $n \ll p$. We are interested in the estimation of the conditional covariance matrix $\Sigma(u)$. In this paper, we focus on univariate dynamical variables where $l = 1$. The result can be easily extended to multivariate cases with $l > 1$ as long as l is fixed.

To motivate the estimates, recall that for fixed p, the usual consistent estimates of the static mean and covariance m and Σ can be written as $\hat{m} = n^{-1}\sum_{i=1}^n Y_i$ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n Y_i Y_i^T - \hat{m}\hat{m}^T$, respectively. The basic idea of kernel smoothing is to replace the weight $1/n$ for each observation in these two expressions by a weight that depends on the distance of the observation to the target point. By having larger weights for the observations closer to the target U, we estimate $m(U)$ and $\Sigma(U)$ locally in a loose sense. More specifically, by noting $m(U) = E(Y|U)$, we can estimate

$m(U)$ at $U = u$ as

$$\hat{m}(u) = \Big\{ \sum_{i=1}^{n} K_h(U_i - u) Y_i \Big\} \Big\{ \sum_{i=1}^{n} K_h(U_i - u) \Big\}^{-1}, \tag{1}$$

where $K_h(\cdot) = K(\cdot/h)/h$ for a kernel function $K(\cdot)$ and h is a bandwidth parameter (Wand and Jones, 1995; Fan and Gijbels, 1996). Similarly, a kernel estimate of $E(Y_{1j} Y_{1k}^T | U = u)$ is simply

$$\Big\{ \sum_{i=1}^{n} K_h(U_i - u) Y_{ij} Y_{ik}^T \Big\} \Big\{ \sum_{i=1}^{n} K_h(U_i - u) \Big\}^{-1},$$

which is consistent at each u under appropriate conditions (Wand and Jones, 1995; Fan and Gijbels, 1996). Putting these pieces together, we have the following empirical sample conditional covariance matrix

$$\begin{aligned} \hat{\Sigma}(u) := & \Big\{ \sum_{i=1}^{n} K_h(U_i - u) Y_i Y_i^T \Big\} \Big\{ \sum_{i=1}^{n} K_h(U_i - u) \Big\}^{-1} \\ & - \Big\{ \sum_{i=1}^{n} K_h(U_i - u) Y_i \Big\} \Big\{ \sum_{i=1}^{n} K_h(U_i - u) Y_i^T \Big\} \Big\{ \sum_{i=1}^{n} K_h(U_i - u) \Big\}^{-2}. \end{aligned} \tag{2}$$

Based on a normality assumption, Yin et al. (2010) derived a slightly different covariance estimator as

$$\hat{\Sigma}_1(u) = \Big[\sum_{i=1}^{n} K_h(U_i - u) \{ Y_i - \hat{m}(U_i) \} \{ Y_i - \hat{m}(U_i) \}^T \Big] \Big\{ \sum_{i=1}^{n} K_h(U_i - u) \Big\}^{-1}.$$

Both $\hat{\Sigma}(u)$ and $\hat{\Sigma}_1(u)$ are consistent for estimating $\Sigma(u)$ when p is fixed and n goes to infinity. However, in high-dimensional settings where the dimension p can vastly outnumber the sample size n, both $\hat{\Sigma}(u)$ and $\hat{\Sigma}_1(u)$ become singular, and neither can be used to estimate the inverse of a covariance matrix. With the increasing availability of large data, it is of great demand to develop new methods to estimate the dynamic covariance matrix with desirable theoretical properties.

Bickel and Levina (2008b) proposed to use hard thresholding on individual entries of the sample covariance matrix. They showed that as long as $\log p/n \to 0$, the thresholded estimator is consistent in the operator norm uniformly over the class of matrices satisfying their notion of sparsity. By using a generalized notion of shrinkage for thresholding that includes hard thresh-

olding, Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), and SCAD (Fan and Li, 2001) as special cases, Rothman et al. (2009) proposed the generalized thresholding operator. Denoted as a function $s_\lambda : R \to R$, this operator satisfies the following three conditions for all $z \in R$: (i) $|s_\lambda(z)| \leq |z|$; (ii) $s_\lambda(z) = 0$ for $z \leq \lambda$; (iii) $|s_\lambda(z) - z| \leq \lambda$. In particular, for hard thresholding, $s_\lambda(z) = zI(|z| \geq \lambda)$; for soft thresholding, $s_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$ with $(z)_+ = z$ if $z > 0$ and $(z)_+ = 0$ if $z \leq 0$; for adaptive lasso, $s_\lambda(z) = \text{sign}(z)(|z| - \lambda^2/|z|)_+$; and for SCAD, $s_\lambda(z)$ is the same as the soft thresholding if $|z| \leq 2\lambda$, and equals $\{2.7z - \text{sign}(z)3.7\lambda\}/1.7$ for $|z| \in [2\lambda, 3.7\lambda]$, and z if $|z| > 3.7\lambda$. We follow this notion of generalized shrinkage to construct our estimator as

$$s_{\lambda(u)}(\hat{\Sigma}(u)) = \left[s_{\lambda(u)}(\hat{\Sigma}_{jk}(u)) \right]_{p \times p},$$

where $\lambda(u)$ is a dynamic thresholding parameter depending on U. We collectively name our covariance estimates as Dynamic Covariance Models (DCMs) to emphasize the dependence of the conditional mean and the conditional covariance matrix on the dynamic index variable U. We remark that we allow the thresholding parameter λ to depend on the dynamic variable U. Thus, the proposed DCMs can be made fully adaptive to the sparsity levels at different U.

3 Theory

Throughout this paper, we implicitly assume $p \gg n$. We first present the exponential-tail condition (Bickel and Levina, 2008b) for deriving our asymptotic result. Namely, for $i = 1, \cdots, n$ and $j = 1, \cdots, p$, it is assumed that

$$Ee^{tY_{ij}^2} \leq K_1 < \infty, \quad \text{for} \quad 0 < |t| < t_0, \tag{3}$$

where t_0 is a positive constant. We establish the convergence of our proposed estimator in the matrix operator norm (spectral norm) defined as $||A||^2 = \lambda_{max}(AA^T)$ for a matrix $A = (a_{ij}) \in R^{p \times r}$. The following conditions are mild and routinely made in the kernel smoothing literature (Fan and Gijbels, 1996; Pagan and Ullah, 1999; Einmahl and Mason, 2005; Fan and Huang, 2005). These conditions may not be the weakest possible conditions for establishing the results of this paper, and are imposed to facilitate the proofs.

Regularity Conditions:

(a) We assume that $U_1, \cdots, U_n$ are independent and identically distributed from a pdf $f(\cdot)$ with compact support Ω. In addition, f is twice continuously differentiable and is bounded away from 0 on its support.

(b) The kernel function $K(\cdot)$ is a symmetric density function about 0 and has bounded variation. Moreover, $\sup_u K(u) < M_3 < \infty$ for a constant M_3.

(c) The bandwidth satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(d) All the components of the mean function $m(u)$ and all the entries of $\Sigma(u)$ have continuous second order derivatives. Moreover, $\sup_u E(Y_{ij}^4 | U_i = u) < M_4 < \infty$, for $i = 1, \cdots, n;\ j = 1, \cdots, p$, where M_4 is a constant.

The following result shows that the proposed estimator converges to the true dynamic covariance matrix uniformly over $u \in \Omega$, which holds uniformly over the set of the covariance matrices defined as

$$\mathcal{U}(q, c_0(p), M_2; \Omega) = \Big\{ \{\Sigma(u), u \in \Omega\} \mid \sup_{u \in \Omega} \sigma_{ii}(u) < M_2 < \infty, \sup_{u \in \Omega} (\sum_{j=1}^{p} |\sigma_{ij}(u)|^q) \le c_0(p),\ \forall i \Big\},$$

where Ω is a compact subset of R and $0 \le q < 1$. When $q = 0$,

$$\mathcal{U}(0, c_0(p), M_2; \Omega) = \Big\{ \{\Sigma(u), u \in \Omega\} \mid \sup_{u \in \Omega} \sigma_{ii}(u) < M_2 < \infty, \sup_{u \in \Omega} (\sum_{j=1}^{p} I\{\sigma_{ij}(u) \neq 0\}) \le c_0(p),\ \forall i \Big\}.$$

The dynamic covariance matrices $\Sigma(u)$ in $\mathcal{U}(q, c_0(p), M_2; \Omega)$ are assumed to satisfy the densest sparse condition over $u \in \Omega$, i.e., $\sup_{u \in \Omega} (\sum_{j=1}^{p} |\sigma_{ij}(u)|^q) \le c_0(p)$. Loosely speaking, for $q = 0$, the densest $\Sigma(u)$ over $u \in \Omega$ has at most $c_0(p)$ nonzero entries on each row. This condition is necessary, because otherwise, with a limited sample size, one can not estimate a dense covariance matrix well. Clearly, the family of covariance matrices over $u \in \Omega$ defined in $\mathcal{U}(q, c_0(p), M_2; \Omega)$ generalizes the notion of static covariance matrices in Bickel and Levina (2008b). We remark that the results presented in this paper apply to any compact subset $\Omega_1 \subset \Omega$ where $\{\Sigma(u), u \in \Omega_1\} \in \mathcal{U}(q, c_0(p), M_2; \Omega_1)$. Thus, even if the densest sparse condition fails on Ω, we can still

apply our method to sub-regions of Ω where this condition holds. We have the following strong uniform results for the consistency of the proposed DCMs.

Theorem 1. *[Uniform consistency in estimation] Under Conditions (a)–(d), suppose that the exponential-tail condition in (3) holds and that s_λ is a generalized shrinkage operator. Uniformly on $\mathcal{U}(q, c_0(p), M_2; \Omega)$, if $\lambda_n(u) = M(u)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$, $\frac{\log p}{nh} \to 0$ and $h^4 \log p \to 0$, we have*

$$\sup_{u\in\Omega} ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big),$$

where $M(u)$ depending on $u \in \Omega$ is large enough and $\sup_{u\in\Omega} M(u) < \infty$.

The proof of Theorem 1 can be found in the Appendix. The uniform convergence rate in Theorem 1 has the familiar bias-variance trade-off in the kernel smoothing literature (Wand and Jones, 1995; Fan and Gijbels, 1996), suggesting that the bandwidth should be selected carefully in order to balance bias and variance for optimally estimating the dynamic covariance matrices. In particular, for $q = 0$, the bias is bounded uniformly by $O(c_0(p)h^2\sqrt{\log p})$ and the variance is of the order $O_p(c_0(p)\sqrt{\frac{\log p}{nh}})$ uniformly. The existence of p here demonstrates clearly the dependence of these two quantities on the dimensionality, the bandwidth and the sample size. From this theorem, we immediately know that the optimal convergence rate is achieved when $h \propto n^{-1/5}$, consistent with the bandwidth choice in the traditional kernel smoothing (Wand and Jones, 1995; Fan and Gijbels, 1996). When the optimal bandwidth parameter $h \propto n^{-1/5}$ is adopted, the uniform convergence rate in Theorem 1 is $O_p(c_0(p)(\frac{\log p}{n^{4/5}})^{(1-q)/2})$. Thus, if $c_0(p)$ is bounded, we can allow the dimension to be of order $o(\exp(n^{4/5}))$ to have a meaningful convergent estimator. Intuitively, our result is also consistent with that of Bickel and Levina (2008b) and Rothman et al. (2009) for estimating a static sparse covariance matrix. A key difference is that the effective sample size for estimating each $\Sigma(u)$ can be seen as $nh \propto n^{4/5}$ with a bandwidth parameter $h \propto n^{-1/5}$ for accounting for the dynamics. Most importantly, our result holds uniformly over a range of the variable used for modeling dynamics. It is noted that the uniform convergence rate depends explicitly on how sparse the truth is through $c_0(p)$, and that the fundamental result underlying this theorem is Lemma 7. This is the first uniform result combining the strength of kernel smoothing and covariance matrix estimation in the challenging high dimensional regime.

We remark that, when the optimal bandwidth parameter $h \propto n^{-1/5}$ is used, the proposed

estimators are also optimal in the sense of Cai et al. (2010) and Cai and Zhou (2012) if $c_0(p)$ is bounded from above by a constant. This is discussed in detail in the Supplementary Materials. Our results generalize the results in these two papers to the situation where the entries in the covariance matrix vary with covariates.

Remark 1. The conditions on the kernel function are mild, and Theorem 1 holds for a wide range of kernel functions. The kernel function used only affects the convergence rate of the estimators up to multiplicative constants and thus has no impact on the rate of convergence. See, for example, Fan and Gijbels (1996) for a detailed analysis when the dimensionality is fixed.

Under appropriate conditions, shrinkage estimates of a large static covariance matrix are consistent in identifying the sparsity pattern (Lam and Fan, 2009; Rothman et al., 2009). The sparsity property of our proposed thresholding estimator of a dynamic covariance matrix also holds in a stronger sense that the proposed DCMs are able to identify the zero entries uniformly over U on a compact set Ω.

Theorem 2. *[Uniform consistency in estimating the sparsity pattern] Under Conditions (a)–(d), suppose that the exponential-tail condition in (3) holds, that s_λ is a generalized shrinkage operator, and that $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ for all i. If $\lambda_n(u) = M(u)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$ with $M(u)$ depending on $u \in \Omega$ large enough and satisfying $\sup_{u\in\Omega} M(u) < \infty$, $\frac{\log p}{nh} \to 0$, and $h^4 \log p \to 0$, we have*

$$s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) = 0 \quad \textit{for } \sigma_{jk}(u) = 0,\ \forall (j,k),$$

with probability tending to 1 uniformly in $u \in \Omega$. If we assume further that, for each $u \in \Omega$, all nonzero elements of $\Sigma(u)$ satisfy $|\sigma_{jk}(u)| > \tau_n(u)$ where $\frac{\log p}{nh \inf_{u\in\Omega}(\tau_n(u)-\lambda_n(u))^2} \to 0$, we have that

$$sign\{s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \cdot \sigma_{jk}(u)\} = 1 \quad \textit{for } \sigma_{jk}(u) \neq 0, \forall (j,k),$$

with probability tending to 1 uniformly in $u \in \Omega$.

