Mach Learn (2010) 80: 295–319
DOI 10.1007/s10994-010-5180-0

Time varying undirected graphs

Shuheng Zhou · John Lafferty · Larry Wasserman

Received: 15 March 2009 / Accepted: 1 November 2009 / Published online: 27 April 2010

Abstract Undirected graphs are often used to describe high dimensional distributions. Under sparsity conditions, the graph can be estimated using ℓ_1 penalization methods. However, current methods assume that the data are independent and identically distributed. If the distribution, and hence the graph, evolves over time then the data are not longer identically distributed. In this paper we develop a nonparametric method for estimating time varying graphical structure for multivariate Gaussian distributions using an ℓ_1 regularization method, and show that, as long as the covariances change smoothly over time, we can estimate the covariance matrix well (in predictive risk) even when p is large.

Keywords Graph selection · ℓ_1 regularization · High dimensional asymptotics · Risk consistency

1 Introduction

Let $Z = (Z_1, \ldots, Z_p)^T$ be a random vector with distribution P. The distribution can be represented by an undirected graph $G = (V, F)$. The vertex set V has one vertex for each component of the vector Z. The edge set F consists of pairs (j, k) that are joined by an edge. If Z_j is independent of Z_k given the other variables, then (j, k) is not in F. When Z

Editors: Sham Kakade and Ping Li.

S. Zhou (✉)
Seminar für Statistik, ETH Zürich HG G 10.2, Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: zhou@stat.math.ethz.ch

J. Lafferty
Computer Science Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
e-mail: lafferty@cs.cmu.edu

L. Wasserman
Department of Statistics, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
e-mail: larry@stat.cmu.edu

is Gaussian, missing edges correspond to zeroes in the inverse covariance matrix Σ^{-1}. Suppose we have independent, identically distributed data $D = (Z^1, \ldots, Z^t, \ldots, Z^n)$ from P. When the dimension p is small, the graph may be estimated from D by testing which partial correlations are not significantly different from zero (Drton and Perlman 2004). When p is large, estimating G is much more difficult. However, if the graph is sparse and the data are Gaussian, then several methods can successfully estimate G (see Meinshausen and Bühlmann 2006; Banerjee et al. 2008; Friedman et al. 2008; Lam and Fan 2009; Bickel and Levina 2008; Rothman et al. 2008).

These recently developed methods assume that the graphical structure is stable over time. But it is easy to imagine cases where such stability would fail. For example, Z^t could represent a large vector of stock prices at time t, and the conditional independence structure between stocks could easily change over time. Another example is gene expression levels. As a cell moves through its metabolic cycle, the conditional independence relations between proteins could change.

In this paper we develop a nonparametric method for estimating time varying graphical structure for multivariate Gaussian distributions using an ℓ_1 regularization method. We show that, as long as the covariances change smoothly over time, we can estimate the covariance matrix well in terms of predictive risk even when p is large. We make the following theoretical contributions: (a) nonparametric predictive risk consistency and rate of convergence of the covariance matrices, (b) consistency and rate of convergence in Frobenius norm of the inverse covariance matrix, (c) large deviation results for covariance matrices for non-identically distributed observations, and (d) conditions that guarantee smoothness of the covariances. In addition, we provide simulation evidence that our method can accurately recover graphical structure. To the best of our knowledge, these are the first such results on time varying undirected graphs in the high dimensional setting.

2 The model and method

Let $Z^t \sim N(0, \Sigma(t))$ be independent. It will be useful to index time as $t = 0, 1/n, 2/n, \ldots, 1$ and thus the data are $D_n = (Z^t : t = 0, 1/n, \ldots, 1)$. Associated with each Z^t is its undirected graph $G(t)$. Under the assumption that the law $\mathcal{L}(Z^t)$ of Z^t changes smoothly, we estimate the graph sequence $G(1), G(2), \ldots$. The graph $G(t)$ is determined by the zeroes of $\Sigma(t)^{-1}$. In this paper we investigate a simple time series model of the following form

$$W^0 \sim N(0, \Sigma(0))$$
$$W^t = W^{t-1} + Z^t, \quad \text{where } Z^t \sim N(0, \Sigma(t)), \text{ for } t > 0.$$

Ultimately, we are interested in the general time series model where the Z^ts are dependent and the graphs change over time. For simplicity, however, we assume independence but allow the graphs to change. Indeed, it is the changing graph, rather than the dependence, that is the biggest hurdle to deal with.

In the i.i.d. case, recent work (Banerjee et al. 2008; Friedman et al. 2008) has considered ℓ_1-penalized maximum likelihood estimators over the entire set of positive definite matrices. These estimators are given by

$$\hat{\Sigma}_n = \arg\min_{\Sigma \succ 0} \{\text{tr}(\Sigma^{-1}\hat{S}_n) + \log|\Sigma| + \lambda|\Sigma^{-1}|_1\} \tag{1}$$

where $\hat{S}_n$ is the sample covariance matrix. In the non-i.i.d. case our approach is to estimate $\Sigma(t)$ at time t by

$$\hat{\Sigma}_n(t) = \arg\min_{\Sigma \succ 0} \left\{ \operatorname{tr}(\Sigma^{-1}\hat{S}_n(t)) + \log|\Sigma| + \lambda|\Sigma^{-1}|_1 \right\}$$

where

$$\hat{S}_n(t) = \frac{\sum_s w_{st} Z_s Z_s^T}{\sum_s w_{st}} \tag{2}$$

is a weighted covariance matrix, with weights $w_{st} = K(|s-t|/h_n)$ given by a symmetric nonnegative kernel over time. In other words, $\hat{S}_n(t)$ is just the regularized kernel estimator of the covariance at time t. An attraction of this approach is that it can use existing software for covariance estimation in the i.i.d. setting.

2.1 Notation

We use the following notation throughout the rest of the paper. For any matrix $W = (w_{ij})$, let $|W|$ denote the determinant of W, $\operatorname{tr}(W)$ the trace of W. Let $\varphi_{\max}(W)$ and $\varphi_{\min}(W)$ be the largest and smallest eigenvalues, respectively. We write $W^{\searrow} = \operatorname{diag}(W)$ for a diagonal matrix with the same diagonal as W, and $W^{\diamondsuit} = W - W^{\searrow}$. The matrix Frobenius norm is given by $\|W\|_F = \sqrt{\sum_i \sum_j w_{ij}^2}$. The operator norm $\|W\|_2^2$ is given by $\varphi_{\max}(WW^T)$. We write $|\cdot|_1$ for the ℓ_1 norm of a matrix vectorized, i.e., for a matrix $|W|_1 = \|\operatorname{vec} W\|_1 = \sum_i \sum_j |w_{ij}|$, and write $\|W\|_0$ for the number of non-zero entries in the matrix. We use $\Theta(t) = \Sigma^{-1}(t)$.

3 Risk consistency

In this section we define the loss and risk. The risk is defined as follows. Let $Z \sim N(0, \Sigma_0)$ and let Σ be a positive definite matrix. Let

$$R(\Sigma) = \operatorname{tr}(\Sigma^{-1}\Sigma_0) + \log|\Sigma|. \tag{3}$$

Note that, up to an additive constant,

$$R(\Sigma) = -2\mathbf{E}_0(\log f_\Sigma(Z)),$$

where f_Σ is the density for $N(0, \Sigma)$. We say that $\hat{G}_n(t)$ is *persistent* (Greenshtein and Ritov 2004) with respect to a class of positive definite matrices $\mathcal{S}_n$ if $R(\hat{\Sigma}_n) - \min_{\Sigma \in \mathcal{S}_n} R(\Sigma) \xrightarrow{P} 0$. In the i.i.d. case, ℓ_1 regularization yields a persistent estimator, as we now show.

The maximum likelihood estimate minimizes

$$\hat{R}_n(\Sigma) = \operatorname{tr}(\Sigma^{-1}\hat{S}_n) + \log|\Sigma|,$$

where $\hat{S}_n$ is the sample covariance matrix. Minimizing $\hat{R}_n(\Sigma)$ without constraints gives $\hat{\Sigma}_n = \hat{S}_n$. We would like to minimize $\hat{R}_n(\Sigma)$ subject to $\|\Sigma^{-1}\|_0 \le L$. This would give the "best" sparse graph G, but it is not a convex optimization problem. Hence we estimate $\hat{\Sigma}_n$ by solving a convex relaxation problem as written in (1) instead. Algorithms for carrying out

this optimization are given by Banerjee et al. (2008), Friedman et al. (2008). Given $L_n, \forall n$, let

$$\mathcal{S}_n = \{\Sigma : \Sigma \succ 0, |\Sigma^{-1}|_1 \leq L_n\}. \tag{4}$$

We define the oracle estimator as (6)

$$\Sigma^*(n) = \arg\min_{\Sigma \in \mathcal{S}_n} R(\Sigma) \tag{5}$$

and write (1) as

$$\hat{\Sigma}_n = \arg\min_{\Sigma \in \mathcal{S}_n} \hat{R}_n(\Sigma). \tag{6}$$

Note that one can choose to only penalize off-diagonal elements of Σ^{-1} as in Rothman et al. (2008), if desired. We have the following result, whose proof appears in Sect. 3.2.