Theorem 2 states that with probability going to one, the proposed DCMs can distinguish the zero and nonzero entries in $\Sigma(u)$. The conditions $|\sigma_{jk}(u)| > \tau_n(u)$ and $\frac{\log p}{nh \inf_{u\in\Omega}(\tau_n(u)-\lambda_n(u))^2} \to 0$ assure that the nonzero elements of $\Sigma(u)$ can be distinguished from the noise stochastically. As is

the case with the thresholding approach for estimating large covariance matrices, one drawback of our proposed approach is that the resulting estimator is not necessarily positive-definite. See Rothman (2012) and Xue et al. (2012) for some examples. To overcome this difficulty, we apply the following ad-hoc step. Let $-\hat{a}(u)$ be the smallest eigenvalue of $s_{\lambda_n(u)}(\hat{\Sigma}(u))$ when $\hat{a}(u) \geq 0$. Let $c_n = O\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big)$ be a positive number. To guarantee positive definiteness, we add $\hat{a}(u) + c_n$ to the diagonals of $s_{\lambda_n(u)}(\hat{\Sigma}(u))$; that is, we define a corrected estimator as

$$\hat{\Sigma}_C(u) = s_{\lambda_n(u)}(\hat{\Sigma}(u)) + \{\hat{a}(u) + c_n\}I_{p\times p},$$

where $I_{p\times p}$ is the $p\times p$ identity matrix. The smallest eigenvalue of $\hat{\Sigma}_C(u)$ is now $c_n > 0$. Therefore, $\hat{\Sigma}_C(u)$ is positive definite. If $s_{\lambda_n(u)}(\hat{\Sigma}(u))$ is already positive definite, no such correction is needed. Taking together, to guarantee positive definiteness, we define a modified estimator of $\Sigma(u)$ as

$$\hat{\Sigma}_M(u) = \hat{\Sigma}_C(u)I[\lambda_{min}\{s_{\lambda_n(u)}(\hat{\Sigma}(u))\} \leq 0] + s_{\lambda_n(u)}(\hat{\Sigma}(u))I[\lambda_{min}\{s_{\lambda_n(u)}(\hat{\Sigma}(u))\} > 0].$$

For any $u \in \Omega$ such that $\lambda_{min}\{s_{\lambda_n(u)}(\hat{\Sigma}(u))\} \leq 0$, it holds that

$$\hat{a}(u) \leq |-\hat{a}(u) - \lambda_{min}(\Sigma(u))| \leq ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)||.$$

Thus, we obtain immediately

$$||\hat{\Sigma}_C(u) - \Sigma(u)|| \leq ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| + \hat{a}(u) + c_n \leq 2\sup_u ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| + c_n.$$

We see that under the conditions in Theorem 1,

$$\sup_{u\in\Omega} ||\hat{\Sigma}_M(u) - \Sigma(u)|| = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big).$$

That is, the modified estimator of $\Sigma(u)$ is guaranteed to be positive definite, with the same convergence rate as that of the original thresholding estimator $s_{\lambda_n}(\hat{\Sigma}(u))$. Since the modified estimating procedure does not change the sparsity pattern of the thresholding estimator when n is large enough, Theorem 2 still holds for $\hat{\Sigma}_M(u)$.

Remark 2. Similar to Yin et al. (2010), Bickel and Levina (2008b), Cai and Liu (2011) and

Xue et al. (2012), we can also show the convergence rate of the proposed estimator under a polynomial-tail condition. To this end, assume that for some $\gamma > 0$ and $c_1 > 0$, $p = c_1 n^{\gamma}$, and for some $\tau > 0$,

$$\sup_{u\in\Omega} E(|Y_{ij}|^{5+5\gamma+\tau} \mid U_i = u) \leq K_2 < \infty, \quad \text{for} \quad i = 1, \cdots, n; j = 1, \cdots, p. \tag{4}$$

Let $h = c_2 n^{-1/5}$ for a positive constant c_2. Under the moment condition (4) and Conditions (a)–(d), if $\lambda_n(u) = M(u)\sqrt{\frac{\log p}{n^{4/5}}}$, we have that, uniformly on $\mathcal{U}(q, c_0(p), M_2; \Omega)$,

$$\sup_{u\in\Omega} ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| = O_p\Big(c_0(p)(\frac{\log p}{n^{4/5}})^{(1-q)/2}\Big), \tag{5}$$

where $M(u)$ depending on $u \in \Omega$ is large enough and $\sup_{u\in\Omega} M(u) < \infty$. The proof of (5) is found in the Appendix.

Define

$$\begin{aligned}\mathcal{U}(q, c_0(p), M_2, \epsilon; \Omega) \;=\; &\Big\{\{\Sigma(u), u \in \Omega\} \mid \{\Sigma(u), u \in \Omega\} \in \mathcal{U}(q, c_0(p), M_2; \Omega), \\ &\inf_u\{\lambda_{min}(\Sigma(u))\} \geq \epsilon > 0\Big\},\end{aligned}$$

which is a set consisting of only positive definite dynamic covariances in $\mathcal{U}(q, c_0(p), M_2; \Omega)$. From Theorem 1, we can derive that the inverse of the covariance matrix estimator converges to the true inverse with convergence rate $O_p(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q})$ uniformly in $u \in \Omega$. The detailed proof of Proposition 3 appears in the Appendix.

Proposition 3. *[Uniform consistency of the inverse of the estimated dynamic matrix] Under Conditions (a)–(d), suppose that the exponential-tail condition in (3) holds and that s_λ is a generalized shrinkage operator. If $\lambda_n(u) = M(u)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$, $\frac{\log p}{nh} \to 0$, $h^4 \log p \to 0$ and $c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q} \to 0$, we have that uniformly in $\mathcal{U}(q, c_0(p), M_2, \epsilon; \Omega)$,*

$$\sup_{u\in\Omega} ||[s_{\lambda_n(u)}(\hat{\Sigma}(u))]^{-1} - \Sigma^{-1}(u)|| = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big),$$

where $M(u)$ depending on $u \in \Omega$ is large enough and $\sup_{u\in\Omega} M(u) < \infty$.

Bandwidth selection and the choice of the threshold

The performance of the proposed DCMs depends critically on the choices of two tuning parameters: the bandwidth parameter h for kernel smoothing and the dynamic thresholding parameter $\lambda(u)$. We propose a simple two-step procedure for choosing them in a sequential manner. Namely, we first determine a data driven choice of the bandwidth parameter, followed by selecting $\lambda(u)$ at each point u.

Since the degree of smoothing is controlled by the bandwidth parameter h, a good bandwidth should reflect the smoothness of the true nonparametric functions $m(u)$ and $\Sigma(u)$. Importantly, in order to minimize the estimating error, the bandwidth parameter should be selected carefully to balance the bias and the variance of the estimate as in Theorem 1. In our implementation, we choose one bandwidth for $\hat{m}(\cdot)$ in (1) and another bandwidth for $\hat{\Sigma}(\cdot)$ in (2). For choosing the bandwidth in estimating the mean function $m(\cdot)$, we use the leave-one-out cross-validation approach in Fan and Gijbels (1996) and denote the chosen bandwidth as h_1. Next we discuss the bandwidth choice for estimating $\hat{\Sigma}(u)$ defined in (2). When the dimension p is large and the sample size n is small, $\hat{\Sigma}(u)$ is not positive definite. Therefore, the usual log-likelihood-type leave-one-out cross-validation (Yin et al., 2010) fails to work. Instead, we propose a subset-y-variables cross-validation procedure to overcome the effect of high dimensionality. Specifically, we choose k ($k < n$) y-variables randomly from $(y_1, \cdots, y_p)^T$ (i.e., $Y_s = (y_{j_1}, \cdots, y_{j_k})^T$) and repeat this N times. Denote $\text{var}(Y_s|U) = \Sigma_s(U)$ and define

$$CV(h) = \frac{1}{N}\sum_{s=1}^{N}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big[\{Y_{is} - \hat{m}_s(U_i)\}^T\hat{\Sigma}^{-1}_{s(-i)}(U_i)\{Y_{is} - \hat{m}_s(U_i)\} + \log(|\hat{\Sigma}_{s(-i)}(U_i)|)\Big]\Big\},$$

where $\hat{\Sigma}_{s(-i)}(\cdot)$ is estimated by leaving out the i-th observation according to (2) using responses Y_s with the bandwidth h, and $\hat{m}_s(u) = \{\sum_{i=1}^n K_{h_1}(U_i - u)Y_{is}\}\{\sum_{i=1}^n K_{h_1}(U_i - u)\}^{-1}$. The optimal bandwidth for estimating the dynamic covariance matrices is the value that minimizes $CV(h)$. We observed empirically that this choice gives good performance in the numerical study.

Now we consider how to choose $\lambda(u)$. For high-dimensional static covariance matrix estimation, Bickel and Levina (2008b) proposed to select the threshold by minimizing the Frobenius norm of the difference between the estimator after thresholding and the sample covariance matrix

Table 1: Average (standard error) MSLs and MFLs for Model 1

Method		MSL			MFL		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		3.18(0.24)	4.05(0.23)	5.69(0.25)	9.10(0.26)	13.5(0.31)	22.2(0.47)
Dynamic	Hard	1.21(0.13)	1.29(0.11)	1.41(0.12)	4.08(0.21)	5.15(0.23)	6.93(0.24)
	Soft	1.28(0.10)	1.37(0.09)	1.49(0.09)	4.61(0.22)	5.95(0.25)	8.31(0.23)
	Adaptive	1.17(0.12)	1.24(0.09)	1.36(0.11)	4.02(0.20)	5.07(0.23)	7.09(0.25)
	SCAD	1.23(0.11)	1.28(0.09)	1.37(0.10)	4.22(0.19)	5.37(0.21)	7.59(0.22)
Modified	Hard	1.23(0.14)	1.29(0.11)	1.43(0.10)	4.27(0.28)	5.48(0.32)	7.62(0.41)
	Soft	1.28(0.10)	1.37(0.09)	1.49(0.09)	4.61(0.22)	5.95(0.25)	8.31(0.23)
	Adaptive	1.17(0.12)	1.24(0.09)	1.36(0.11)	4.02(0.20)	5.07(0.23)	7.09(0.25)
	SCAD	1.23(0.11)	1.28(0.09)	1.37(0.10)	4.22(0.19)	5.37(0.21)	7.59(0.22)
Static	Hard	1.42(0.11)	1.51(0.11)	1.63(0.10)	4.86(0.19)	6.05(0.21)	8.25(0.25)
	Soft	1.37(0.08)	1.41(0.08)	1.51(0.09)	4.79(0.17)	6.11(0.27)	8.51(0.25)
	Adaptive	1.39(0.10)	1.43(0.10)	1.51(0.09)	4.79(0.13)	5.88(0.14)	7.79(0.16)
	SCAD	1.46(0.10)	1.51(0.09)	1.58(0.09)	5.06(0.14)	6.27(0.15)	8.34(0.16)

computed from an independent data. We adopt this idea. Specifically, we divide the original sample into two samples at random of size n_1 and n_2, where $n_1 = n(1 - \frac{1}{\log n})$ and $n_2 = \frac{n}{\log n}$, and repeat this N_1 times. Let $\hat{\Sigma}_{1,s}(u)$ and $\hat{\Sigma}_{2,s}(u)$ be the empirical dynamic covariance estimators according to (2) based on n_1 and n_2 observations respectively with the bandwidth selected by the subset-y-variables cross validation. Given u, we select the thresholding parameter $\hat{\lambda}(u)$ by minimizing

$$R(\lambda, u) := \frac{1}{N_1} \sum_{s=1}^{N_1} ||s_\lambda(\hat{\Sigma}_{1,s}(u)) - \hat{\Sigma}_{2,s}(u)||_F^2,$$

where $||M||_F^2 = \mathrm{tr}(MM^T)$ is the squared Frobenius norm of a matrix.

4 Numerical Studies

In this section we investigate the finite sample performance of the proposed procedure with Monte Carlo simulation studies. We compare our method to the static covariance matrix estimates in Rothman et al. (2009) when generalized thresholding, including hard, soft, adaptive lasso, and SCAD thresholding, is considered. We also include the modified estimator $\hat{\Sigma}_M(u)$ that guarantees positive definiteness and the empirical sample dynamic covariance matrix in (2) for comparison purposes. Throughout the numerical demonstration, the Gaussian kernel function $K(a) = \frac{1}{\sqrt{2\pi}} \exp(-a^2/2)$ was used for kernel smoothing.

Table 2: Average (standard error) MSLs and MFLs for Model 2

Method		MSL			MFL		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		2.81(0.18)	3.76(0.15)	5.20(0.22)	8.91(0.22)	13.5(0.29)	22.2(0.47)
Dynamic	Hard	1.11(0.08)	1.15(0.06)	1.23(0.06)	4.12(0.20)	5.33(0.23)	7.29(0.26)
	Soft	1.22(0.06)	1.27(0.05)	1.36(0.04)	4.33(0.16)	5.56(0.16)	7.64(0.20)
	Adaptive	1.11(0.07)	1.15(0.06)	1.26(0.05)	4.02(0.16)	5.13(0.16)	7.01(0.19)
	SCAD	1.09(0.06)	1.12(0.05)	1.20(0.05)	4.16(0.15)	5.34(0.14)	7.26(0.16)
Modified	Hard	1.11(0.08)	1.15(0.06)	1.23(0.06)	4.12(0.20)	5.33(0.23)	7.29(0.26)
	Soft	1.22(0.06)	1.27(0.05)	1.36(0.04)	4.33(0.16)	5.56(0.16)	7.64(0.20)
	Adaptive	1.11(0.07)	1.15(0.06)	1.26(0.05)	4.02(0.16)	5.13(0.16)	7.01(0.19)
	SCAD	1.09(0.06)	1.12(0.05)	1.20(0.05)	4.16(0.15)	5.34(0.14)	7.26(0.16)
Static	Hard	1.26(0.09)	1.29(0.07)	1.36(0.08)	4.75(0.18)	5.99(0.21)	7.97(0.23)
	Soft	1.24(0.06)	1.28(0.05)	1.38(0.04)	4.42(0.17)	5.65(0.17)	7.80(0.19)
	Adaptive	1.18(0.09)	1.20(0.06)	1.27(0.05)	4.40(0.13)	5.47(0.13)	7.24(0.17)
	SCAD	1.19(0.09)	1.19(0.07)	1.24(0.05)	4.58(0.13)	5.73(0.13)	7.55(0.15)

4.1 Simulation studies

Study 1.

We consider two dynamic covariance matrices in this study to investigate the accuracy of our proposed estimating approach in terms of the spectral loss and the Frobenius loss of a matrix.

Model 1: (Dynamic banded covariance matrices). Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1\le i,j\le p}$, where $\sigma_{ij}(u) = \exp(u/2)[\{\phi(u)+0.1\}I(|i-j|=1)+\phi(u)I(|i-j|=2)+I(i=j)]$ and $\phi(u)$ is the density of the standard normal distribution.

Model 2: (Dynamic AR(1) covariance model). Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1\le i,j\le p}$, where $\sigma_{ij}(u) = \exp(u/2)\phi(u)^{|i-j|}$.

Model 2 is not sparse, although many of the entries in $\Sigma(u)$ are close to zero for large p. This model is used to assess the accuracy of the sparse DCMs for approximating non-sparse matrices.

For each covariance model, we generate 50 datasets, each consisting of $n = 150$ observations. We sample $U_i,\ i = 1, \cdots, n$, independently from the uniform distribution with support $[-1, 1]$. The response variable is generated according to $Y_i \sim N(\mathbf{0}, \Sigma(U_i)),\ i = 1, \cdots, n$, for $p = 100, 150$, or 250, respectively. The k and N in the subset-y-variables cross-validation are set to be $[p/12]$ and p, respectively, with $[p/12]$ denoting the largest integer no greater than $p/12$, and the N_1 in cross validation for choosing $\lambda(u)$ is set to be 100. We use the

Table 3: Average (standard error) spectral loss, Frobenius loss, TPR and FPR when $U = -0.75$ for Model 3 in Study 2

Method		spectral loss			Frobenius loss		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		2.54(0.18)	3.30(0.23)	4.68(0.27)	8.28(0.34)	12.3(0.48)	20.1(0.62)
Dynamic	Hard	0.49(0.13)	0.51(0.15)	0.55(0.14)	1.78(0.25)	2.19(0.33)	2.78(0.32)
	Soft	0.48(0.08)	0.46(0.09)	0.48(0.08)	1.71(0.17)	2.06(0.18)	2.57(0.13)
	Adaptive	0.53(0.10)	0.53(0.10)	0.55(0.11)	1.67(0.27)	1.99(0.32)	2.42(0.30)
	SCAD	0.55(0.08)	0.56(0.08)	0.61(0.09)	1.93(0.22)	2.39(0.29)	3.08(0.32)
Modified	Hard	0.49(0.13)	0.51(0.15)	0.55(0.14)	1.78(0.25)	2.19(0.33)	2.78(0.32)
	Soft	0.48(0.08)	0.46(0.09)	0.48(0.08)	1.71(0.17)	2.06(0.18)	2.57(0.13)
	Adaptive	0.53(0.10)	0.53(0.10)	0.55(0.11)	1.67(0.27)	1.99(0.32)	2.42(0.30)
	SCAD	0.55(0.08)	0.56(0.08)	0.61(0.09)	1.93(0.22)	2.39(0.29)	3.08(0.32)
Static	Hard	1.11(0.32)	1.16(0.29)	1.11(0.36)	4.34(0.77)	5.14(0.71)	6.26(0.71)
	Soft	0.92(0.13)	0.95(0.12)	0.93(0.15)	3.11(0.41)	3.60(0.37)	4.28(0.43)
	Adaptive	1.06(0.15)	1.10(0.14)	1.11(0.19)	3.89(0.47)	4.62(0.44)	5.69(0.51)
	SCAD	1.05(0.15)	1.09(0.14)	1.10(0.19)	4.12(0.42)	5.02(0.38)	6.37(0.45)
		TPR			FPR		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		NA	NA	NA	NA	NA	NA
Dynamic	Hard	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.07(0.01)	0.04(0.01)	0.03(0.00)
	Adaptive	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.01(0.01)	0.01(0.00)	0.00(0.00)
	SCAD	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.03(0.01)	0.02(0.01)	0.01(0.00)
Static	Hard	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.01(0.01)	0.00(0.00)	0.00(0.00)
	Soft	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.08(0.01)	0.05(0.01)	0.03(0.01)
	Adaptive	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.03(0.01)	0.02(0.00)	0.01(0.00)
	SCAD	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.05(0.01)	0.03(0.01)	0.02(0.00)

spectral and Frobenius losses as the criteria to compare the estimators produced by various approaches. Specifically, for each dataset, we estimate the DCMs at the following 20 points $u_i \in \mathcal{A} = \{-0.95, -0.85, \cdots, -0.05, 0.05, 0.15, \cdots, 0.85, 0.95\}$. Then for each method, we calculate the medians of 20 spectral and Frobenius losses, defined as

$$\text{Median Spectral Loss} = \text{median}\{\bigtriangledown_S(u_i),\ i = 1, \cdots, 20\},$$

$$\text{Median Frobenius Loss} = \text{median}\{\bigtriangledown_F(u_i),\ i = 1, \cdots, 20\},$$

where $\bigtriangledown_S(u) = \max_{1 \le j \le p} |\lambda_j\{\hat{\Sigma}(u) - \Sigma(u)\}|$ and $\bigtriangledown_F(u) = \sqrt{\text{trace}[\{\hat{\Sigma}(u) - \Sigma(u)\}^2]}$ are spectral loss and Frobenius loss, respectively. For brevity, the two losses, Median Spectral Loss and Median Frobenius Loss, are referred to as MSL and MFL, respectively.

Table 1 and Table 2 summarize the results of the spectral (Frobenius) losses for various

Table 4: Average (standard error) spectral loss, Frobenius loss, TPR and FPR when $U = 0.25$ for Model 3 in Study 2

Method		spectral loss			Frobenius loss		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		3.03(0.26)	4.11(0.26)	6.01(0.35)	10.8(0.32)	15.9(0.39)	26.1(0.68)
Dynamic	Hard	1.24(0.07)	1.26(0.04)	1.29(0.04)	6.61(0.88)	8.45(0.74)	11.4(0.77)
	Soft	1.35(0.07)	1.39(0.05)	1.46(0.04)	6.63(0.36)	8.42(0.36)	11.2(0.41)
	Adaptive	1.25(0.08)	1.29(0.05)	1.34(0.04)	6.28(0.46)	8.00(0.44)	10.7(0.50)
	SCAD	1.19(0.08)	1.22(0.05)	1.27(0.05)	6.39(0.38)	8.08(0.35)	10.7(0.39)
Modified	Hard	1.24(0.07)	1.26(0.04)	1.29(0.04)	6.61(0.88)	8.45(0.74)	11.4(0.77)
	Soft	1.35(0.07)	1.39(0.05)	1.46(0.04)	6.63(0.36)	8.42(0.36)	11.2(0.41)
	Adaptive	1.25(0.08)	1.29(0.05)	1.34(0.04)	6.28(0.46)	8.00(0.44)	10.7(0.50)
	SCAD	1.19(0.08)	1.22(0.05)	1.27(0.05)	6.39(0.38)	8.08(0.35)	10.7(0.39)
Static	Hard	1.30(0.06)	1.31(0.04)	1.32(0.03)	7.45(0.70)	9.48(0.49)	12.5(0.46)
	Soft	1.41(0.06)	1.44(0.04)	1.49(0.03)	7.37(0.33)	9.29(0.33)	12.3(0.36)
	Adaptive	1.32(0.07)	1.34(0.04)	1.38(0.04)	7.11(0.40)	8.99(0.39)	11.8(0.41)
	SCAD	1.26(0.06)	1.27(0.04)	1.31(0.04)	7.10(0.33)	8.91(0.31)	11.7(0.32)
		TPR			FPR		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		NA	NA	NA	NA	NA	NA
Dynamic	Hard	0.58(0.14)	0.54(0.10)	0.49(0.08)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.95(0.03)	0.93(0.03)	0.91(0.03)	0.07(0.01)	0.04(0.01)	0.03(0.00)
	Adaptive	0.86(0.06)	0.83(0.06)	0.80(0.05)	0.02(0.01)	0.01(0.00)	0.01(0.00)
	SCAD	0.92(0.04)	0.90(0.04)	0.88(0.04)	0.04(0.01)	0.03(0.01)	0.02(0.00)
Static	Hard	0.46(0.13)	0.40(0.07)	0.36(0.05)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.89(0.05)	0.85(0.05)	0.82(0.05)	0.07(0.01)	0.04(0.01)	0.03(0.00)
	Adaptive	0.75(0.08)	0.70(0.08)	0.66(0.07)	0.02(0.01)	0.01(0.00)	0.01(0.00)
	SCAD	0.84(0.07)	0.80(0.07)	0.77(0.06)	0.04(0.01)	0.03(0.01)	0.02(0.00)

estimators of the dynamic covariance matrices in Model 1 and Model 2, respectively. Here, Sample represents the sample conditional covariance estimate in (2). Several conclusions can be drawn from Table 1 and Table 2. First, there is a drastic improvement in accuracy by using thresholded estimators over the kernel smoothed conditional covariance matrix in (2), and this improvement increases with dimension p. Second, as is expected, our proposed estimating method produces more accurate estimators than the static covariance estimation approach independent of the thresholding rule used. Third, the modified estimators perform similarly as the unmodified dynamic estimates. However, we observe that the unmodified estimate is not positive definite sometimes. For example, when $n = 100$ in Model 2, we observe that about 0.6% of the estimated covariance matrices are not positive definite.

Study 2.

In this study, we consider a dynamic covariance model whose sparsity pattern varies as a function of the covariate U. The purpose of this study is to assess the ability of our proposed method for recovering the varying sparsity, evaluated via the true positive rate (TPR) and the false positive rate (FPR), defined as

$$TPR(u) = \frac{\#\{(i,j) : s_{\lambda_n(u)}(\hat{\sigma}_{ij}(u)) \neq 0 \ and \ \sigma_{ij}(u) \neq 0\}}{\#\{(i,j) : \sigma_{ij}(u) \neq 0\}},$$

$$FPR(u) = \frac{\#\{(i,j) : s_{\lambda_n(u)}(\hat{\sigma}_{ij}(u)) \neq 0 \ and \ \sigma_{ij}(u) = 0\}}{\#\{(i,j) : \sigma_{ij}(u) = 0\}},$$

respectively (Rothman et al., 2009). For each given point u, we also evaluate the estimation accuracy of various approaches in terms of the spectral loss and the Frobenius loss.

Model 3: (Varying-sparsity covariance model) Let $\Sigma(u) = \{\sigma_{ij}(u)\}_{1\leq i,j\leq p}$, where

$$\begin{aligned}\sigma_{ij}(u) &= \exp(u/2)\Big[0.5\exp\{-\frac{(u-0.25)^2}{0.75^2-(u-0.25)^2}\}I(-0.5\leq u\leq 1)I(|i-j|=1)\\ &\quad +0.4\exp\{-\frac{(u-0.65)^2}{0.35^2-(u-0.65)^2}\}I(0.3\leq u\leq 1)I(|i-j|=2)+I(i=j)\Big].\end{aligned}$$

For this model, we assess the estimated covariance matrices at three points $\{-0.75, 0.25, 0.65\}$. Note that from the data generating process, the sparsity of this dynamic covariance model varies with the value of U, and that the covariance matrices at -0.75, 0.25 and 0.65 are diagonal, tridiagonal and five-diagonal respectively.