Theorem 1 *Suppose that* $p_n \leq n^\xi$ *for some* $\xi \geq 0$ *and*

$$L_n = o\left(\frac{n}{\log p_n}\right)^{1/2}$$

in (4). *Then for the sequence of empirical estimators as defined in* (6) *and the oracle* $\Sigma^*(n)$ *as in* (5),

$$R(\hat{\Sigma}_n) - R(\Sigma^*(n)) \xrightarrow{P} 0.$$

3.1 Risk consistency for the non-identical case

In the non-i.i.d. case we estimate $\Sigma(t)$ at time $t \in [0, 1]$. Given $\Sigma(t)$, let

$$\hat{R}_n(\Sigma(t)) = \operatorname{tr}(\Sigma(t)^{-1}\hat{S}_n(t)) + \log|\Sigma(t)|.$$

For a given ℓ_1 bound L_n, we define $\hat{\Sigma}_n(t)$ as the minimizer of $\hat{R}_n(\Sigma)$ subject to $\Sigma \in \mathcal{S}_n$,

$$\hat{\Sigma}_n(t) = \arg\min_{\Sigma \in \mathcal{S}_n} \{\operatorname{tr}(\Sigma^{-1}\hat{S}_n(t)) + \log|\Sigma|\} \tag{7}$$

where $\hat{S}_n(t)$ is given in (2), with $K(\cdot)$ a symmetric nonnegative function that satisfies

Assumption 1 *The kernel function* K *has a compact support* $[-1, 1]$.

Throughout this section, we assume that the constants do not depend on n (or p).

Lemma 1 *Let* $\Sigma(t) = [\sigma_{jk}(t)]$. *Suppose the following conditions hold*:

1. *There exists* $C_0, C > 0$ *such that* $\max_{j,k} \sup_t |\sigma'_{jk}(t)| \leq C_0$ *and* $\max_{j,k} \sup_t |\sigma''_{jk}(t)| \leq C$;
2. $p_n \leq n^\xi$ *for some* $\xi \geq 0$;
3. $h_n \asymp n^{-1/3}$.

Then

$$\max_{j,k} |\hat{S}_n(t, j, k) - \Sigma(t, j, k)| = O_P\left(\frac{\sqrt{\log n}}{n^{1/3}}\right) \tag{8}$$

for all $t > 0$.

Proof By the triangle inequality,

$$|\hat{S}_n(t,j,k) - \Sigma(t,j,k)| \le |\hat{S}_n(t,j,k) - \mathbf{E}\hat{S}_n(t,j,k)| + |\mathbf{E}\hat{S}_n(t,j,k) - \Sigma(t,j,k)|.$$

By Lemma 5, we have for $h_n \asymp n^{-1/3}$,

$$\max_{j,k} \sup_t |\mathbf{E}\hat{S}_n(t,j,k) - \Sigma(t,j,k)| = O(h_n). \tag{9}$$

In Lemma 6, we show that for some constant $c_1 > 0$,

$$\mathbf{P}\big(|\hat{S}_n(t,j,k) - \mathbf{E}\hat{S}_n(t,j,k)| > \epsilon\big) \le \exp\{-c_1 h_n n \epsilon^2\}.$$

Hence,

$$\mathbf{P}\big(\max_{j,k} |\hat{S}_n(t,j,k) - \mathbf{E}\hat{S}_n(t,j,k)| > \epsilon\big) \le \exp\{-nh_n(c_1\epsilon^2 - 2\xi \log n/(nh_n))\}$$

and $\max_{j,k} |\hat{S}_n(t,j,k) - \mathbf{E}\hat{S}_n(t,j,k)| = O_P(\sqrt{\frac{\log n}{nh_n}})$ for $h_n \asymp n^{-1/3}$. □

With the use of Lemma 1, the proof of the following follows the same lines as that of Theorem 1.

Theorem 2 *Suppose all conditions in Lemma 1 and the following hold*:

$$L_n = o\left(\frac{n^{1/3}}{\sqrt{\log n}}\right). \tag{10}$$

Then, for all $t > 0$, for the sequence of estimators as in (7),

$$R(\hat{\Sigma}_n(t)) - R(\Sigma^*(t)) \xrightarrow{P} 0.$$

Remark 1 If a local linear smoother is substituted for a kernel smoother, the rate can be improved from $n^{1/3}$ to $n^{2/5}$ as the bias will be bounded as $O(h^2)$ in (9).

Remark 2 Suppose that $\forall i, j$, if $\theta_{ij} \neq 0$, we have $\theta_{ij} = \Omega(1)$. Then condition (10) allows that $|\Theta|_1 = L_n$; hence if $p = n^\xi$ and $\xi < 1/3$, we have that $\|\Theta\|_0 = \Omega(p)$. Thus, the family of graphs that we can guarantee persistency for, although sparse, is likely to include connected graphs.

Remark 3 Under appropriate assumptions, we can achieve a sparsistency property defined as follows. Consider estimates $\hat{G}_n(t) = (V, \hat{F}_n)$. Consider the size of the symmetric difference between the true and estimated edge sets:

$$L(G(t), \hat{G}_n(t)) = |F(t) \Delta \hat{F}_n(t)|.$$

We say that $\hat{G}_n(t)$ is *sparsistent* if $L(G(t), \hat{G}_n(t)) \xrightarrow{P} 0$ as $n \to \infty$. Now assume that the graph structure does not change in a neighborhood of t, that is, we assume that $G(t)$ remains invariant in $(t-\epsilon, t+\epsilon)$ for some $\epsilon > 0$. Under suitable incoherence assumptions as in Ravikumar et al. (2008) (Theorem 2), one can derive sparsistency results as in that paper given the bound in (8). We omit a formal theorem regarding this property and refer to Ravikumar et al. (2008) for details instead; our experiments in Sect. 8 evaluate two metrics pertaining to this property.

The smoothness condition in Lemma 1 is expressed in terms of the elements of $\Sigma(t) = [\sigma_{ij}(t)]$. It might be more natural to impose smoothness on $\Theta(t) = \Sigma(t)^{-1}$ instead. In fact, smoothness of Θ_t implies smoothness of Σ_t as the next result shows. Let us first specify two assumptions. We use $\sigma_i^2(x)$ as an alternate notation for $\sigma_{ii}(x)$.

Definition 1 For a function $u : [0, 1] \to \mathbf{R}$, let $\|u\|_\infty = \sup_{x\in[0,1]} |u(x)|$.

We emphasize that all constants below do not depend on n (or p).

Assumption 2 *There exists some constant $S_0 < \infty$ such that*

$$\max_{i=1\ldots,p} \sup_{t\in[0,1]} |\sigma_i(t)| \le S_0 < \infty, \quad \textit{hence} \quad \max_{i=1\ldots,p} \|\sigma_i\|_\infty \le S_0. \tag{11}$$

Assumption 3 *Let $\theta_{ij}(t), \forall i, j$, be twice differentiable functions such that $\theta'_{ij}(t) < \infty$ and $\theta''_{ij}(t) < \infty, \forall t \in [0, 1]$. In addition, there exist constants $S_1, S_2 < \infty$ such that*

$$\sup_{t\in[0,1]} \sum_{k=1}^{p}\sum_{\ell=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p} |\theta'_{ki}(t)\theta'_{\ell j}(t)| \le S_1$$

$$\sup_{t\in[0,1]} \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta''_{k\ell}(t)| \le S_2,$$

where the first inequality guarantees that $\sup_{t\in[0,1]} \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta'_{k\ell}(t)| < \sqrt{S_1} < \infty$.

Lemma 2 *Denote the elements of $\Theta(t) = \Sigma(t)^{-1}$ by $\theta_{jk}(t)$. Under Assumptions 2 and 3, the smoothness condition in Lemma 1 holds.*

The proof is in Sect. 6. In Sect. 7, we show some preliminary results on achieving upper bounds on quantities that appear in Condition 1 of Lemma 1 through the sparsity level of the inverse covariance matrix, i.e., $\|\Theta_t\|_0, \forall t \in [0, 1]$.

3.2 Proof of Theorem 1

Note that for all n and for all Σ, we have

$$|R(\Sigma) - \hat{R}_n(\Sigma)| \le \sum_{j,k} |\Sigma^{-1}_{jk}|\, |\hat{S}_n(j,k) - \Sigma_0(j,k)| \le \delta_n |\Sigma^{-1}|_1,$$

where it follows from Rothman et al. (2008) that

$$\delta_n = \max_{j,k} |\hat{S}_n(j,k) - \Sigma_0(j,k)| = O_P(\sqrt{\log p/n}).$$

Hence, minimizing over $\mathcal{S}_n$ with $L_n = o(\frac{n}{\log p_n})^{1/2}$,

$$\sup_{\Sigma\in\mathcal{S}_n} |R(\Sigma) - \hat{R}_n(\Sigma)| = o_P(1).$$

By the definitions of $\Sigma^*(n) \in \mathcal{S}_n$ and $\hat{\Sigma}_n \in \mathcal{S}_n$, we immediately have $R(\Sigma^*(n)) \leq R(\hat{\Sigma}_n)$ and $\hat{R}_n(\hat{\Sigma}_n) \leq \hat{R}_n(\Sigma^*(n))$; thus

$$\begin{aligned} 0 &\leq R(\hat{\Sigma}_n) - R(\Sigma^*(n)) \\ &= R(\hat{\Sigma}_n) - \hat{R}_n(\hat{\Sigma}_n) + \hat{R}_n(\hat{\Sigma}_n) - R(\Sigma^*(n)) \\ &\leq R(\hat{\Sigma}_n) - \hat{R}_n(\hat{\Sigma}_n) + \hat{R}_n(\Sigma^*(n)) - R(\Sigma^*(n)). \end{aligned}$$

Using the triangle inequality and $\hat{\Sigma}_n, \Sigma^*(n) \in \mathcal{S}_n$,

$$\begin{aligned} |R(\hat{\Sigma}_n) - R(\Sigma^*(n))| &\leq |R(\hat{\Sigma}_n) - \hat{R}_n(\hat{\Sigma}_n) + \hat{R}_n(\Sigma^*(n)) - R(\Sigma^*(n))| \\ &\leq |R(\hat{\Sigma}_n) - \hat{R}_n(\hat{\Sigma}_n)| + |\hat{R}_n(\Sigma^*(n)) - R(\Sigma^*(n))| \\ &\leq 2 \sup_{\Sigma \in \mathcal{S}_n} |R(\Sigma) - \hat{R}_n(\Sigma)|. \end{aligned}$$

Thus, for all $\epsilon > 0$, the event $\{|R(\hat{\Sigma}_n) - R(\Sigma^*(n))| > \epsilon\}$ is contained in the event

$$\Big\{ \sup_{\Sigma \in \mathcal{S}_n} |R(\Sigma) - \hat{R}_n(\Sigma)| > \epsilon/2 \Big\}.$$

Thus, for $L_n = o((n/\log n)^{1/2})$, and for all $\epsilon > 0$,

$$\mathbf{P}\big(|R(\hat{\Sigma}_n) - R(\Sigma^*(n))| > \epsilon\big) \leq \mathbf{P}\Big(\sup_{\Sigma \in \mathcal{S}_n} |R(\Sigma) - \hat{R}_n(\Sigma)| > \epsilon/2 \Big) \to 0$$

as $n \to \infty$. □

4 Frobenius norm consistency

In this section, we show an explicit convergence rate in the Frobenius norm for estimating $\Theta(t), \forall t$, where $p, |F|$ grow with n, so long as the covariances change smoothly over t. Note that certain smoothness assumptions on a matrix W would guarantee the corresponding smoothness conditions on its inverse W^{-1}, so long as W is non-singular, as we show in Sect. 6. We first write our time-varying estimator $\hat{\Theta}_n(t)$ for $\Sigma^{-1}(t)$ at time $t \in [0, 1]$ as the minimizer of the ℓ_1 regularized negative smoothed log-likelihood over the entire set of positive definite matrices,

$$\hat{\Theta}_n(t) = \arg\min_{\Theta \succ 0} \{\mathrm{tr}(\Theta \hat{S}_n(t)) - \log|\Theta| + \lambda_n |\Theta|_1\} \tag{12}$$

where λ_n is a non-negative regularization parameter, and $\hat{S}_n(t)$ is the smoothed sample covariance matrix using a kernel function as defined in (2).

Now fix a point of interest t_0. In the following, we use $\Sigma_0 = (\sigma_{ij}(t_0))$ to denote the true covariance matrix at this time. Let $\Theta_0 = \Sigma_0^{-1}$ be its inverse matrix. Define the set $S = \{(i, j) : \theta_{ij}(t_0) \neq 0,\ i \neq j\}$. Then $|S| = s$. Note that $|S|$ is twice the number of edges in the graph $G(t_0)$. We make the following assumptions.

Assumption 4 *Assume* $p + s = o(n^{(2/3)}/\log n)$ *and* $\varphi_{\min}(\Sigma_0) \geq \underline{k} > 0$, *hence* $\varphi_{\max}(\Theta_0) \leq 1/\underline{k}$, *and* $\varphi_{\min}(\Theta_0) = \Omega(2\sqrt{\frac{(p+s)\log n}{n^{2/3}}})$.

The proof draws upon techniques from Rothman et al. (2008), with modifications necessary to handle the fact that we penalize $|\Theta|_1$ rather than $|\Theta^{\Diamond}|_1$ as in their case.

Theorem 3 *Let $\hat{\Theta}_n(t)$ be the minimizer defined by (12). Suppose all conditions in Lemma 1 and Assumption 4 hold. If $\lambda_n \asymp \sqrt{\frac{\log n}{n^{2/3}}}$, then*

$$\|\hat{\Theta}_n(t) - \Theta_0\|_F = O_P\left(\sqrt{\frac{(p+s)\log n}{n^{2/3}}}\right).$$

Proof Let $\underline{0}$ be a matrix with all entries being zero. Let

$$\begin{aligned} Q(\Theta) &= \operatorname{tr}(\Theta \hat{S}_n(t_0)) - \log|\Theta| + \lambda|\Theta| - \operatorname{tr}(\Theta_0 \hat{S}_n(t_0)) + \log|\Theta_0| - \lambda|\Theta_0|_1 \\ &= \operatorname{tr}((\Theta - \Theta_0)(\hat{S}_n(t) - \Sigma_0)) - (\log|\Theta| - \log|\Theta_0|) + \operatorname{tr}((\Theta - \Theta_0)\Sigma_0) \\ &\quad + \lambda(|\Theta|_1 - |\Theta_0|_1). \end{aligned}$$

Now, $\hat{\Theta}$ minimizes $Q(\Theta)$, or equivalently $\hat{\Delta}_n = \hat{\Theta} - \Theta_0$ minimizes $G(\Delta) \equiv Q(\Theta_0 + \Delta)$. Hence $G(\underline{0}) = 0$ and $G(\hat{\Theta}_n) \leq G(\underline{0}) = 0$ by definition. Define for some constant C_3, $\delta_n = C_3\sqrt{\frac{\log n}{n^{2/3}}}$. Now, let

$$\lambda_n = \frac{C_3}{\varepsilon}\sqrt{\frac{\log n}{n^{2/3}}} = \frac{\delta_n}{\varepsilon} \quad \text{for some } 0 < \varepsilon < 1.$$

Consider now the set

$$\mathcal{T}_n = \{\Delta : \Delta = B - \Theta_0,\, B, \Theta_0 \succ 0,\, \|\Delta\|_F = Mr_n\},$$

where

$$r_n = \sqrt{\frac{(p+s)\log n}{n^{2/3}}} \asymp \delta_n\sqrt{p+s} \to 0. \tag{13}$$

Proposition 1 *Under Assumption 4, for all $\Delta \in \mathcal{T}_n$ such that $\|\Delta\|_F = o(1)$ as in (13), $\Theta_0 + v\Delta \succ 0, \forall v \in I \supset [0,1]$.*

Proof It is sufficient to show that $\Theta_0 + (1+\varepsilon)\Delta \succ 0$ and $\Theta_0 - \varepsilon\Delta \succ 0$ for some $1 > \varepsilon > 0$. Indeed, $\varphi_{\min}(\Theta_0 + (1+\varepsilon)\Delta) \geq \varphi_{\min}(\Theta_0) - (1+\varepsilon)\|\Delta\|_2 > 0$ for $\varepsilon < 1$, given that $\varphi_{\min}(\Theta_0) = \Omega(2Mr_n)$ and $\|\Delta\|_2 \leq \|\Delta\|_F = Mr_n$. Similarly, $\varphi_{\min}(\Theta_0 - \varepsilon\Delta) \geq \varphi_{\min}(\Theta_0) - \varepsilon\|\Delta\|_2 > 0$ for $\varepsilon < 1$. □

Thus we have that $\log\det(\Theta_0 + v\Delta)$ is infinitely differentiable on the open interval $I \supset [0,1]$ of v. This allows us to use the Taylor's formula with integral remainder to obtain the following lemma:

Lemma 3 *With probability $1 - 1/n^c$ for some $c \geq 2$, $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$.*

Proof Let us use A as a shorthand for

$$\operatorname{vec}\Delta^T\left(\int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv\right)\operatorname{vec}\Delta,$$

where $\otimes$ is the Kronecker product (if $W = (w_{ij})_{m\times n}$, $P = (b_{k\ell})_{p\times q}$, then $W \otimes P = (w_{ij}P)_{mp\times nq}$), and $\mathrm{vec}\Delta \in \mathbf{R}^{\mathbf{p^2}}$ is $\Delta_{p\times p}$ vectorized. Now, the Taylor expansion gives

$$\begin{aligned}\log|\Theta_0+\Delta| - \log|\Theta_0| &= \frac{d}{dv}\log|\Theta_0+v\Delta||_{v=0}\Delta + \int_0^1 (1-v)\frac{d^2}{dv^2}\log\det(\Theta_0+v\Delta)dv\\ &= \mathrm{tr}(\Sigma_0\Delta) + A,\end{aligned}$$

where by symmetry, $\mathrm{tr}(\Sigma_0\Delta) = \mathrm{tr}(\Theta-\Theta_0)\Sigma_0$. Hence

$$G(\Delta) = A + \mathrm{tr}(\Delta(\hat{S}_n - \Sigma_0)) + \lambda_n(|\Theta_0+\Delta|_1 - |\Theta_0|_1). \tag{14}$$