The data is generated following the procedure in Study 1. We report the spectral losses, Frobenius losses, TPRs and FPRs in Table 3, Table 4 and Table 5 at point -0.75, 0.25 and 0.65, respectively. In these tables, "NA" means "not applicable". Since the modified dynamic covariance estimator does not change the sparsity, we do not report the performance of this method for sparsity identification. The following conclusions can be drawn from the three tables. First, the accuracy statement in terms of the spectral loss and Frobenius loss made in Study 1 continues to hold in this study. Second, for each thresholding rule, our proposed dynamic covariance estimating method has generally higher true positive rates compared to the method for estimating static covariance matrices. Third, our proposed approach using the soft and the

Table 5: Average (standard error) spectral loss, Frobenius loss, TPR and FPR when $U = 0.65$ for Model 3 in Study 2

Method		spectral loss			Frobenius loss		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		3.79(0.37)	5.21(0.41)	7.74(0.53)	13.8(0.49)	20.4(0.75)	33.2(1.33)
Dynamic	Hard	2.25(0.08)	2.26(0.03)	2.29(0.03)	10.1(0.36)	12.5(0.20)	16.3(0.17)
	Soft	2.37(0.09)	2.43(0.06)	2.51(0.04)	10.0(0.36)	12.5(0.32)	16.6(0.33)
	Adaptive	2.27(0.09)	2.32(0.06)	2.38(0.04)	9.69(0.36)	12.1(0.32)	16.0(0.35)
	SCAD	2.17(0.09)	2.21(0.07)	2.29(0.05)	9.42(0.30)	11.7(0.28)	15.5(0.30)
Modified	Hard	2.25(0.08)	2.26(0.03)	2.29(0.03)	10.1(0.36)	12.5(0.19)	16.3(0.17)
	Soft	2.37(0.09)	2.43(0.06)	2.51(0.04)	10.0(0.36)	12.5(0.32)	16.6(0.33)
	Adaptive	2.27(0.09)	2.32(0.06)	2.38(0.04)	9.69(0.36)	12.1(0.32)	16.0(0.35)
	SCAD	2.17(0.09)	2.21(0.07)	2.29(0.05)	9.42(0.30)	11.7(0.28)	15.5(0.30)
Static	Hard	2.38(0.09)	2.40(0.04)	2.43(0.03)	10.4(0.44)	12.9(0.32)	16.9(0.29)
	Soft	2.46(0.07)	2.51(0.05)	2.56(0.04)	10.6(0.30)	13.3(0.29)	17.5(0.30)
	Adaptive	2.38(0.07)	2.41(0.05)	2.45(0.04)	10.2(0.32)	12.7(0.30)	16.7(0.32)
	SCAD	2.32(0.07)	2.34(0.04)	2.38(0.04)	10.1(0.28)	12.5(0.26)	16.3(0.27)
		TPR			FPR		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		NA	NA	NA	NA	NA	NA
Dynamic	Hard	0.28(0.07)	0.25(0.03)	0.23(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.68(0.07)	0.64(0.06)	0.59(0.04)	0.06(0.01)	0.04(0.01)	0.02(0.00)
	Adaptive	0.52(0.07)	0.48(0.06)	0.44(0.05)	0.01(0.00)	0.01(0.00)	0.00(0.00)
	SCAD	0.63(0.06)	0.60(0.05)	0.57(0.04)	0.03(0.01)	0.03(0.00)	0.02(0.00)
Static	Hard	0.28(0.08)	0.25(0.05)	0.22(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.65(0.06)	0.61(0.05)	0.57(0.04)	0.06(0.01)	0.04(0.01)	0.03(0.00)
	Adaptive	0.50(0.07)	0.46(0.06)	0.43(0.05)	0.01(0.00)	0.01(0.00)	0.00(0.00)
	SCAD	0.59(0.06)	0.55(0.06)	0.52(0.05)	0.03(0.01)	0.03(0.01)	0.02(0.00)

SCAD thresholding rules seems to have higher TPRs than using the hard and the adaptive thresholding rules.

Study 3.

In this study, we demonstrate the effectiveness of our proposed method for estimating a dynamic covariance model whose positions of the nonzero elements varied as a function of the covariate U.

Model 4: (Varying-nonzero-position covariance model) The dynamic covariance model is similar to the random graph model in Zhou et al. (2010). Specifically, we examine 9 time points $\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$. Let $R(u)$ be the correlation matrix at point u. We randomly choose p entries $\{r_{ij} : i = 2, \cdots, p; j < i\}$ of $R(-1)$ such that each of these p elements was a random variable generated from the uniform distribution with support $[0.1, 0.3]$. The other correlations in $R(-1)$ are all set to zero. For each of the other 8 time points $\{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$, we change $p/10$ existing nonzero correlations to zero and add $p/10$ new nonzero correlations. For each of the $p/10$ new entries having nonzero correlations, we choose a target correlation, and the correlation on the entry is gradually changed to ensure smoothness. Similarly, for each of the $p/10$ entries to be set as zero, the correlation decays to zero gradually. Thus, there exist $p + p/10$ nonzero correlations and there exist $p/5$ correlations that varied smoothly. The covariance matrix is then set as $\exp(u/2)R(u)$. We generate data following the procedure in Study 1 with $n = 100$ or $n = 150$. The results for estimating the covariance matrix at point $U = -1$ (Zhou et al., 2010) are reported in Table 6 for $n = 150$ and Table 7 for $n = 100$. We find that the proposed method performs better than the sample estimates and the static estimates in terms of the spectral loss, Frobenius loss and TPR.

Finally, we investigate the performance of the proposed bandwidth selection procedure using the model in this study. For a given bandwidth parameter h, our proposed estimator of $\Sigma(u)$ is denoted as $s_{\lambda_n(u)}(\hat{\Sigma}(u; h))$. The oracle that knows the true dynamic covariance matrix $\Sigma(u)$ prefers to select the bandwidth parameter h (i.e., h_{oracle}) that minimizes $\sum_{i=1}^{n} ||s_{\lambda_n(U_i)}(\hat{\Sigma}(U_i; h)) - \Sigma(U_i)||_F$, where $|| \cdot ||_F$ is the Frobenius loss. The bandwidth parameter selected by our proposed cross-validation procedure is denoted as h_{CV}. We use the absolute relative error $|h_{CV} -$

Table 6: Average (standard error) spectral loss, Frobenius loss, TPR and FPR when $U = -1$ for Model 4 in Study 3 with $n = 150$

Methods		spectral loss			Frobenius loss		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		2.65(0.23)	3.34(0.18)	4.69(0.23)	8.26(0.33)	12.1(0.34)	19.8(0.53)
Dynamic	Hard	0.68(0.09)	0.75(0.07)	0.76(0.07)	2.86(0.13)	3.84(0.16)	4.95(0.20)
	Soft	0.55(0.09)	0.59(0.06)	0.59(0.05)	2.17(0.10)	2.74(0.13)	3.44(0.10)
	Adaptive	0.63(0.09)	0.68(0.05)	0.69(0.06)	2.50(0.12)	3.34(0.15)	4.22(0.17)
	SCAD	0.66(0.09)	0.70(0.04)	0.73(0.05)	2.81(0.13)	3.77(0.16)	4.88(0.20)
Modified	Hard	0.68(0.09)	0.75(0.07)	0.76(0.07)	2.86(0.13)	3.84(0.16)	4.95(0.20)
	Soft	0.55(0.09)	0.59(0.06)	0.59(0.05)	2.17(0.10)	2.74(0.13)	3.44(0.10)
	Adaptive	0.63(0.09)	0.68(0.05)	0.69(0.06)	2.50(0.12)	3.34(0.15)	4.22(0.17)
	SCAD	0.66(0.09)	0.70(0.04)	0.73(0.05)	2.81(0.13)	3.77(0.16)	4.88(0.20)
Static	Hard	0.99(0.11)	1.02(0.11)	1.02(0.09)	4.87(0.27)	5.93(0.27)	7.65(0.34)
	Soft	0.79(0.09)	0.79(0.06)	0.80(0.07)	3.38(0.23)	3.95(0.21)	4.91(0.25)
	Adaptive	0.91(0.09)	0.90(0.07)	0.93(0.08)	4.32(0.25)	5.17(0.26)	6.58(0.31)
	SCAD	0.94(0.09)	0.94(0.06)	0.98(0.07)	4.83(0.27)	5.86(0.26)	7.60(0.35)
		TPR			FPR		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		NA	NA	NA	NA	NA	NA
Dynamic	Hard	0.34(0.01)	0.34(0.01)	0.34(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.60(0.03)	0.59(0.03)	0.54(0.02)	0.06(0.01)	0.04(0.01)	0.02(0.00)
	Adaptive	0.47(0.03)	0.46(0.03)	0.43(0.02)	0.01(0.00)	0.01(0.00)	0.00(0.00)
	SCAD	0.53(0.03)	0.52(0.03)	0.49(0.02)	0.02(0.01)	0.02(0.00)	0.01(0.00)
Static	Hard	0.34(0.01)	0.34(0.01)	0.34(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.56(0.03)	0.55(0.03)	0.50(0.02)	0.05(0.01)	0.03(0.01)	0.02(0.00)
	Adaptive	0.44(0.03)	0.43(0.02)	0.41(0.02)	0.01(0.00)	0.00(0.00)	0.00(0.00)
	SCAD	0.50(0.03)	0.49(0.03)	0.46(0.02)	0.02(0.01)	0.01(0.00)	0.01(0.00)

Table 7: Average (standard error) spectral loss, Frobenius loss, TPR and FPR when $U = -1$ for Model 4 in Study 3 with $n = 100$

Method		spectral loss			Frobenius loss		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		3.25(0.25)	4.12(0.22)	5.83(0.33)	9.82(0.39)	14.6(0.42)	23.8(0.86)
Dynamic	Hard	0.78(0.10)	0.83(0.07)	0.88(0.12)	3.26(0.22)	4.40(0.24)	5.68(0.36)
	Soft	0.62(0.08)	0.67(0.08)	0.69(0.10)	2.37(0.12)	2.96(0.13)	3.75(0.16)
	Adaptive	0.72(0.09)	0.78(0.08)	0.80(0.11)	2.77(0.19)	3.68(0.21)	4.65(0.28)
	SCAD	0.78(0.09)	0.85(0.08)	0.90(0.11)	3.26(0.22)	4.39(0.24)	5.63(0.35)
Modified	Hard	0.78(0.10)	0.83(0.07)	0.88(0.12)	3.26(0.22)	4.40(0.24)	5.68(0.36)
	Soft	0.62(0.08)	0.67(0.08)	0.69(0.10)	2.37(0.12)	2.96(0.13)	3.75(0.16)
	Adaptive	0.72(0.09)	0.78(0.08)	0.80(0.11)	2.77(0.19)	3.68(0.21)	4.65(0.28)
	SCAD	0.78(0.09)	0.85(0.08)	0.90(0.11)	3.26(0.22)	4.39(0.24)	5.63(0.35)
Static	Hard	1.03(0.15)	1.05(0.12)	1.08(0.15)	4.87(0.31)	5.96(0.36)	7.64(0.50)
	Soft	0.82(0.09)	0.83(0.09)	0.86(0.09)	3.13(0.23)	3.68(0.24)	4.55(0.27)
	Adaptive	0.94(0.12)	0.95(0.10)	1.00(0.10)	4.06(0.29)	4.87(0.33)	6.12(0.42)
	SCAD	1.03(0.11)	1.06(0.09)	1.12(0.09)	4.85(0.31)	5.90(0.35)	7.46(0.49)
		TPR			FPR		
		$p = 100$	$p = 150$	$p = 250$	$p = 100$	$p = 150$	$p = 250$
Sample		NA	NA	NA	NA	NA	NA
Dynamic	Hard	0.34(0.00)	0.33(0.00)	0.33(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.53(0.03)	0.50(0.02)	0.47(0.02)	0.05(0.01)	0.03(0.00)	0.02(0.00)
	Adaptive	0.41(0.02)	0.41(0.02)	0.39(0.01)	0.01(0.00)	0.00(0.00)	0.00(0.00)
	SCAD	0.50(0.03)	0.48(0.02)	0.46(0.02)	0.03(0.01)	0.02(0.00)	0.01(0.00)
Static	Hard	0.34(0.01)	0.34(0.00)	0.33(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
	Soft	0.49(0.03)	0.47(0.02)	0.44(0.02)	0.04(0.01)	0.02(0.00)	0.01(0.00)
	Adaptive	0.39(0.02)	0.38(0.02)	0.37(0.01)	0.01(0.00)	0.00(0.00)	0.00(0.00)
	SCAD	0.47(0.03)	0.45(0.02)	0.44(0.02)	0.03(0.01)	0.02(0.00)	0.01(0.00)

$h_{oracle}|/h_{oracle}$ as the criterion to measure the performance of h_{CV}. Setting $p = 100$, we explore the performance based on 50 simulations for $n = 100$, 150 and 200, respectively. The medians of 50 absolute relative errors for sample sizes 100, 150 and 200 are 0.12, 0.06 and 0.01, respectively. The percentages of absolute relative errors less than 20% for sample sizes 100, 150 and 200 are 0.82, 0.96, 0.98, respectively. It is concluded that the estimated bandwidth converges fast to its oracle counterpart when the sample size grows.