For an index set S and a matrix $W = [w_{ij}]$, write $W_S \equiv (w_{ij}I((i,j)\in S))$, where $I(\cdot)$ is an indicator function. Recall $S = \{(i,j) : \Theta_{0ij} \neq 0,\ i \neq j\}$ and let $S^c = \{(i,j) : \Theta_{0ij} = 0,\ i \neq j\}$. Hence in our notation,

$$\Theta = \Theta^{\searrow} + \Theta_S^{\diamondsuit} + \Theta_{S^c}^{\diamondsuit}.$$

Note that we have $\Theta_{0S^c}^{\diamondsuit} = \underline{0}$,

$$\begin{aligned}|\Theta_0^{\diamondsuit} + \Delta^{\diamondsuit}|_1 &= |\Theta_{0S}^{\diamondsuit} + \Delta_S^{\diamondsuit}|_1 + |\Delta_{S^c}^{\diamondsuit}|_1,\\ |\Theta_0^{\diamondsuit}|_1 &= |\Theta_{0S}^{\diamondsuit}|_1, \quad \text{hence}\\ |\Theta_0^{\diamondsuit} + \Delta^{\diamondsuit}|_1 - |\Theta_0^{\diamondsuit}|_1 &\geq |\Delta_{S^c}^{\diamondsuit}|_1 - |\Delta_S^{\diamondsuit}|_1,\\ |\Theta_0^{\searrow} + \Delta^{\searrow}|_1 - |\Theta_0^{\searrow}|_1 &\geq -|\Delta^{\searrow}|_1,\end{aligned}$$

where the last two steps follow from the triangle inequality. Therefore

$$\begin{aligned}|\Theta_0+\Delta|_1 - |\Theta_0|_1 &= |\Theta_0^{\diamondsuit} + \Delta^{\diamondsuit}|_1 - |\Theta_0^{\diamondsuit}|_1 + |\Theta_0^{\searrow} + \Delta^{\searrow}|_1 - |\Theta_0^{\searrow}|_1\\ &\geq |\Delta_{S^c}^{\diamondsuit}|_1 - |\Delta_S^{\diamondsuit}|_1 - |\Delta^{\searrow}|_1.\end{aligned} \tag{15}$$

Now, from Lemma 1,

$$\max_{j,k}|\hat{S}_n(t,j,k) - \sigma(t,j,k)| = O_P\left(\frac{\sqrt{\log n}}{n^{1/3}}\right) = O_P(\delta_n).$$

By (10), with probability $1-\frac{1}{n^2}$ we have $|\mathrm{tr}(\Delta(\hat{S}_n - \Sigma_0))| \leq \delta_n|\Delta|_1$; hence by (15),

$$\begin{aligned}&\mathrm{tr}(\Delta(\hat{S}_n - \Sigma_0)) + \lambda_n(|\Theta_0+\Delta|_1 - |\Theta_0|_1)\\ &\quad\geq -\delta_n|\Delta^{\searrow}|_1 - \delta_n|\Delta_{S^c}^{\diamondsuit}|_1 - \delta_n|\Delta_S^{\diamondsuit}|_1 - \lambda_n|\Delta^{\searrow}|_1 + \lambda_n|\Delta_{S^c}^{\diamondsuit}|_1 - \lambda_n|\Delta_S^{\diamondsuit}|_1\\ &\quad\geq -(\delta_n+\lambda_n)(|\Delta^{\searrow}|_1 + |\Delta_S^{\diamondsuit}|_1) + (\lambda_n - \delta_n)|\Delta_{S^c}^{\diamondsuit}|_1\\ &\quad\geq -(\delta_n+\lambda_n)(|\Delta^{\searrow}|_1 + |\Delta_S^{\diamondsuit}|_1),\end{aligned} \tag{16}$$

where

$$\begin{aligned}(\delta_n+\lambda_n)(|\Delta^{\searrow}|_1 + |\Delta_S^{\diamondsuit}|_1) &\leq (\delta_n+\lambda_n)(\sqrt{p}\|\Delta^{\searrow}\|_F + \sqrt{s}\|\Delta_S^{\diamondsuit}\|_F)\\ &\leq (\delta_n+\lambda_n)(\sqrt{p}\|\Delta^{\searrow}\|_F + \sqrt{s}\|\Delta^{\diamondsuit}\|_F)\end{aligned}$$

$$\leq (\delta_n + \lambda_n) \max\{\sqrt{p}, \sqrt{s}\}(\|\Delta^{\searrow}\|_F + \|\Delta^{\diamondsuit}\|_F)$$
$$\leq (\delta_n + \lambda_n) \max\{\sqrt{p}, \sqrt{s}\}\sqrt{2}\|\Delta\|_F$$
$$\leq \delta_n \frac{1+\varepsilon}{\varepsilon}\sqrt{p+s}\sqrt{2}\|\Delta\|_F. \tag{17}$$

Combining (14), (16), and (17), we have with probability $1 - \frac{1}{n^c}$, for all $\Delta \in \mathcal{T}_n$,

$$G(\Delta) \geq A - (\delta_n + \lambda_n)(|\Delta^{\searrow}|_1 + |\Delta^{\diamondsuit}_S|_1)$$
$$\geq \frac{\underline{k}^2}{2+\tau}\|\Delta\|_F^2 - \delta_n \frac{1+\varepsilon}{\varepsilon}\sqrt{p+s}\sqrt{2}\|\Delta\|_F$$
$$= \|\Delta\|_F^2\left(\frac{\underline{k}^2}{2+\tau} - \delta_n \frac{\sqrt{2}(1+\varepsilon)}{\varepsilon\|\Delta\|_F}\sqrt{p+s}\right)$$
$$= \|\Delta\|_F^2\left(\frac{\underline{k}^2}{2+\tau} - \frac{\delta_n\sqrt{2}(1+\varepsilon)}{\varepsilon M r_n}\sqrt{p+s}\right) > 0$$

for M sufficiently large, where the bound on A comes from Rothman et al. (2008) (see p. 502, proof of Theorem 1 therein). □

Lemma 4 (Rothman et al. 2008) *For some $\tau = o(1)$, under Assumption 4,*

$$\mathrm{vec}\Delta^T\left(\int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv\right)\mathrm{vec}\Delta \geq \|\Delta\|_F^2 \frac{\underline{k}^2}{2+\tau}$$

for all $\Delta \in \mathcal{T}_n$.

We next show the following proposition.

Proposition 2 *If $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all Δ in $\mathcal{V}_n = \{\Delta : \Delta = D - \Theta_0, D \succ 0, \|\Delta\|_F > M r_n$, for r_n as in (13)$\}$. Hence if $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n \cup \mathcal{V}_n$.*

Proof Suppose $G(\Delta') \leq 0$ for some $\Delta' \in \mathcal{V}_n$. Let $\Delta_0 = \frac{M r_n}{\|\Delta'\|_F}\Delta'$. Thus $\Delta_0 = \theta\underline{0} + (1-\theta)\Delta'$, where $0 < 1 - \theta = \frac{M r_n}{\|\Delta'\|_F} < 1$ by definition of Δ_0. Hence $\Delta_0 \in \mathcal{T}_n$ given that $\Theta_0 + \Delta_0 \succ 0$ by Proposition 3. Hence by convexity of $G(\Delta)$, we have that $G(\Delta_0) \leq \theta G(\underline{0}) + (1-\theta)G(\Delta') \leq 0$, contradicting that $G(\Delta_0) > 0$ for $\Delta_0 \in \mathcal{T}_n$. □

By Proposition 2 and the fact that $G(\hat{\Delta}_n) \leq G(0) = 0$, we have the following: If $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $\hat{\Delta}_n \notin (\mathcal{T}_n \cup \mathcal{V}_n)$, that is, $\|\hat{\Delta}_n\|_F < M r_n$, given that $\hat{\Delta}_n = \hat{\Theta}_n - \Theta_0$, where $\hat{\Theta}_n, \Theta_0 \succ 0$. Therefore

$$\mathbf{P}(\|\hat{\Delta}_n\|_F \geq M r_n) = 1 - \mathbf{P}(\|\hat{\Delta}_n\|_F < M r_n)$$
$$\leq 1 - \mathbf{P}(G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n)$$
$$= \mathbf{P}(G(\Delta) \leq 0 \text{ for some } \Delta \in \mathcal{T}_n) < \frac{1}{n^c}.$$

We thus establish that $\|\hat{\Delta}_n\|_F \leq O_P(M r_n)$ and hence the theorem holds. □

Proposition 3 *Let B be a $p \times p$ matrix. If $B \succ 0$ and $B + D \succ 0$, then $B + vD \succ 0$ for all $v \in [0, 1]$.*

Proof We only need to check for $v \in (0, 1)$; $\forall x \in \mathbf{R^p}$, by $B \succ 0$ and $B + D \succ 0$, $x^T Bx > 0$ and $x^T(B + D)x > 0$; hence $x^T Dx > -x^T Bx$. Thus $x^T(B + vD)x = x^T Bx + vx^T Dx > (1 - v)x^T Bx > 0$. □

5 Large deviation inequalities

Before continuing, we explain the notation that we follow throughout this section. We switch notation from t to x and form a regression problem for non-i.i.d. data. Given an interval of $[0, 1]$, the point of interest is $x_0 = 1$. We form a design matrix by sampling a set of n p-dimensional Gaussian random vectors Z^t at $t = 0, 1/n, 2/n, \ldots, 1$, where $Z^t \sim N(0, \Sigma_t)$ are independently distributed. In this section, we index the random vectors Z with $k = 0, 1, \ldots, n$ such that $Z_k = Z^t$ for $k = nt$, with corresponding covariance matrix denoted by Σ_k. Hence

$$Z_k = (Z_{k1}, \ldots, Z_{kp})^T \sim N(0, \Sigma_k), \quad \forall k.$$

These are independent but not identically distributed. We will need to generalize the usual inequalities. In Appendix, via a boxcar kernel function, we use moment generating functions to show that for $\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} Z_k Z_k^T$,

$$P^n(|\hat{\Sigma}_{ij} - \Sigma_{ij}(x_0)| > \epsilon) < e^{-cn\epsilon^2}$$

where $P^n = P_1 \times \cdots \times P_n$ denotes the product measure. We look across n time-varying Gaussian vectors, and roughly, we compare $\hat{\Sigma}_{ij}$ with $\Sigma_{ij}(x_0)$, where $\Sigma(x_0) = \Sigma_n$ is the covariance matrix in the end of the window for $t_0 = n$. Furthermore, we derive inequalities in Sect. 5.1 for a general kernel function.

5.1 Bounds for kernel smoothing

In this section, we derive large deviation inequalities for the covariance matrix based on kernel regression estimations. Recall that we assume that the symmetric nonnegative kernel function K has a bounded support $[-1, 1]$ in Assumption 1. This kernel satisfies

$$\int_{-1}^{1} vK(v)dv = 1$$

$$2\int_{-1}^{0} vK(v)dv \le 2\int_{-1}^{0} K(v)dv = 1$$

$$2\int_{-1}^{0} v^2 K(v)dv \le 1.$$

In order to estimate t_0, instead of taking an average of sample variances/covariances over the last n samples, we use the weighting scheme such that data close to t_0 receives larger weights than those that are far away. Let $\Sigma(x) = (\sigma_{ij}(x))$. Let us define $x_0 = \frac{t_0}{n} = 1$, and $\forall i = 1, \ldots, n$, $x_i = \frac{t_0 - i}{n}$ and

$$\ell_i(x_0) = \frac{2}{nh} K\left(\frac{x_i - x_0}{h}\right) \approx \frac{K(\frac{x_i - x_0}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x_0}{h})}$$

where the approximation is due to replacing the sum with a Riemann integral:

$$\sum_{i=1}^{n} \ell_i(x_0) = \sum_{i=1}^{n} \frac{2}{nh} K\left(\frac{x_i - x_0}{h}\right) \approx 2\int_{-1}^{0} K(v)dv = 1,$$

due to the fact that $K(v)$ has compact support in $[-1, 1]$ and $h \leq 1$. Let $\Sigma_k = (\sigma_{ij}(x_k))$, $\forall k = 1, \ldots, n$, where $\sigma_{ij}(x_k) = \mathrm{cov}(Z_{ki}, Z_{kj}) = \rho_{ij}(x_k)\sigma_i(x_k)\sigma_j(x_k)$ and $\rho_{ij}(x_k)$ is the correlation coefficient between Z_i and Z_j at time x_k. Recall that we have independent $(Z_{ki}Z_{kj})$ for all $k = 1, \ldots, n$ such that $\mathbf{E}(Z_{ki}Z_{kj}) = \sigma_{ij}(x_k)$. Let

$$\Phi_1(i, j) = \frac{1}{n}\sum_{k=1}^{n} \frac{2}{h} K\left(\frac{x_k - x_0}{h}\right)\sigma_{ij}(x_k)$$

hence

$$\mathbf{E}\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj} = \sum_{k=1}^{n} \ell_k(x_0)\sigma_{ij}(x_k) = \Phi_1(i, j).$$

We thus decompose and bound for the point of interest x_0

$$\begin{aligned}
&\left|\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj} - \sigma_{ij}(x_0)\right| \\
&\quad\leq \left|\mathbf{E}\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj} - \sigma_{ij}(x_0)\right| + \left|\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj} - \mathbf{E}\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj}\right| \\
&\quad= \left|\sum_{k=1}^{n} \ell_k(x_0)Z_{ki}Z_{kj} - \Phi_1(i, j)\right| + |\Phi_1(i, j) - \sigma_{ij}(x_0)|. \qquad (18)
\end{aligned}$$

Before we start our analysis on large deviations, we first look at the bias term.

Lemma 5 *Suppose there exists $C > 0$ such that*

$$\max_{i,j} \sup_{t} |\sigma''(t, i, j)| \leq C.$$

Then for $K(\cdot)$ that satisfies

$$\sup_{u\in[x_n, x_0]} K''\left(\frac{u - x_0}{h}\right) = O\left(\frac{1}{h^4}\right), \qquad (19)$$

we have

$$\sup_{t\in[0,1]} \max_{i,j} |\mathbf{E}\hat{S}_n(t, i, j) - \sigma_{ij}(t)| = O(h) + O\left(\frac{1}{n^2h^5}\right).$$

Remark 4 Most smooth kernel functions including the Gaussian kernel satisfy (19).

Proof of Lemma 5 W.l.o.g., let $t = t_0$, hence $\mathbf{E}\hat{S}_n(t,i,j) = \Phi_1(i,j)$. We use the Riemann integral to approximate the sum,

$$\begin{aligned}\Phi_1(i,j) &= \frac{1}{n}\sum_{k=1}^{n}\frac{2}{h}K\left(\frac{x_k - x_0}{h}\right)\sigma_{ij}(x_k)\\ &= \int_{x_n}^{x_0}\frac{2}{h}K\left(\frac{u-x_0}{h}\right)\sigma_{ij}(u)du + O\left(\frac{2}{h}\sup_{u\in[x_n,x_0]}\frac{(K(\frac{u-x_0}{h})\sigma_{ij}(u))''}{n^2}\right)\\ &= 2\int_{-1/h}^{0}K(v)\sigma_{ij}(x_0+hv)dv + O\left(\frac{1}{n^2h^5}\right).\end{aligned}$$

We now use Taylor's formula to replace $\sigma_{ij}(x_0+hv)$ and obtain

$$\begin{aligned}2\int_{-1/h}^{0}K(v)\sigma_{ij}(x_0+hv)dv &= 2\int_{-1}^{0}K(v)\left(\sigma_{ij}(x_0)+hv\sigma'_{ij}(x_0)+\frac{\sigma''_{ij}(y(v))(hv)^2}{2}\right)dv\\ &= \sigma_{ij}(x_0) + 2\int_{-1}^{0}K(v)\left(hv\sigma'_{ij}(x_0)+\frac{C(hv)^2}{2}\right)dv,\end{aligned}$$

where

$$\begin{aligned}2\int_{-1}^{0}K(v)\left(hv\sigma'_{ij}(x_0)+\frac{C(hv)^2}{2}\right)dv &= 2h\sigma'_{ij}(x_0)\int_{-1}^{0}vK(v)dv + \frac{Ch^2}{2}\int_{-1}^{0}v^2K(v)dv\\ &\le h\sigma'_{ij}(x_0)+\frac{Ch^2}{4}\end{aligned}$$

with $y(v) - x_0 < hv$. Thus $\Phi_1(i,j) - \sigma_{ij}(x_0) = O(h) + O(\frac{1}{n^2h^5})$ and the lemma holds. □

We now move on to the large deviation bound for all entries of the smoothed empirical covariance matrix.

Lemma 6 *For* $\epsilon < \frac{C_1(\sigma_i^2(x_0)\sigma_j^2(x_0)+\sigma_{ij}^2(x_0))}{\max_{k=1,\ldots,n}(2K(\frac{x_k-x_0}{h})\sigma_i(x_k)\sigma_j(x_k))}$, *where* C_1 *is defined in Proposition 5, for some* $c_1 > 0$,

$$\mathbf{P}\big(|\hat{S}_n(t,i,j) - \mathbf{E}\hat{S}_n(t,i,j)| > \epsilon\big) \le \exp\{-c_1nh\epsilon^2\}.$$

Proof Let us define $A_k = Z_{ki}Z_{kj} - \sigma_{ij}(x_k)$.

$$\mathbf{P}\big(|\hat{S}_n(t,i,j) - \mathbf{E}\hat{S}_n(t,i,j)| > \epsilon\big) = \mathbf{P}\left(\sum_{k=1}^{n}\ell_k(x_0)Z_{ki}Z_{kj} - \sum_{k=1}^{n}\ell_k(x_0)\sigma_{ij}(x_k) > \epsilon\right).$$

For every $t > 0$, we have by Markov's inequality

$$\begin{aligned}\mathbf{P}\left(\sum_{k=1}^{n}n\ell_k(x_0)A_k > n\epsilon\right) &= \mathbf{P}\left(\exp\left(t\sum_{k=1}^{n}\frac{2}{h}K\left(\frac{x_i-x_0}{h}\right)A_k\right) > e^{nt\epsilon}\right)\\ &\le \frac{\mathbf{E}\exp(t\sum_{k=1}^{n}\frac{2}{h}K(\frac{x_i-x_0}{h})A_k)}{e^{nt\epsilon}}.\end{aligned} \tag{20}$$

Before we continue, for a given t, let us first define the following quantities, where i, j are omitted from $\Phi_1(i, j)$

$$a_k = \frac{2t}{h} K\left(\frac{x_k - x_0}{h}\right)(\sigma_i(x_k)\sigma_j(x_k) + \sigma_{ij}(x_k)),$$

$$b_k = \frac{2t}{h} K\left(\frac{x_k - x_0}{h}\right)(\sigma_i(x_k)\sigma_j(x_k) - \sigma_{ij}(x_k)),$$

$$\Phi_1 = \frac{1}{n}\sum_{k=1}^{n} \frac{a_k - b_k}{2t}, \qquad \Phi_2 = \frac{1}{n}\sum_{k=1}^{n} \frac{a_k^2 + b_k^2}{4t^2},$$

$$\Phi_3 = \frac{1}{n}\sum_{k=1}^{n} \frac{a_k^3 - b_k^3}{6t^3}, \qquad \Phi_4 = \frac{1}{n}\sum_{k=1}^{n} \frac{a_k^4 + b_k^4}{8t^4},$$

$$M = \max_{k=1,\ldots,n}\left(\frac{2}{h} K\left(\frac{x_k - x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right).$$

We now establish some convenient comparisons; see Sects. 5.2 and 5.3 for their proofs.