4.2 Real data analysis

As an illustration, we apply the dynamic covariance method to the resting state fMRI data obtained from the attention deficit hyperactivity disorder (ADHD) study conducted by New York University Child Study Center. ADHD is one of the most common childhood and adolescents

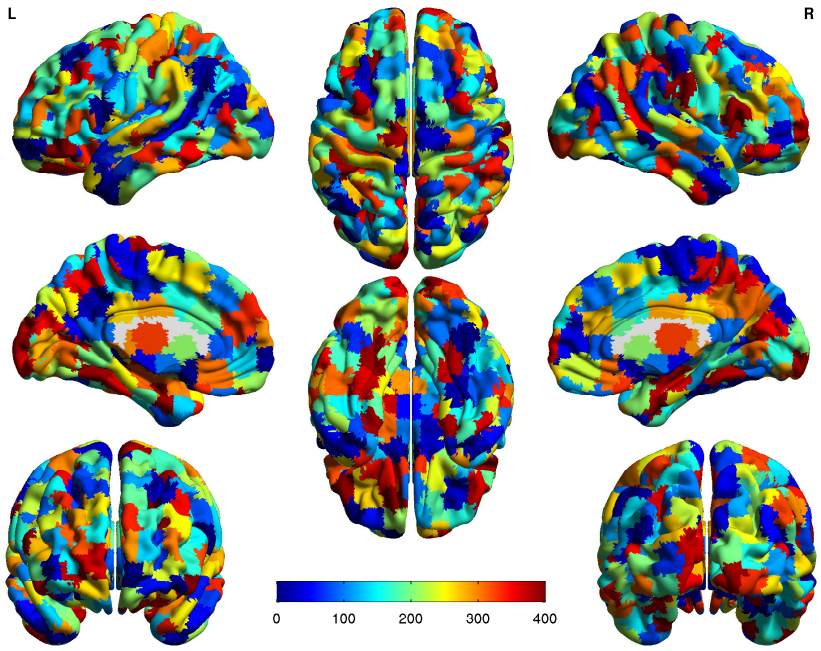


Figure 1: The ROIs from the CC400 functional parcellation atlases.

disorders and can continue through adulthood. Symptoms of ADHD include difficulty staying focused and paying attention, difficulty controlling behavior, and over-activity. An ADHD patient tends to have high variability in brain activities over time. Because fMRI measures brain activity by detecting associated changes in blood flow through low frequency BOLD signal in the brain (Biswal et al., 1995), it is believed that the temporally varying information in fMRI data may provide insight into the fundamental workings of brain networks (Calhoun et al., 2014; Lindquist et al., 2014). Thus, it is of great interest to study the dynamic changes of association among different regions of interest (ROIs) of the brain for an ADHD patient at the resting state. For this dataset, we examine the so-called CC400 ROI atlases with 351 ROIs derived by functionally parcellating the resting state data as discussed in Craddock et al. (2012). An illustration of these ROIs is found in Figure 1.

The experiment included 222 children and adolescents. We focus on Individual 0010001. The BOLD signals of $p = 351$ ROIs of the brain were recorded over $n = 172$ scans equally spaced in time. We treat the time as the index variables U after normalizing the 172 scanning time points onto $[0, 1]$. The main aim is to assess how the correlations of BOLD signals change with the scanning time, as changing correlations can illustrate the existence of distinctive temporal associations. Based on the time index variable and 351 ROIs, we apply the proposed dynamic

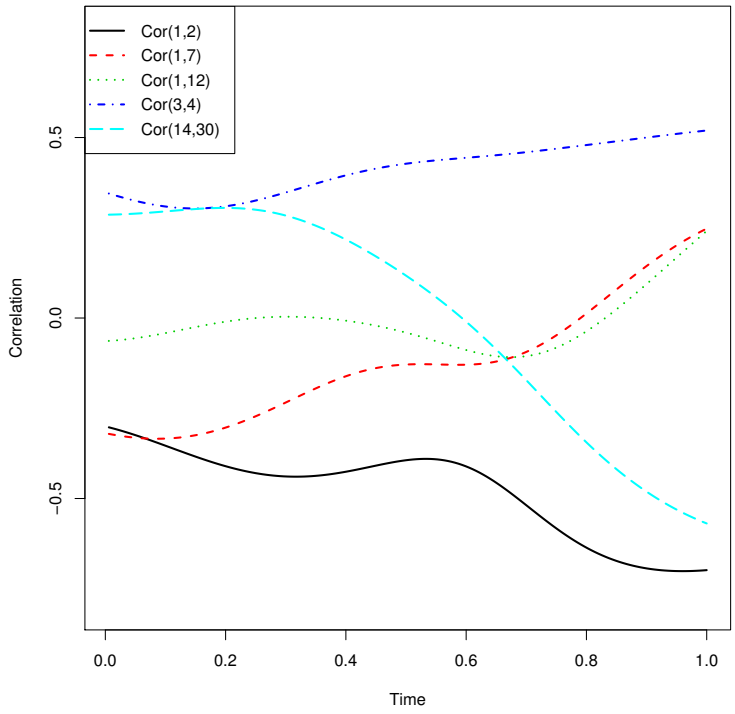


Figure 2: Selected entries of the correlation matrix as functions of time using (2) for the ADHD data. $\mathrm{Cor}(k, l)$ represents the (k, l)-th element of the correlation matrix.

covariance estimating method and compare it to the static method in Rothman et al. (2009).

We first obtain the sample covariance estimate as defined in (2), and plot selected entries of this matrix in Figure 2. We can see that the correlations of BOLD signals for different pairs of ROIs vary with time. For example, the entries $(1, 7)$ and $(14, 30)$ as functions of the time change signs, one from negative to positive and one from positive to negative. Entries $(1, 2)$ and $(3, 4)$ seem to remain negative and positive respectively over the entire time, while entry $(1, 12)$ is very close to zero before becoming positive. As discussed in the Introduction, a test of the equality of the two covariance matrices, one for the first 86 scans and the other for the last 86 scans, is rejected. These motivate the use of the proposed dynamic covariance method. For the dynamic sparse estimates, we only report the results using the soft thresholding rule, since simulation studies indicate that this thresholding rule performs satisfactorily in recovering the true sparsity of the covariance matrices. We examine the estimated dynamic covariance matrices at the 50-th, 90-th and 130-th scans.

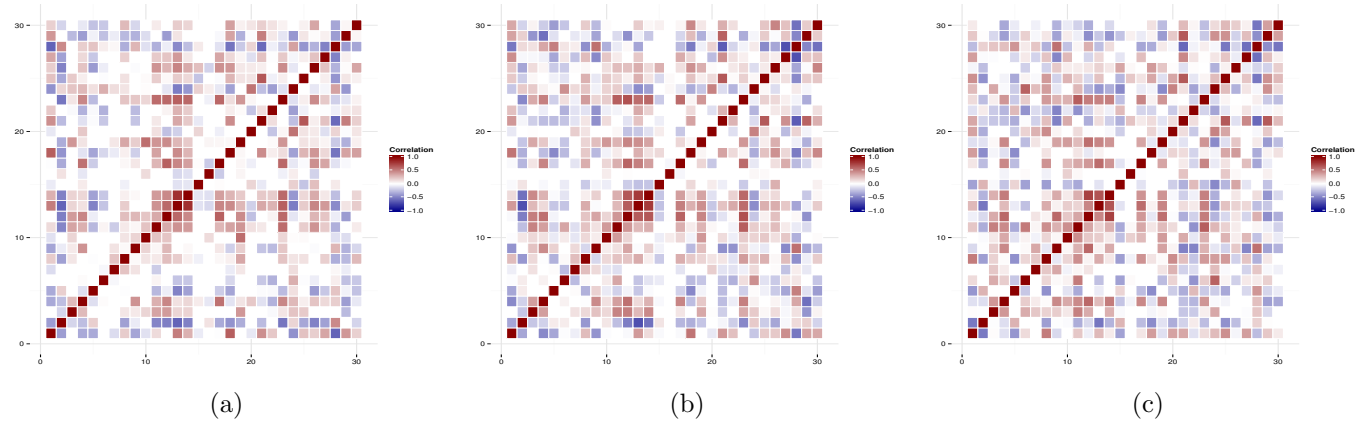
The heatmaps of the estimated dynamic correlation matrices of the first 30 ROIs are shown

Table 8: Qualitative summary of the estimated correlations at time 50 and 130

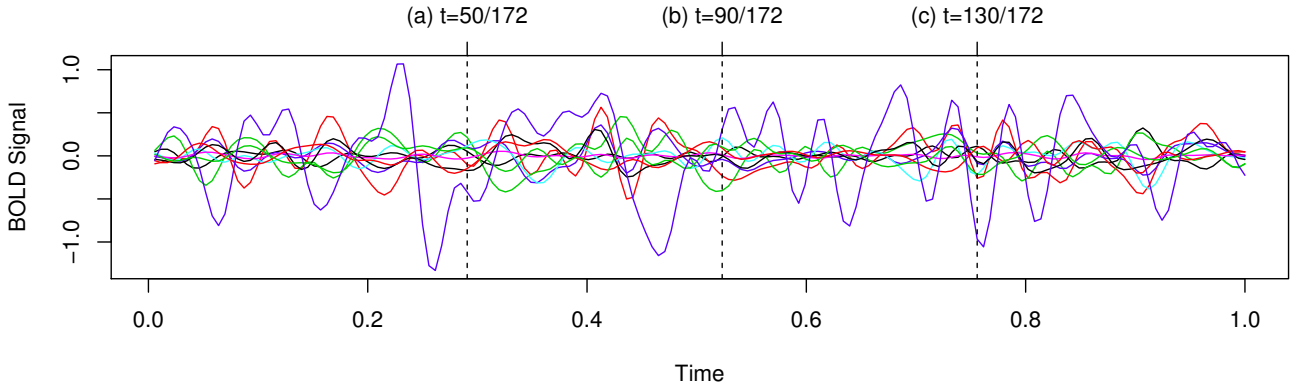
	Time=50			
Time=130	Positive	Negative	Nonzero	Zero
Positive	14739	4068	–	–
Negative	3934	13973	–	–
Nonzero	–	–	36714	11423
Zero	–	–	5901	7387

in Figure 3 (a), (b) and (c) for time 50, 90 and 130 respectively. We can see a clear varying pattern in panel (a)–(c), as compared to the static covariance matrix estimate in panel (e). To appreciate the dynamic characteristics of the BOLD signals, in panel (d), we use K-means clustering to cluster the 351 time series, one from each ROI, with $K = 10$, and plot these ten centroids. We can see different correlation patterns for the BOLD signals during different time periods, indicating the need for dynamically capturing the time varying phenomenon.

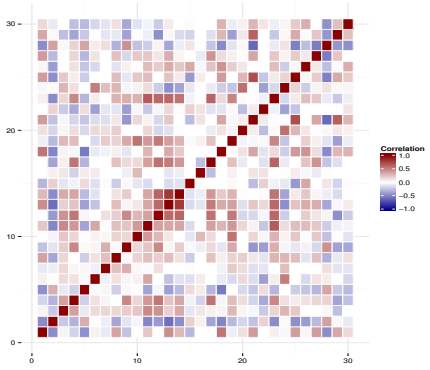
We comment on the dynamic nature of the three estimated matrices. The matrices at time 50, 90 and 130 have 42615, 46778 and 48137 non-zero correlations, respectively. That is, the covariance matrix at time = 130 is denser than those at time = 50 and 90. This tells that the sparsity of the covariance varies with time. Moreover, the positions of the nonzero (or zero) correlations change with time. For example, as Table 8 shows, there are 7387 and 36714 entries (correlations) of the estimated covariance matrices at time = 50 and 130 to be simultaneously zero and nonzero, respectively. However, there are 5901 entries (correlations) that are nonzero in the estimated covariance matrix at time = 50 but are zero in the estimated matrix at time = 130. There are 11423 entries (correlations) that are zero in the estimated matrix at time = 50 but are zero in the estimated matrix at time = 130. Furthermore, the signs of the correlations are time varying. As Table 8 shows, the numbers of the entries (correlations) of the two estimated covariance matrices being simultaneously positive and negative are 14739 and 13973, respectively. Meanwhile, there are 4068 entries (correlations) that are negative in the estimated covariance matrix at time = 50 but are positive in the estimated matrix at time = 130. And there are 3934 entries (correlations) that are positive in the estimated covariance matrix at time = 50 but are negative in the estimated matrix at time = 130. We found that, at time = 50, 2 ROIs have more than 325 associations with other ROIs, and that 8 ROIs have fewer than 10 associations with other ROIs. When time = 130, the number of the ROIs having more than 325 associations



(a) (b) (c)



(d)



(e)

Figure 3: Heatmaps of the estimated correlations of the ADHD data. (a), (b) and (c) are the estimated correlation matrices by our proposed dynamic covariance estimating method when the index random variable U is 50/172, 90/172 and 130/172, respectively; (d) shows the cluster centroids using K-means clustering with $K = 10$; (e) is the estimated correlation matrix by the static covariance estimating method.

becomes 11, and the number of the ROIs having fewer than 10 associations is 3, indicating that there are more active ROIs at this time. However, the static covariance estimating method can not show these dynamics.

As brain activities are often measured through low frequency BOLD signals in the brain, our model indicates that the correlations of different areas of the brain varied over time, which coincides with the high variability of brain function for an ADHD patient and makes sense for the purpose of locating the ADHD pathology.

5 Discussion

We study for the first time a novel uniform theory of the dynamic covariance model that combines the strength of kernel smoothing and modelling sparsity in high dimensional covariances. Our numerical results show that our proposed method can capture dynamic behaviors of these varying matrices and outperforms its competitors. We are currently studying similar uniform theory for high dimensional regression and classification where dynamics are incorporated.

We identify several directions for future study. First, the kernel smoothing employed in this paper uses local constant fitting. It is interesting to study local linear models that is known to reduce bias (Fan and Gijbels, 1996). A first step was taken by Chen and Leng (2015) when the dimensionality is fixed, but more research is warranted. Second, it is of great interest to develop more adaptive thresholding rules such as those in Cai and Liu (2011). A difficulty in extending our method in that direction is that the sample dynamic covariance matrix in (2) is biased entrywise, unlike the sample static covariance matrix in Cai and Liu (2011). Third, it is of great interest to study a new notion of rank-based estimation of a large dimensional covariance matrix in a dynamic setting (Liu et al., 2012; Xue and Zou, 2012) that is more robust to the distributional assumption of the variables. Fourth, it is possible to study dynamic estimation of the inverse of a covariance matrix that elucidates conditional independence structures among the variables (Yuan and Lin, 2007). These topics are beyond the scope of the current paper and will be pursued elsewhere.

Appendix

Lemma 4. *Under Conditions (a)–(d), suppose that (3) is satisfied and* $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ *for all* i. *If* $x > 0$ *and* $x \to 0$ *as* $n \to \infty$, *we have, for* $1 \le i \le n$, $1 \le j \le p$ *and* $1 \le k \le p$,

$$Ee^{xK(\frac{U_i-u}{h})Y_{ij}Y_{ik}} < \infty$$

for each $u \in \Omega$.

Proof. When n is large enough, we have

$$\begin{aligned} Ee^{xY_{ij}Y_{ik}} &= 1 + xEY_{ij}Y_{ik} + \frac{x^2}{2}EY_{ij}^2Y_{ik}^2 + \frac{x^3}{3!}EY_{ij}^3Y_{ik}^3 + \cdots \\ &\le 1 + xEY_{ij}^2 + \frac{x^2}{2}EY_{ij}^4 + \frac{x^3}{3!}EY_{ij}^6 + \cdots \\ &\quad +1 + xEY_{ik}^2 + \frac{x^2}{2}EY_{ik}^4 + \frac{x^3}{3!}EY_{ik}^6 + \cdots \\ &= Ee^{xY_{ij}^2} + Ee^{xY_{ik}^2} < \infty. \end{aligned}$$

By simple calculation, it is seen $Ee^{|xY_{ij}Y_{ik}|} < \infty$. Due to the boundedness of the kernel function $K(\cdot)$, we obtain $Ee^{|xK(\frac{U_i-u}{h})Y_{ij}Y_{ik}|} < \infty$. The result follows. □

Define $W_{ijk}(u,h) := K\{(U_i-u)/h\}Y_{ij}Y_{ik} - EK\{(U_i-u)/h\}Y_{ij}Y_{ik}$, for $1 \le i \le n$, $1 \le j \le p$ and $1 \le k \le p$. By Lemma 4, we easily have the following lemma.

Lemma 5. *Under Conditions (a)–(d), if (3) holds and* $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ *for all* i, *then, if* $x > 0$ *and* $x \to 0$ *as* $n \to \infty$, *there exists a positive constant* $g_i = O(var(W_{ijk}(u,h)) = O(h)$, *which does not depend on* u, *such that*

$$E\exp\{xW_{ijk}(u,h)\} \le \exp\{g_i x^2\}$$

for each $u \in \Omega$.

The exponential convergence rate for $|\frac{1}{n}\sum_{i=1}^n K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}|$ is important for deriving our theorems. First, we obtain its point-wise convergence rate in the following lemma.

Lemma 6. *Under Conditions (a)–(d), suppose that (3) holds and* $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ *for all* i. *If* $\lambda_n \to 0$, *there exists a constant* $C > 0$, *which does not depend on* u, *such that*

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big| \geq \lambda_n\Big) \leq 2\exp\{-Cnh\lambda_n^2\}$$

for each $u \in \Omega$.

Proof. Define $G := \sum_{i=1}^{n} g_i = O(nh)$. By Lemma 5, for $x > 0$ and $x \to 0$ as $n \to \infty$, we have

$$\begin{aligned} P\Big(\sum_{i=1}^{n} W_{ijk}(u,h) \geq \lambda_n\Big) &\leq \exp\{-x\lambda_n\}E\exp\{x\sum_{i=1}^{n} W_{ijk}(u,h)\} \\ &= \exp\{-x\lambda_n\}\Pi_i^n E\exp\{xW_{ijk}(u,h)\} \\ &\leq \exp\{-\lambda_n x + Gx^2\}. \end{aligned} \tag{6}$$

Note that (6) is maximized when $x = \lambda_n/2G \to 0$ and that the maximizer is $\exp\{-\frac{\lambda_n^2}{4G}\}$. Thus, there exists a positive constant C such that

$$P\Big(\Big\{\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big\} \geq \lambda_n\Big) \leq \exp\{-Cnh\lambda_n^2\}.$$

Similarly, we obtain

$$P\Big(\Big\{\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big\} \leq -\lambda_n\Big) \leq \exp\{-Cnh\lambda_n^2\}.$$

Combining the above two equations, the result is established. □

We discuss the uniform exponential convergence rate for $|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - E(Y_{ij}Y_{ik}|U=u)f(u)|$, which plays an important role in deriving our theorems.

Lemma 7. *Under Conditions (a)–(d), suppose that (3) holds,* $\sup_u |K'(u)| < M_5 < \infty$ *and* $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ *for all* i. *For sufficient large* M', *if* $\lambda_n = M'(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$, $\frac{\log p}{nh} \to 0$ *and* $h^4 \log p \to 0$, *there exist* $C_1 > 0$ *and* $C_2 > 0$ *such that*

$$P\Big(\sup_{u\in\Omega}\Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - E(Y_{ij}Y_{ik}|U=u)f(u)\Big| \geq \lambda_n\Big) \leq C_2 h^{-4}\exp\{-C_1 nh\lambda_n^2\}.$$

Proof. Without loss of generality, we let $\Omega = [a, b]$. Decompose $[a, b] = \cup_{l=1}^{q_n}[u_{n,l}-r_n, u_{n,l}+r_n]$, which contains q_n intervals of length $2r_n$. That is $2q_n r_n = b - a$. Then,

$$\begin{aligned}
&\sup_{u\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)Y_{ij}Y_{ik} - EK_h(U_i - u)Y_{ij}Y_{ik}\Big| \\
&\quad\le \max_{1\le l\le q_n} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h\{U_i - u_{n,l}\}Y_{ij}Y_{ik} - EK_h\{U_i - u_{n,l}\}Y_{ij}Y_{ik}\Big| \\
&\qquad + \max_{1\le l\le q_n} \sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h\{U_i - u_{n,l}\}Y_{ij}Y_{ik} \\
&\qquad\qquad - \{EK_h(U_i - u)Y_{ij}Y_{ik} - EK_h(U_i - u_{n,l})Y_{ij}Y_{ik}\}\Big|.
\end{aligned}$$

Let $r_n = h^4$, then $q_n = \frac{b-a}{2h^4}$. Let $\lambda_n = M'(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$, where M' is sufficiently large, $\frac{\log p}{nh} \to 0$ and $h^4 \log p \to 0$. Define $\mathcal{B}_1 = \{\omega : \max_{1\le l\le q_n} \sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} |\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_{n,l})Y_{ij}Y_{ik} - \{EK_h(U_i - u)Y_{ij}Y_{ik} - EK_h(U_i - u_{n,l})Y_{ij}Y_{ik}\}| < \lambda_n/3\}$ and $\mathcal{B}_2 = \mathcal{B}_1^C$. By Taylor's expansion, we have

$$\begin{aligned}
&\sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_{n,l})Y_{ij}Y_{ik} \\
&\qquad\qquad - \{EK_h(U_i - u)Y_{ij}Y_{ik} - EK_h(U_i - u_{n,l})Y_{ij}Y_{ik}\}\Big| \\
&\le h^4 \sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{nh^2}\sum_{i=1}^{n} \Big\{K'(\frac{U_i - u_{n,l} + R_i(u_{n,l} - u)}{h})Y_{ij}Y_{ik} \\
&\qquad\qquad - EK'(\frac{U_i - u_{n,l} + R_i(u_{n,l} - u)}{h})Y_{ij}Y_{ik}\Big\}\Big|, \qquad (7)
\end{aligned}$$

where $0 < R_i < 1$ is a random scalar depending on U_i, for $i = 1, \cdots, n$. Since $\sup_u |K'(u)| < M_5 < \infty$, (7) is bounded by

$$\frac{h^2 M_5}{n}\sum_{i=1}^{n}(Y_{ij}^2 + Y_{ik}^2) + h^2 M_5(EY_{ij}^2 + EY_{ik}^2). \qquad (8)$$

Since $\lambda_n \gg h^2$, by (7), (8) and similar arguments in Lemma 6, there exists a positive constant

A (for all l) such that

$$P\Big(\sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_{n,l})Y_{ij}Y_{ik} - \{EK_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u_{n,l})Y_{ij}Y_{ik}\}\Big| \geq \lambda_n/3\Big)$$
$$\leq 2\exp\{-\frac{An}{h^4}\lambda_n^2\}.$$

Thus, we obtain immediately

$$\begin{aligned} P(\mathcal{B}_2) &= P\Big(\max_{1\leq l\leq q_n} \sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_{n,l})Y_{ij}Y_{ik} \\ &\qquad - \{EK_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u_{n,l})Y_{ij}Y_{ik}\}\Big| \geq \lambda_n/3\Big) \\ &= q_n \max_{1\leq l\leq q_n} P\Big(\sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_{n,l})Y_{ij}Y_{ik} \\ &\qquad - \{EK_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u_{n,l})Y_{ij}Y_{ik}\}\Big| \geq \lambda_n/3\Big) \\ &\leq 2q_n\exp\{-\frac{An}{h^4}\lambda_n^2\}. \end{aligned} \tag{9}$$

Note

$$\begin{aligned} &P\Big(\sup_{u\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big| \geq \lambda_n\Big) \\ &\leq P\Big(\Big\{ \sup_{u\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big| \geq \lambda_n\Big\} \cap \mathcal{B}_1\Big) \\ &\quad + P\Big(\Big\{ \sup_{u\in[a,b]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u)Y_{ij}Y_{ik}\Big| \geq \lambda_n\Big\} \cap \mathcal{B}_2\Big) \\ &:= J_1 + J_2. \end{aligned}$$

We know that J_1 is bounded (using Lemma 6) as

$$\begin{aligned} &P\Big(\max_{1\leq l\leq q_n} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h\{U_i-u_{n,l}\}Y_{ij}Y_{ik} - EK_h\{U_i-u_{n,l}\}Y_{ij}Y_{ik}\Big| \geq (\lambda_n - \lambda_n/3)\Big) \\ &\leq q_n \max_{1\leq l\leq q_n} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_h\{U_i-u_{n,l}\}Y_{ij}Y_{ik} - EK_h\{U_i-u_{n,l}\}Y_{ij}Y_{ik}\Big| \geq \lambda_n/2\Big) \\ &\leq 2q_n\exp\{-Cnh\lambda_n^2/4\}, \end{aligned}$$

and that J_2 is bounded by (9). Therefore, for large enough n,

$$P\Big(\sup_{u\in[a,b]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-EK_h(U_i-u)Y_{ij}Y_{ik}\Big|\geq\lambda_n\Big)$$
$$\leq 4q_n\exp\{-Cnh\lambda_n^2/4\}\leq\frac{2(b-a)}{h^4}\exp\{-Cnh\lambda_n^2/4\}.$$

It is well known that (Pagan and Ullah, 1999; Fan and Huang, 2005)

$$\sup_{u\in[a,b]}\Big|EK_h(U_i-u)Y_{ij}Y_{ik}-E(Y_{ij}Y_{ik}|U=u)f(u)\Big|=O(h^2).$$

Since

$$\sup_{u\in[a,b]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-E(Y_{ij}Y_{ik}|U=u)f(u)\Big|$$
$$\leq\sup_{u\in[a,b]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-EK_h(U_i-u)Y_{ij}Y_{ik}\Big|$$
$$+\sup_{u\in[a,b]}\Big|EK_h(U_i-u)Y_{ij}Y_{ik}-E(Y_{ij}Y_{ik}|U=u)f(u)\Big|$$

and $\lambda_n\gg h^2$, we have immediately

$$P\Big(\sup_{u\in[a,b]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-E(Y_{ij}Y_{ik}|U=u)f(u)\Big|\geq\lambda_n\Big)$$
$$\leq P\Big(\sup_{u\in[a,b]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-EK_h(U_i-u)Y_{ij}Y_{ik}\Big|\geq\lambda_n/2\Big)$$
$$\leq\frac{2(b-a)}{h^4}\exp\{-Cnh\lambda_n^2/16\}.$$

Letting $C_1=C/16$ and $C_2=2(b-a)$, we obtain the conclusion. □

Remark 3. We now outline the main steps of our proofs. Along the way, we highlight the main theoretical innovations and comment that existing results in the literature may not apply to the high-dimensional problems we are interested in studying. In the first step, we obtain the exponential convergence rate of $|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik}-EK_h(U_i-u)Y_{ij}Y_{ik}|$ in Lemma 6 for each u on a discrete grid. After decomposing Ω, we transform the uniform exponential convergence rate problem in Lemma 7 to the discrete-point exponential convergence rate problem in Lemma

6. The exponential uniform convergence result of the kernel estimators in Lemma 7, essential for establishing the main conclusions of the paper, plays an important role in deriving the rates of convergence in Theorem 1 and Proposition 3 as well as the result in Theorem 2, when one deals with problems with the dimensionality exponentially high in relation to the sample size. In contrast, Einmahl and Mason (2005) used the exponential inequality of Talagrand (1994) to construct the uniform exponential convergence rate. According to their arguments, the right hand side of (7) should be much larger than $h^{-4}\exp\{-C_1 nh\lambda_n^2\}$. This indicates that the result implied by the theorems in Einmahl and Mason (2005) is not enough for deriving the uniform convergence results in our paper and thus does not apply to the challenging high-dimensional setup.