Proposition 4 $\frac{\Phi_3}{\Phi_2} \le \frac{4M}{3}$ *and* $\frac{\Phi_4}{\Phi_2} \le 2M^2$, *where both equalities are established at* $\rho_{ij}(x_k) = 1, \forall k$.

Lemma 7 *For* $b_k \le a_k \le \frac{1}{2}, \forall k$, $\frac{1}{2}\sum_{k=1}^{n} \ln \frac{1}{(1-a_k)(1+b_k)} \le nt\Phi_1 + nt^2\Phi_2 + nt^3\Phi_3 + \frac{9}{5}nt^4\Phi_4$.

To show the following, we first replace the sum with a Riemann integral, and then use a Taylor expansion to approximate $\sigma_i(x_k), \sigma_j(x_k)$, and $\sigma_{ij}(x_k), \forall k = 1, \ldots, n$ with σ_i, σ_j σ_{ij} and their first derivatives at x_0 respectively, plus some remainder terms; see Sect. 5.4 for details.

Proposition 5 *Let* $K(\cdot)$ *satisfy* (19) *and* $1 > h > n^{-(2/5)}$. *Then there exists some constant* $C_1 > 0$ *such that*

$$\Phi_2(i, j) = \frac{C_1(\sigma_i^2(x_0)\sigma_j^2(x_0) + \sigma_{ij}^2(x_0))}{h}.$$

Lemma 8 computes the moment generating function for $\frac{2}{h}K(\frac{x_k - x_0}{h})Z_{ki} \cdot Z_{kj}$. The proof proceeds exactly as in the proof of Lemma 9 after substituting t with $\frac{2t}{h}K(\frac{x_k - x_0}{h})$ everywhere.

Lemma 8 *Let* $\frac{2t}{h}K(\frac{x_k - x_0}{h})(1 + \rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k) < 1, \forall k$. *For* $b_k \le a_k < 1$.

$$\mathbf{E}e^{\frac{2t}{h}K(\frac{x_k - x_0}{h}t)Z_{ki}Z_{kj}} = ((1 - a_k)(1 + b_k))^{-1/2}.$$

Remark 5 When we set $t = \frac{\epsilon}{4\Phi_2}$, the bound on ϵ implies that $b_k \le a_k \le 1/2$:

$$a_k = t(1 + \rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k) \le 2t\sigma_i(x_k)\sigma_j(x_k) = \frac{\epsilon\sigma_i(x_k)\sigma_j(x_k)}{2\Phi_2} \le \frac{1}{2}.$$

We can now finish showing the large deviation bound for $\max_{i,j}|\hat{S}_{i,j}-\mathbf{E}S_{i,j}|$. Given that $A_1,\ldots,A_n$ are independent, we have

$$\begin{aligned}\mathbf{E}e^{t\sum_{k=1}^{n}\frac{2}{h}K(\frac{x_k-x_0}{h})A_k} &= \prod_{k=1}^{n}\mathbf{E}e^{\frac{2t}{h}K(\frac{x_1-x_0}{h})A_k}\\ &= \prod_{k=1}^{n}\exp\left(-\frac{2t}{h}K\left(\frac{x_k-x_0}{h}\right)\sigma_{ij}(x_k)\right)\prod_{k=1}^{n}\mathbf{E}e^{\frac{2t}{h}K(\frac{x_k-x_0}{h})Z_{ki}Z_{kj}}. \quad (21)\end{aligned}$$

By (20), (21), and Lemma 8, for $t\le\frac{\epsilon}{4\Phi_2}$,

$$\begin{aligned}\mathbf{P}\left(\sum_{k=1}^{n}\frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)A_k>n\epsilon\right) &\le \frac{\mathbf{E}e^{t\sum_{k=1}^{n}\frac{2}{h}K(\frac{x_k-x_0}{h})A_k}}{e^{-nt\epsilon}}\\ &= e^{-nt\epsilon}\prod_{k=1}^{n}e^{-\frac{2t}{h}K(\frac{x_k-x_0}{h})\sigma_{ij}(x_k)}\mathbf{E}e^{\frac{2t}{h}K(\frac{x_k-x_0}{h})Z_{ki}Z_{kj}}\\ &= \exp\left(-nt\epsilon-nt\Phi_1(i,j)+\frac{1}{2}\sum_{k=1}^{n}\ln\frac{1}{(1-a_k)(1+b_k)}\right)\\ &\le \exp\left(-nt\epsilon+nt^2\Phi_2+nt^3\Phi_3+\frac{9}{5}nt^4\Phi_4\right),\end{aligned}$$

where the last step is due to Remark 5 and Lemma 7. Now let us consider taking t that minimizes $\exp(-nt\epsilon+nt^2\Phi_2+nt^3\Phi_3+\frac{9}{5}nt^4\Phi_4)$. Let $t=\frac{\epsilon}{4\Phi_2}$, then

$$\frac{d}{dt}\left(-nt\epsilon+nt^2\Phi_2+nt^3\Phi_3+\frac{9}{5}nt^4\Phi_4\right)\le-\frac{\epsilon}{40}.$$

Given that, $\frac{\epsilon^2}{\Phi_2}<\frac{1}{M}$, we have

$$\begin{aligned}\mathbf{P}\left(\sum_{k=1}^{n}\frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)A_k>n\epsilon\right) &\le \exp\left(-nt\epsilon+nt^2\Phi_2+nt^3\Phi_3+\frac{9}{5}nt^4\Phi_4\right)\\ &\le \exp\left(\frac{-n\epsilon^2}{4\Phi_2}+\frac{n\epsilon^2}{16\Phi_2}+\frac{n\epsilon^2}{64\Phi_2}\frac{\epsilon\Phi_3}{\Phi_2^2}+\frac{9}{5}\frac{n\epsilon^2}{256\Phi_2}\frac{\epsilon^2\Phi_4}{\Phi_2^3}\right)\\ &\le \exp\left(\frac{-3n\epsilon^2}{20\Phi_2}\right)\\ &\le \exp\left(-\frac{3nh\epsilon^2}{20C_1(\sigma_i^2(x_0)\sigma_j^2(x_0)+\sigma_{ij}^2(x_0))}\right).\end{aligned}$$

Finally, we verify the requirement on $\epsilon\le\frac{\Phi_2}{M}$,

$$\epsilon\le\frac{(C_1(1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0))/h}{\max_{k=1,\ldots,n}(\frac{2}{h}K(\frac{x_k-x_0}{h})\sigma_i(x_k)\sigma_j(x_k))}=\frac{(C_1(1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0))}{\max_{k=1,\ldots,n}(2K(\frac{x_k-x_0}{h})\sigma_i(x_k)\sigma_j(x_k))}. \quad \square$$

For completeness, we compute the moment generating function for $Z_{k,i}Z_{k,j}$.

Lemma 9 *Let* $t(1+\rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k) < 1, \forall k$, *so that* $b_k \le a_k < 1$, *omitting* x_k *everywhere*,

$$\mathbf{E}e^{tZ_{k,i}Z_{k,j}} = \left(\frac{1}{(1-t(\sigma_i\sigma_j+\sigma_{ij}))(1+t(\sigma_i\sigma_j-\sigma_{ij}))}\right)^{1/2}.$$

Proof W.l.o.g., let $i=1$ and $j=2$.

$$\begin{aligned}
\mathbf{E}(e^{tZ_1Z_2}) &= \mathbf{E}(\mathbf{E}(e^{tZ_2Z_1}|Z_2)) \\
&= \mathbf{E}\exp\left(\left(\frac{t\rho_{12}\sigma_1}{\sigma_2} + \frac{t^2\sigma_1^2(1-\rho_{12}^2)}{2}\right)Z_2^2\right) \\
&= \left(1-2\left(\frac{t\rho_{12}\sigma_1}{\sigma_2} + \frac{t^2\sigma_1^2(1-\rho_{12}^2)}{2}\right)\sigma_2^2\right)^{-1/2} \\
&= \left(\frac{1}{1-(2t\rho_{12}\sigma_1\sigma_2 + t^2\sigma_1^2\sigma_2^2(1-\rho_{12}^2))}\right)^{1/2} \\
&= \left(\frac{1}{(1-t(1+\rho_{12})\sigma_1\sigma_2)(1+t(1-\rho_{12})\sigma_1\sigma_2)}\right)^{1/2}
\end{aligned}$$

where $2t\rho_{12}\sigma_1\sigma_2 + t^2\sigma_1^2\sigma_2^2(1-\rho_{12}^2) < 1$. This requires that $t < \frac{1}{(1+\rho_{12})\sigma_1\sigma_2}$ which is equivalent to $2t\rho_{12}\sigma_1\sigma_2 + t^2\sigma_1^2\sigma_2^2(1-\rho_{12}^2) - 1 < 0$. One can check that if we require $t(1+\rho_{12})\sigma_1\sigma_2 \le 1$, which implies that $t\sigma_1\sigma_2 \le 1 - t\rho_{12}\sigma_1\sigma_2$ and $t^2\sigma_1^2\sigma_2^2 \le (1-t\rho_{12}\sigma_1\sigma_2)^2$; hence the lemma holds. □