Remark 4. To facilitate the proof of Lemma 7, we assume $\sup_u |K^{'}(u)| < M_5 < \infty$. However, this does not mean that some commonly used kernel functions (e.g., the boxcar kernel) can not be used. We now examine the boxcar kernel specifically. Write $K(u) = \frac{1}{2}I(|u| \leq 1)$ and note

$$
\begin{aligned}
&\sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_{n,l})Y_{ij}Y_{ik} \\
&\qquad\qquad - \{EK_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u_{n,l})Y_{ij}Y_{ik}\}\Big| \\
\leq\ & \frac{1}{2}\sup_{u\in[u_{n,l}-r_n,u_{n,l}]} \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I\{U_i\in[u+h,u_{n,l}+h]\}Y_{ij}Y_{ik} - E\frac{1}{h} I\{U_i\in[u+h,u_{n,l}+h]\}Y_{ij}Y_{ik}\Big| \\
&+\frac{1}{2}\sup_{u\in[u_{n,l}-r_n,u_{n,l}]} \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I\{U_i\in[u-h,u_{n,l}-h]\}Y_{ij}Y_{ik} - E\frac{1}{h} I\{U_i\in[u-h,u_{n,l}-h]\}Y_{ij}Y_{ik}\Big| \\
&+\frac{1}{2}\sup_{u\in[u_{n,l},u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I\{U_i\in[u_{n,l}+h,u+h]\}Y_{ij}Y_{ik} - E\frac{1}{h} I\{U_i\in[u_{n,l}+h,u+h]\}Y_{ij}Y_{ik}\Big| \\
&+\frac{1}{2}\sup_{u\in[u_{n,l},u_{n,l}+r_n]} \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I\{U_i\in[u_{n,l}-h,u-h]\}Y_{ij}Y_{ik} - E\frac{1}{h} I\{U_i\in[u_{n,l}-h,u-h]\}Y_{ij}Y_{ik}\Big| \\
:=\ & A_1+A_2+A_3+A_4. \qquad (10)
\end{aligned}
$$

We have

$$
\begin{aligned}
A_1 \ \leq\ & \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I\{U_i\in[u_{n,l}+h-r_n,u_{n,l}+h]\}(Y_{ij}^2+Y_{ik}^2)\Big| \\
&+\Big|E\{\frac{1}{h} I\{U_i\in[u_{n,l}+h-r_n,u_{n,l}+h]\}(Y_{ij}^2+Y_{ik}^2)\Big| := B_1. \qquad (11)
\end{aligned}
$$

There exists $M > 0$ such that

$$\begin{aligned}\mathrm{var}\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}I\{U_i \in [u_{n,l}+h-r_n, u_{n,l}+h]\}Y_{ij}^2\Big\} &\leq \frac{1}{nh^2}E\Big\{I\{U_i \in [u_{n,l}+h-r_n, u_{n,l}+h]\}Y_{ij}^4\Big\} \\ &\leq \frac{h^4}{nh^2}\{\sup_u f(u)\}\{\sup_u E(Y_{ij}^4|U_i=u)\} \\ &\leq \frac{Mh^2}{n}.\end{aligned}$$

Thus, by the similar arguments in Lemma 6 and Lemma 7,

$$P\Big(B_1 \geq \frac{\lambda_n}{12}\Big) \leq 2\exp\{-\frac{Hn}{h^2}\lambda_n^2\}. \tag{12}$$

By (10), (11) and (12), we have

$$\begin{aligned}P\Big(\sup_{u\in[u_{n,l}-r_n,u_{n,l}+r_n]}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik} - \frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_{n,l})Y_{ij}Y_{ik} \\ - \{EK_h(U_i-u)Y_{ij}Y_{ik} - EK_h(U_i-u_{n,l})Y_{ij}Y_{ik}\}\Big| \geq \lambda_n/3\Big) \\ \leq 8\exp\{-\frac{Hn}{h^2}\lambda_n^2\}.\end{aligned}$$

Taking the similar procedure below (9), we prove the result of Lemma 7. We conclude that Lemma 7 as well as the results in Theorem 1 and 2 and Proposition 3 hold for the boxcar kernel.

Now, we have the following lemma.

Lemma 8. *Under Conditions (a)–(d), suppose that (3) holds and* $\sup_{u\in\Omega}\sigma_{ii}(u) < M_2 < \infty$ *for all* i. *For sufficient large* M', *if* $\lambda_n = M'(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$, $\frac{\log p}{nh} \to 0$ *and* $h^4\log p \to 0$, *there exist* $C_1 > 0$ *and* $C_2 > 0$ *such that*

$$P\Big(\max_{j,k}\sup_{u\in\Omega}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik} - E(Y_{ij}Y_{ik}|U=u)f(u)\Big| \geq \lambda_n\Big) \leq C_2p^2h^{-4}\exp\{-C_1nh\lambda_n^2\}.$$

Equivalently,

$$\max_{j,k}\sup_{u\in\Omega}\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u)Y_{ij}Y_{ik} - E(Y_{ij}Y_{ik}|U=u)f(u)\Big| = O_p(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p}).$$

Since $\frac{1}{n}\sum_{i=1}^n K_h(U_i-u)$ converges to $f(u)$ with convergent rate $\sqrt{\frac{\log n}{nh}}+h^2$ uniformly in u (Silverman, 1978; Pagan and Ullah, 1999) and $f(\cdot)$ is bounded away from 0, the following is true.

Lemma 9. *Under the conditions in Lemma 8, we have*

$$\max_{j,k}\sup_{u\in\Omega}\Big|\frac{\sum_{i=1}^n K_h(U_i-u)Y_{ij}Y_{ik}}{\sum_{i=1}^n K_h(U_i-u)}-E(Y_{ij}Y_{ik}|U=u)\Big|=O_p(\sqrt{\frac{\log p}{nh}}+h^2\sqrt{\log p}).$$

Following the similar arguments for deriving Lemma 8, we have the following lemma.

Lemma 10. *Under the conditions in Lemma 8, there exist $C_3>0$ and $C_4>0$ such that*

$$P\Big(\max_j\sup_{u\in\Omega}\Big|\frac{1}{n}\sum_{i=1}^n K_h(U_i-u)Y_{ij}-E(Y_{1j}|U=u)f(u)\Big|\geq\lambda_n\Big)\leq C_4ph^{-4}\exp\{-C_3nh\lambda_n^2\}.$$

Since $\sup_u|\frac{1}{n}\sum_{i=1}^n K_h(U_i-u)-f(u)|=O_p(\sqrt{\frac{\log n}{nh}}+h^2)$ and $f(\cdot)$ is bounded away from 0, by Lemma 10, the following result is immediate.

Lemma 11. *Under the conditions in Lemma 8, we have*

$$\max_j\sup_{u\in\Omega}\Big|\frac{\sum_{i=1}^n K_h(U_i-u)Y_{ij}}{\sum_{i=1}^n K_h(U_i-u)}-m_j(u)\Big|=O_p(\sqrt{\frac{\log p}{nh}}+h^2\sqrt{\log p}).$$

Proof of Theorem 1. By Lemma 9 and Lemma 11, we know

$$\max_{j,k}\sup_{u\in\Omega}|\hat{\sigma}_{jk}(u)-\sigma_{jk}(u)|=O_p(\sqrt{\frac{\log p}{nh}}+h^2\sqrt{\log p}). \tag{13}$$

Let $M_1:=\inf_{u\in\Omega}M(u)$ and $M_2:=\sup_{u\in\Omega}M(u)<\infty$. Then $\lambda_{1n}:=M_1(\sqrt{\frac{\log p}{nh}}+h^2\sqrt{\log p})\leq\lambda_n(u)\leq M_2(\sqrt{\frac{\log p}{nh}}++h^2\sqrt{\log p}):=\lambda_{2n}$, where M_1 is sufficiently large. We have, for each u,

$$||s_{\lambda_n(u)}(\hat{\Sigma}(u))-\Sigma(u)||\leq||s_{\lambda_n(u)}(\hat{\Sigma}(u))-s_{\lambda_n(u)}(\Sigma(u))||+||s_{\lambda_n(u)}(\Sigma(u))-\Sigma(u)||. \tag{14}$$

The second term of (14) is bounded by

$$
\begin{aligned}
&\max_j \sum_{k=1}^p |s_{\lambda_n(u)}(\sigma_{jk}(u)) - \sigma_{jk}(u)| \\
&\le \max_j \sum_{k=1}^p \lambda_n(u) I\{|\sigma_{jk}(u)| > \lambda_n(u)\} + \max_j \sum_{k=1}^p |\sigma_{jk}(u)| I\{|\sigma_{jk}(u)| \le \lambda_n(u)\} \\
&\le \max_j \sum_{k=1}^p \lambda_n(u)^q \lambda_n(u)^{1-q} I\{|\sigma_{jk}(u)| > \lambda_n(u)\} + \max_j \sum_{k=1}^p |\sigma_{jk}(u)| \frac{\lambda_n(u)^{1-q}}{|\sigma_{jk}(u)|^{1-q}} \\
&\le 2 \max_j (\sum_{k=1}^p |\sigma_{jk}(u)|^q) \lambda_n(u)^{1-q} \le 2 \max_j \sup_u (\sum_{k=1}^p |\sigma_{jk}(u)|^q) \lambda_n(u)^{1-q} \\
&\le 2c_0(p) \lambda_n(u)^{1-q} \le 2c_0(p) \lambda_{2n}^{1-q}. \qquad (15)
\end{aligned}
$$

On the other hand, the first term in (14) satisfies

$$
\begin{aligned}
||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - s_{\lambda_n(u)}(\Sigma(u))|| &\le \max_j \sum_{k=1}^p |s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) - s_{\lambda_n(u)}(\sigma_{jk}(u))| \\
&\le \max_j \sum_{k=1}^p |\hat{\sigma}_{jk}(u)| I\{|\hat{\sigma}_{jk}(u)| \ge \lambda_n(u), |\sigma_{jk}(u)| < \lambda_n(u)\} \\
&\quad + \max_j \sum_{k=1}^p |\sigma_{jk}(u)| I\{|\hat{\sigma}_{jk}(u)| < \lambda_n(u), |\sigma_{jk}(u)| \ge \lambda_n(u)\} \\
&\quad + \max_j \sum_{k=1}^p \{|\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| + |s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) - \hat{\sigma}_{jk}(u)| \\
&\quad + |s_{\lambda_n(u)}(\sigma_{jk}(u)) - \sigma_{jk}(u)|\} I\{|\hat{\sigma}_{jk}(u)| \ge \lambda_n(u), |\sigma_{jk}(u)| \ge \lambda_n(u)\} \\
&:= I_1 + I_2 + I_3. \qquad (16)
\end{aligned}
$$

By using (13), we have that the first term of I_3 is bounded by

$$
\begin{aligned}
&\max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \max_j \sup_u (\sum_{k=1}^p |\sigma_{jk}(u)|^q) \lambda_n(u)^{-q} \\
&= O_p\Big(c_0(p) (\sqrt{\frac{\log p}{nh}} + h^2 \sqrt{\log p})^{1-q}\Big).
\end{aligned}
$$

For the second term of I_3, we see

$$\begin{aligned}
&\max_j \sum_{k=1}^{p} |s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) - \hat{\sigma}_{jk}(u)| I\{|\hat{\sigma}_{jk}(u)| \geq \lambda_n(u), |\sigma_{jk}(u)| \geq \lambda_n(u)\} \\
&\leq \max_j \sum_{k=1}^{p} \lambda_n(u)^q \lambda_n(u)^{1-q} I\{|\hat{\sigma}_{jk}(u)| \geq \lambda_n(u), |\sigma_{jk}(u)| \geq \lambda_n(u)\} \\
&\leq \lambda_n(u)^{1-q} \max_j \sum_{k=1}^{p} |\sigma_{jk}(u)|^q I\{|\sigma_{jk}(u)| \geq \lambda_n(u)\} \\
&\leq \lambda_n(u)^{1-q} \max_j \sup_u (\sum_{k=1}^{p} |\sigma_{jk}(u)|^q) \leq c_0(p) \lambda_n(u)^{1-q} \leq c_0(p) \lambda_{2n}^{1-q}.
\end{aligned}$$

By (15), the third term of I_3 is less than $2c_0(p)\lambda_{2n}^{1-q}$. We have proven up to now

$$I_3 = O_p(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}). \tag{17}$$

The following can be derived

$$\begin{aligned}
I_2 &\leq \max_j \sum_{k=1}^{p} [|\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| + |\hat{\sigma}_{jk}(u)|] I\{|\hat{\sigma}_{jk}(u)| < \lambda_n(u), |\sigma_{jk}(u)| \geq \lambda_n(u)\} \\
&\leq \max_{j,k} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \max_j \sum_{k=1}^{p} I\{|\sigma_{jk}(u)| \geq \lambda_n(u)\} + \lambda_n(u) \max_j \sum_{k=1}^{p} I\{|\sigma_{jk}(u)| \geq \lambda_n(u)\} \\
&\leq \max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \frac{\max_j \sup_u (\sum_{k=1}^{p} |\sigma_{jk}(u)|^q)}{\lambda_n(u)^q} + \lambda_n(u)^{1-q} \max_j \sup_u (\sum_{k=1}^{p} |\sigma_{jk}(u)|^q) \\
&= O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big).
\end{aligned}$$

Thus, the following is satisfied

$$I_2 = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big). \tag{18}$$

Note

$$\begin{aligned}
I_1 &\leq \max_j \sum_{k=1}^{p} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| I(|\hat{\sigma}_{jk}(u)| \geq \lambda_n(u), |\sigma_{jk}(u)| < \lambda_n(u)) \\
&\quad + \max_j \sum_{k=1}^{p} |\sigma_{jk}(u)| I(|\sigma_{jk}(u)| < \lambda_n(u)) := I_4 + I_5.
\end{aligned}$$

By (15), we obtain $I_5 \le c_0(p)\lambda_n(u)^{1-q} \le c_0(p)\lambda_{2n}^{1-q}$. Taking $\alpha \in (0,1)$, then

$$\begin{aligned}
I_4 &= \max_j \sum_{k=1}^{p} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| I(|\hat{\sigma}_{jk}(u)| \ge \lambda_n(u), |\sigma_{jk}(u)| < \alpha\lambda_n(u)) \\
&\quad + \max_j \sum_{k=1}^{p} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| I(|\hat{\sigma}_{jk}(u)| \ge \lambda_n(u), \alpha\lambda_n(u) \le |\sigma_{jk}(u)| < \lambda_n(u)) \\
&\le \max_{j,k} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \max_j \sum_{k=1}^{p} I(|\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \ge (1-\alpha)\lambda_n(u)) \\
&\quad + \max_{j,k} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \max_j \sum_{k=1}^{p} \frac{|\sigma_{jk}(u)|^q}{(\alpha\lambda_n(u))^q} \\
&\le \max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \max_j \sum_{k=1}^{p} I(\sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \ge (1-\alpha)\lambda_n(u)) \\
&\quad + \max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| (\alpha\lambda_n(u))^{-q} \max_j \sup_u (\sum_{k=1}^{p} |\sigma_{jk}(u)|^q) \\
&:= I_6 + I_7.
\end{aligned}$$