5.2 Proof of Proposition 4

We show one inequality; the other one is bounded similarly. $\forall k$, we compare the kth elements $\Phi_{2,k}, \Phi_{4,k}$ that appear in the sum for Φ_2 and Φ_4 respectively:

$$\begin{aligned}
\frac{\Phi_{4,k}}{\Phi_{2,k}} &= \frac{(a_k^4+b_k^4)4t^2}{(a_k^2+b_k^2)4t^4} \\
&= \left(\frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)^2 \frac{2((1+\rho_{ij}(x_k))^4 + (1-\rho_{ij}(x_k))^4)}{8(1+\rho_{ij}^2(x_k))} \\
&\le \max_k\left(\frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)^2 \max_{0\le\rho\le1}\frac{(1+\rho)^4+(1-\rho)^4}{4(1+\rho^2)} = 2M^2. \quad \square
\end{aligned}$$

5.3 Proof of Lemma 7

We first use the Taylor expansions to obtain:

$$\ln(1-a_k) = -a_k - \frac{a_k^2}{2} - \frac{a_k^3}{3} - \frac{a_k^4}{4} - \sum_{l=5}^{\infty}\frac{(a_k)^l}{l},$$

where,

$$\sum_{l=5}^{\infty}\frac{(a_k)^l}{l} \le \frac{1}{5}\sum_{l=5}^{\infty}(a_k)^5 = \frac{a_k^5}{5(1-a_k)} \le \frac{2a_k^5}{5} \le \frac{a_k^4}{5}$$

for $a_k < 1/2$; Similarly,

$$\ln(1+b_k) = \sum_{n=1}^{\infty} \frac{(-1)^{l-1}(b_k)^l}{l},$$

where

$$\sum_{l=4}^{\infty} \frac{(-1)^l(b_k)^l}{l} > 0 \quad \text{and} \quad \sum_{l=5}^{\infty} \frac{(-1)^l(b_k)^l}{l} < 0.$$

Hence for $b_k \le a_k \le \frac{1}{2}, \forall k$,

$$\begin{aligned}\frac{1}{2}\sum_{k=1}^{n} \ln \frac{1}{(1-a_k)(1+b_k)} &\le \sum_{k=1}^{n} \frac{a_k - b_k}{2} + \frac{a_k^2 + b_k^2}{4} + \frac{a_k^3 - b_k^3}{6} + \frac{9}{5}\frac{a_k^4 + b_k^4}{8} \\ &= nt\Phi_1 + nt^2\Phi_2 + nt^3\Phi_3 + \frac{9}{5}nt^4\Phi_4. \quad \square\end{aligned}$$

5.4 Proof of Proposition 5

We replace the sum with the Riemann integral, and then use Taylor's formula to replace $\sigma_i(x_k)$, $\sigma_j(x_k)$, and $\sigma_{ij}(x_k)$,

$$\begin{aligned}\Phi_2(i,j) &= \frac{1}{n}\sum_{k=1}^{n} \frac{2}{h^2} K^2\left(\frac{x_k - x_0}{h}\right)\left(\sigma_i^2(x_k)\sigma_j^2(x_k) + \sigma_{ij}^2(x_k)\right) \\ &= \int_{x_n}^{x_0} \frac{2}{h^2} K^2\left(\frac{u - x_0}{h}\right)\left(\sigma_i^2(u)\sigma_j^2(u) + \sigma_{ij}^2(u)\right)du + O\left(\frac{1}{h^6 n^2}\right)\end{aligned}$$

where $O(\frac{1}{h^6 n^2})$ is $o(1/h)$ given that $h > 1/n^{2/5}$ and

$$\begin{aligned}&\int_{x_n}^{x_0} \frac{2}{h^2} K^2\left(\frac{u - x_0}{h}\right)\left(\sigma_i^2(u)\sigma_j^2(u) + \sigma_{ij}^2(u)\right)du \\ &\frac{2}{h}\int_{-\frac{1}{h}}^{0} K^2(v)\left(\sigma_i^2(x_0 + hv)\sigma_j^2(x_0 + hv) + \sigma_{ij}^2(x_0 + hv)\right)dv \\ &= \frac{2}{h}\int_{-1}^{0} K^2(v)\left(\sigma_i(x_0) + hv\sigma_i'(x_0) + \frac{\sigma_i''(y_1)(hv)^2}{2}\right)^2 \\ &\quad \times \left\{\left(\sigma_j(x_0) + hv\sigma_j'(x_0) + \frac{\sigma_j''(y_2)(hv)^2}{2}\right)^2\right. \\ &\quad \left. + \left(\sigma_{ij}(x_0) + hv\sigma_{ij}'(x_0) + \frac{\sigma_{ij}''(y_3)(hv)^2}{2}\right)^2\right\}dv \\ &= \frac{2}{h}\int_{-1}^{0} K^2(v)\left((1 + \rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_k^2(x_0)\right)dv + C_2\int_{-1}^{0} vK^2(v)dv + O(h) \\ &= \frac{C_1(1 + \rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0)}{h},\end{aligned}$$

where $y_0, y_1, y_2 \leq hv + x_0$ and C_1, C_2 are some constants chosen so that all equalities hold. Thus the proposition holds. □

6 Smoothness of Σ_t

In this section we prove Lemma 2, which is a corollary of Lemma 12 and Theorem 4. Hence we show that if we assume $\Theta(x) = (\theta_{ij}(x))$ are smooth and twice differentiable functions of $x \in [0, 1]$, i.e., $\theta'_{ij}(x) < \infty$ and $\theta''_{ij}(x) < \infty$ for $x \in [0, 1], \forall i, j$, and satisfy Assumption 3, then the smoothness conditions of Lemma 1 are satisfied.

The following is a standard result in matrix analysis.

Lemma 10 *Let $\Theta(t) \in R^{p\times p}$ has entries that are differentiable functions of $t \in [0, 1]$. Assuming that $\Theta(t)$ is always non-singular, then*

$$\frac{d}{dt}[\Sigma(t)] = -\Sigma(t)\frac{d}{dt}[\Theta(t)]\Sigma(t).$$

Lemma 11 *Suppose $\Theta(t) \in R^{p\times p}$ has entries that each are twice differentiable functions of t. Assuming that $\Theta(t)$ is always non-singular, then*

$$\frac{d^2}{dt^2}[\Sigma(t)] = \Sigma(t)D(t)\Sigma(t),$$

where

$$D(t) = 2\frac{d}{dt}[\Theta(t)]\Sigma(t)\frac{d}{dt}[\Theta(t)] - \frac{d^2}{dt^2}[\Theta(t)].$$

Proof The existence of the second order derivatives for entries of $\Sigma(t)$ is due to the fact that $\Sigma(t)$ and $\frac{d}{dt}[\Theta(t)]$ are both differentiable $\forall t \in [0, 1]$; indeed by Lemma 10,

$$\begin{aligned}
\frac{d^2}{dt^2}[\Sigma(t)] &= \frac{d}{dt}\left[-\Sigma(t)\frac{d}{dt}[\Theta(t)]\Sigma(t)\right] \\
&= -\frac{d}{dt}[\Sigma(t)]\frac{d}{dt}[\Theta(t)]\Sigma(t) - \Sigma(t)\frac{d}{dt}\left[\frac{d}{dt}[\Theta(t)]\Sigma(t)\right] \\
&= -\frac{d}{dt}[\Sigma(t)]\frac{d}{dt}[\Theta(t)]\Sigma(t) - \Sigma(t)\frac{d^2}{dt^2}[\Theta(t)]\Sigma(t) - \Sigma(t)\frac{d}{dt}[\Theta(t)]\frac{d}{dt}[\Sigma(t)] \\
&= \Sigma(t)\left(2\frac{d}{dt}[\Theta(t)]\Sigma(t)\frac{d}{dt}[\Theta(t)] - \frac{d^2}{dt^2}[\Theta(t)]\right)\Sigma(t),
\end{aligned}$$

hence the lemma holds by the definition of $D(t)$. □

Let $\Sigma(x) = (\sigma_{ij}(x)), \forall x \in [0, 1]$. Let $\Sigma(x) = (\Sigma_1(x), \Sigma_2(x), \ldots, \Sigma_p(x))$, where $\Sigma_i(x) \in R^p$ denotes a column vector. By Lemma 11,

$$\begin{aligned}
\sigma'_{ij}(x) &= -\Sigma_i^T(x)\Theta'(x)\Sigma_j(x), \\
\sigma''_{ij}(x) &= \Sigma_i^T(x)D(x)\Sigma_j(x),
\end{aligned} \tag{22}$$

where $\Theta'(x) = (\theta'_{ij}(x)), \forall x \in [0, 1]$.