Using (13), we have that $I_7 = O_p\Big(c_0(p)\lambda_{1n}^{-q}(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})\Big)$. By Lemma 8 and Lemma 10, there exist $C_1^* > 0$ and $C_2^* > 0$ such that

$$\begin{aligned}
&P\Big(\max_j \sum_{k=1}^{p} I\{\sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \ge (1-\alpha)\lambda_n(u)\} > 0\Big) \\
&= P\Big(\max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \ge (1-\alpha)\lambda_n(u)\Big) \\
&\le \frac{C_2^* p^2}{h^4} \exp\{-nhC_1^*(1-\alpha)^2\lambda_n(u)^2\} \\
&\le \frac{C_2^* p^2}{h^4} \exp\{-nhC_1^*(1-\alpha)^2\lambda_{1n}^2\} \to 0.
\end{aligned}$$

Thus, we obtain $I_6 = o_p(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})$ and then

$$I_1 = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big). \tag{19}$$

Combining (14)–(19), we prove the result. □

Proof of Theorem 2. Note for each $u \in \Omega$,

$$\begin{aligned}\{(j,k) : s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \neq 0, \sigma_{jk}(u) = 0\} &= \{(j,k) : |\hat{\sigma}_{jk}(u)| > \lambda_n(u), \sigma_{jk}(u) = 0\} \\ &\subseteq \{(j,k) : |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| > \lambda_n(u)\}.\end{aligned}$$

Therefore,

$$\begin{aligned}P\Big(\sum_{j,k} I(s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \neq 0, \sigma_{jk}(u) = 0) > 0\Big) &\leq P(\max_{j,k} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| > \lambda_n(u)) \\ &\leq P(\max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| > \lambda_n(u)).\end{aligned}$$

By Lemma 8 and Lemma 10, there exist $C_1^* > 0$ and $C_2^* > 0$ such that the right-hand side of the above equation is dominated by $C_2^* p^2 h^{-4} \exp\{-C_1^* nh\lambda_{1n}^2\}$. Since M_1 can be chosen to be large enough, we have

$$\sup_u P\Big(\sum_{j,k} I(s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \neq 0, \sigma_{jk}(u) = 0) > 0\Big) \to 0.$$

Thus, the first result is shown.

For proving the second result of this theorem, we note, for each $u \in \Omega$,

$$\begin{aligned}&\Big\{(j,k) : s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \leq 0, \sigma_{jk}(u) > 0 \text{ or } s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \geq 0, \sigma_{jk}(u) < 0\Big\} \\ &\subseteq \Big\{(j,k) : |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| > \tau_n(u) - \lambda_n(u)\Big\}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}P\Big(\cup_{j,k} \{|\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \geq \tau_n(u) - \lambda_n(u)\}\Big) &= P\Big(\max_{j,k} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \geq \tau_n(u) - \lambda_n(u)\Big) \\ &\leq P\Big(\max_{j,k} \sup_u |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| \geq \tau_n(u) - \lambda_n(u)\Big).\end{aligned}$$

By Lemma 8 and Lemma 10 again, the right-hand side of the above equation is dominated by $C_2^* p^2 h^{-4} \exp\{-C_1^* nh \inf_{u\in\Omega}(\tau_n(u) - \lambda_n(u))^2\}$. Since $\frac{\log p}{nh \inf_{u\in\Omega}(\tau_n(u)-\lambda_n(u))^2} \to 0$, we have

$$\sup_u P\Big(\cup_{j,k} \{s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \leq 0, \sigma_{jk}(u) > 0 \text{ or } s_{\lambda_n(u)}(\hat{\sigma}_{jk}(u)) \geq 0, \sigma_{jk}(u) < 0\}\Big) \to 0.$$

The second result of this theorem follows immediately. □

Proof of (5). Denote $\delta_n = \frac{n^{1/5}}{\sqrt{\log n}}$. Let

$$T_{ij} = Y_{ij} I\{|Y_{ij}| \le \delta_n\} \quad \text{and} \quad T_{ik} = Y_{ik} I\{|Y_{ik}| \le \delta_n\},$$

and

$$R_{ijk} = Y_{ij}Y_{ik} I\{|Y_{ij}| > \delta_n \text{ or } |Y_{ik}| > \delta_n\}.$$

Then, $Y_{ij}Y_{ik} = T_{ij}T_{ik} + R_{ijk}$.

Let $\epsilon_n = c_3\sqrt{\frac{\log p}{n^{4/5}}}$ with a large enough c_3. Using the notations in the proof of Lemma 7, we have

$$\begin{aligned}
& P\Big(\max_{j,k}\max_{0\le l\le q_n} |\sum_{i=1}^n (K(\frac{U_i-u_{n,l}}{h})Y_{ij}Y_{ik} - EK(\frac{U_i-u_{n,l}}{h})Y_{ij}Y_{ik})| \ge n^{4/5}\epsilon_n\Big) \\
\le\ & P\Big(\max_{j,k}\max_{0\le l\le q_n} |\sum_{i=1}^n (K(\frac{U_i-u_{n,l}}{h})T_{ij}T_{ik} - EK(\frac{U_i-u_{n,l}}{h})T_{ij}T_{ik})| \ge \frac{1}{2}n^{4/5}\epsilon_n\Big) \\
& + P\Big(\max_{j,k}\max_{0\le l\le q_n} |\sum_{i=1}^n (K(\frac{U_i-u_{n,l}}{h})R_{ijk} - EK(\frac{U_i-u_{n,l}}{h})R_{ijk})| \ge \frac{1}{2}n^{4/5}\epsilon_n\Big) \\
:=\ & H_1 + H_2. \qquad (20)
\end{aligned}$$

By Bernstein's inequality (Bennett, 1962), there exist constants $K_3 > 0$ and $K_4 > 0$ such that

$$\begin{aligned}
H_1 &\le (b-a)p^2 n^{4/5} \exp\{-\frac{c_3^2 n^{2/5}\log p}{K_3 n^{2/5} + c_3 K_4 \frac{n^{2/5}}{\log n}\sqrt{\log p}}\} \\
&= (b-a)p^2 n^{4/5} \exp\{-\frac{c_3^2 \log p}{K_3 + c_3 K_4 \frac{\sqrt{\log p}}{\log n}}\}. \qquad (21)
\end{aligned}$$

By Hölder's inequality and (4), we have

$$\begin{aligned}
& |EK(\frac{U_i-u_{n,l}}{h})R_{ijk}| \\
\le\ & |EK(\frac{U_i-u_{n,l}}{h})Y_{ij}Y_{ik}I\{|Y_{ij}| > \delta_n\}| + |EK(\frac{U_i-u_{n,l}}{h})Y_{ij}Y_{ik}I\{|Y_{ik}| > \delta_n\}| \\
\le\ & EK(\frac{U_i-u_{n,l}}{h})\frac{|Y_{ij}|^{3+\gamma}}{\delta_n^{2+\gamma}}|Y_{ik}|I\{|Y_{ij}| > \delta_n\} + EK(\frac{U_i-u_{n,l}}{h})|Y_{ij}|\frac{|Y_{ik}|^{3+\gamma}}{\delta_n^{2+\gamma}}I\{|Y_{ik}| > \delta_n\} \\
=\ & o(\frac{\sqrt{\log p}}{n^{3/5}}).
\end{aligned}$$

Thus, we have

$$H_2 \leq \sum_{i,j} P(|Y_{ij}| > \delta_n) \leq \frac{npE|Y_{ij}|^{5+5\gamma+\tau}}{\delta_n^{5+5\gamma+\tau}} \leq \frac{K_2 p(\log n)^{(5+5\gamma+\tau)/2}}{n^{\gamma+\tau/5}}. \tag{22}$$

We note that the upper bounds of H_1 in (21) and H_2 in (22) do not depend on u. Combining (20), (21) and (22), following similar arguments in the proofs of Lemma 7 and Lemma 8, we have

$$\max_{j,k} \sup_{u\in\Omega} |\hat{\sigma}_{jk}(u) - \sigma_{jk}(u)| = O_p(\frac{\sqrt{\log p}}{n^{2/5}}).$$

Then, following the same arguments in the proof of Theorem 1, we obtain the result. □

Proof of Proposition 3. It is easy to see

$$\begin{aligned} &\sup_u ||[s_{\lambda_n(u)}(\hat{\Sigma}(u))]^{-1} - \Sigma^{-1}(u)|| \\ &\quad \leq \sup_u ||[s_{\lambda_n(u)}(\hat{\Sigma}(u))]^{-1}|| \cdot \sup_u ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| \cdot \sup_u ||\Sigma^{-1}(u)||. \end{aligned} \tag{23}$$

Since $||\Sigma^{-1}(u)|| = \frac{1}{\lambda_{min}(\Sigma(u))}$, we have

$$\sup_u ||\Sigma^{-1}(u)|| \leq \epsilon^{-1} < \infty. \tag{24}$$

Recall $\sup_u ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| = O_p\Big(c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}\Big)$. For any $\eta > 0$ which is small enough, there exists $M^* > 0$ large enough such that $P\Big(\sup_u ||s_{\lambda_n(u)}(\hat{\Sigma}(u)) - \Sigma(u)|| \leq r_n\Big) \geq 1 - \eta$, where $r_n := M^* c_0(p)(\sqrt{\frac{\log p}{nh}} + h^2\sqrt{\log p})^{1-q}$. Thus, for any vector v with $\sum_{i=1}^p v_i^2 = 1$, we have, for given u,

$$v^T s_{\lambda_n(u)}(\hat{\Sigma}(u))v \geq v^T\Sigma(u)v - r_n \geq \lambda_{min}(\Sigma(u)) - r_n > \epsilon/2 > 0. \tag{25}$$

It follows

$$\inf_u \lambda_{min}\{s_{\lambda_n(u)}(\hat{\Sigma}(u))\} \geq \epsilon/3 > 0. \tag{26}$$

Since $||[s_{\lambda_n(u)}(\hat{\Sigma}(u))]^{-1}|| = \frac{1}{\lambda_{min}\{s_{\lambda_n(u)}(\hat{\Sigma}(u))\}}$, it is seen

$$\sup_u ||[s_{\lambda_n(u)}(\hat{\Sigma}(u))]^{-1}|| \leq (\epsilon/3)^{-1} < \infty. \tag{27}$$

We note that the inequalities in (25), (26), and (27) hold with probabilities larger than $1 - \eta$, respectively. Since $\eta > 0$ is arbitrarily small, combining (23), (24), (27) and Theorem 1, we have the result. □

References

Ahmed, A. and Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. Proceedings of the National Academy of Sciences, 106, 11878–11883.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association, 57, 33–45.

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

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

Biswal, B., Yetkin F. Z., Haughton V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic Resonance in Medicine, 34, 537–541.

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

Cai, T. T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106, 594–607.

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

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

Calhoun, V. D., Miller, R., Pearlson, G., and Adall, T. (2014). The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron, 84, 262-274.

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

Chen, Z. and Leng, C. (2015). Local linear estimation of covariance matrices via Cholesky decomposition. Statisitca Sinica, to appear.

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering, Human Brain Mapping, 33, 1914-1928.

Cui, Y., Leng, C., and Sun, D. (2015). Sparse estimation of high-dimensional correlation matrices. Computational Statistics and Data Analysis, to appear.

Danaher, D., Wang, P. and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society B, 76, 373–397.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. The Annals of Statistics, 33, 1380–1403.

Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. Chapman and Hall/CRC.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli, 11, 1031–1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348–1360.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. Biometrika, 98, 1–15.

Kolar, M., Song, L., Ahmed, A., and Xing, E. (2010). Estimating time-varying networks. The Annals of Applied Statistics, 4, 94–123.

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

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. The Annals of Statistics, 40, 908–940.

Lindquist, M. A., Xu, Y., Nebel, M. B., and Caffo, B. S. (2014). Evaluating dynamic bivariate correlations in resting-state fMRI: A comparison study and a new approach. NeuroImage, 101, 531–546.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics, 40, 2293–2326.

Pagan, A. and Ullah, A. (1999). Nonparametric econometrics. Cambridge University Press, Cambridge.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. Biometrika, 99, 733–740.

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

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. The Annals of Statistics, **6**, 177–184.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. The Annals of Probablity, 22, 28–76.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B, 58, 267–288.

Wand, M. M. and Jones, M. C. (1995). Kernel smoothing. CRC Press.

Xue, L., Ma, S., and Zou, H. (2012). Positive definite ℓ_1 penalized estimation of large covariance matrices. Journal of the American Statistical Association, 107, 1480–1491.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. The Annals of Statistics, 40, 2541–2571.

Yin, J., Geng, Z., Li, R., and Wang, H. (2010). Nonparametric covariance model. Statistica Sinica, 20, 469–479.

Yu, K. and Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. Journal of the American Statistical Association, 99, 139–144.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research, 99, 2261–2286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B, 68, 49–67.

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

Zhou, S., Lafferty, J. and Wasserman, L. (2010). Time varying undirected graphs. Machine Learning Journal, 80, 295–319.

Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418–1429.