Lemma 12 *Given Assumptions 2 and 3,* $\forall x \in [0,1]$,

$$|\sigma'_{ij}(x)| \le S_0^2 \sqrt{S_1} < \infty.$$

Proof $|\sigma'_{ij}(x)| = |\Sigma_i^T(x)\Theta'(x)\Sigma_j(x)| \le \max_{i=1,\dots,p} |\sigma_i^2(x)| \sum_{k=1}^p \sum_{\ell=1}^p |\theta'_{k\ell}(x)| \le S_0^2\sqrt{S_1}$. □

We denote the elements of $\Theta(x)$ by $\theta_{jk}(x)$. Let θ'_ℓ represent a column vector of Θ'.

Theorem 4 *Given Assumptions 2 and 3,* $\forall i, j, \forall x \in [0,1]$,

$$\sup_{x\in[0,1]} |\sigma''_{ij}(x)| < 2S_0^3 S_1 + S_0^2 S_2 < \infty.$$

Proof By (22) and the triangle inequality,

$$|\sigma''_{ij}(x)| = |\Sigma_i^T(x) D(x) \Sigma_j(x)| \le \max_{i=1,\dots,p} |\sigma_i^2(x)| \sum_{k=1}^p \sum_{\ell=1}^p |D_{k\ell}(x)|$$

$$\le S_0^2 \sum_{k=1}^p \sum_{\ell=1}^p 2|\theta_k'^T(x)\Sigma(x)\theta'_\ell(x)| + |\theta''_{k\ell}(x)| = 2S_0^3 S_1 + S_0^2 S_2,$$

where by Assumption 3, $\sum_{k=1}^p \sum_{\ell=1}^p |\theta''_{k\ell}(x)| \le S_2$, and

$$\sum_{k=1}^p \sum_{\ell=1}^p |\theta_k'^T(x)\Sigma(x)\theta'_\ell(x)| = \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p |\theta'_{ki}(x)\theta'_{\ell j}(x)\sigma_{ij}(x)|$$

$$\le \max_{i=1,\dots,p} |\sigma_i(x)| \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p |\theta'_{ki}(x)\theta'_{\ell j}(x)| \le S_0 S_1.$$ □

7 Some implications of a very sparse Θ

We use $\mathcal{L}^1$ to denote Lebesgue measure on $\mathbf{R}$. The aim of this section is to prove some bounds that correspond to Assumption 3, but only for $\mathcal{L}^1$ a.e. $x \in [0,1]$, based on a single sparsity assumption on Θ as in Assumption 5. We let $E \subset [0,1]$ represent the "bad" set with $\mathcal{L}^1(E) = 0$. and $\mathcal{L}^1$ a.e. $x \in [0,1]$ refer to points in the set $[0,1]\setminus E$ such that $\mathcal{L}^1([0,1]\setminus E) = 1$. When $\|\Theta(x)\|_0 \le s + p$ for all $x \in [0,1]$, we immediately obtain Theorem 5, whose proof appears in Sect. 7.1. We like to point out that although we apply Theorem 5 to Θ and deduce smoothness of Σ, we could apply it the other way around. In particular, it might be interesting to apply it to the correlation coefficient matrix (ρ_{ij}), where the diagonal entries remain invariant. We use $\Theta'(x)$ and $\Theta''(x)$ to denote $(\theta'_{ij}(x))$ and $(\theta''_{ij}(x))$ respectively $\forall x$.

Assumption 5 *Assume that* $\|\Theta(x)\|_0 \le s + p$ $\forall x \in [0,1]$, *where s is the number of off-diagonal non-zero entries in* $\Theta(x)$.

Assumption 6 $\exists S_4, S_5 < \infty$ *such that*

$$S_4 = \max_{ij} \|\theta'_{ij}\|_\infty^2 \quad \textit{and} \quad S_5 = \max_{ij} \|\theta''_{ij}\|_\infty.$$

We state a theorem, the proof of which is in Sect. 7.1 and a corollary.

Theorem 5 *Under Assumption 5, we have* $\|\Theta''(x)\|_0 \le \|\Theta'(x)\|_0 \le \|\Theta(x)\|_0 \le s + p$ *for* $\mathcal{L}^1$ *a.e.* $x \in [0,1]$.

Corollary 1 *Given Assumptions 2 and 5, for* $\mathcal{L}^1$ *a.e.* $x \in [0,1]$

$$|\sigma'_{ij}(x)| \le S_0^2 \sqrt{S_4}(s+p) < \infty.$$

Proof By the proof of Lemma 12,

$$|\sigma'_{ij}(x)| \le \max_{i=1\ldots,p} \|\sigma_i^2\|_\infty \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta'_{k\ell}(x)|.$$

Hence by Theorem 5, for $\mathcal{L}^1$ a.e. $x \in [0,1]$,

$$\begin{aligned}|\sigma'_{ij}(x)| &\le \max_{i=1\ldots,p} \|\sigma_i^2\|_\infty \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta'_{k\ell}(x)| \\ &\le S_0^2 \max_{k,\ell} \|\theta'_{k\ell}\|_\infty \|\Theta'(x)\|_0 \le S_0^2 \sqrt{S_4}(s+p).\end{aligned}$$

□

Lemma 13 *Under Assumptions 5 and 6, for* $\mathcal{L}^1$ *a.e.* $x \in [0,1]$,

$$\sum_{k=1}^{p}\sum_{\ell=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p} |\theta'_{ki}(x)\theta'_{\ell j}(x)| \le (s+p)^2 \max_{ij} \|\theta'_{ij}\|_\infty^2 \sum_{k=1}^{p}\sum_{\ell=1}^{p} \theta''_{k\ell} \le (s+p) \max_{ij} \|\theta''_{ij}\|_\infty$$

hence

$$\operatorname*{ess\,sup}_{x\in[0,1]} \sigma''_{ij}(x) \le 2S_0^3(s+p)^2 S_4 + S_0^2(s+p)S_5.$$

Proof By the triangle inequality, for $\mathcal{L}^1$ a.e. $x \in [0,1]$,

$$\begin{aligned}|\sigma''_{ij}(x)| = |\Sigma_i^T D \Sigma_j| &= \left| \sum_{k=1}^{p}\sum_{\ell=1}^{p} \sigma_{ik}(x)\sigma_{j\ell}(x) D_{k\ell}(x) \right| \\ &\le \max_{i=1\ldots,p} \|\sigma_i^2\|_\infty \sum_{k=1}^{p}\sum_{\ell=1}^{p} |D_{k\ell}(x)| \\ &\le 2S_0^2 \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta_k'^T \Sigma \theta'_\ell| + S_0^2 \sum_{k=1}^{p}\sum_{\ell=1}^{p} |\theta''_{k\ell}| \\ &= 2S_0^3(s+p)^2 S_4 + S_0^2(s+p)S_5,\end{aligned}$$

where for $\mathcal{L}^1$ a.e. $x \in [0,1]$,

$$\sum_{k=1}^{p}\sum_{\ell=1}^{p}|\theta_k'^T\Sigma\theta_\ell'| \le \sum_{k=1}^{p}\sum_{\ell=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p}|\theta_{ki}'\theta_{\ell j}'\sigma_{ij}|$$

$$\le \max_{i=1\ldots,p}\|\sigma_i\|_\infty \sum_{k=1}^{p}\sum_{\ell=1}^{p}\sum_{i=1}^{p}\sum_{j=1}^{p}|\theta_{ki}'\theta_{\ell j}'| \le S_0(s+p)^2S_4$$

and $\sum_{k=1}^{p}\sum_{\ell=1}^{p}|\theta_{k\ell}''| \le (s+p)S_5$. The first inequality is due to the following observation: at most $(s+p)^2$ elements in the sum of $\sum_k\sum_i\sum_\ell\sum_j|\theta_{ki}'(x)\theta_{\ell j}'(x)|$ for $\mathcal{L}^1$ a.e. $x\in[0,1]$, that is, except for E, are non-zero, due to the fact that for $x\in[0,1]\setminus N$, $\|\Theta'(x)\|_0 \le \|\Theta(x)\|_0 \le s+p$ as in Theorem 5. The second inequality is obtained similarly using the fact that for $\mathcal{L}^1$ a.e. $x\in[0,1]$, $\|\Theta''(x)\|_0 \le \|\Theta(x)\|_0 \le s+p$. □

Remark 6 For the bad set $E\subset[0,1]$ with $\mathcal{L}^1(E)=0$, $\sigma_{ij}'(x)$ is well defined as shown in Lemma 10, but it can only be loosely bounded by $O(p^2)$, as $\|\Theta'(x)\|_0 = O(p^2)$, instead of $s+p$, for $x\in E$; similarly, $\sigma_{ij}''(x)$ can only be loosely bounded by $O(p^4)$.

By Lemma 13, using the Lebesgue integral, we can derive the following corollary.

Corollary 2 *Under Assumptions 2, 5, and 6,*

$$\int_0^1(\sigma_{ij}''(x))^2dx \le 2S_0^3S_4s+p^2+S_0^2S_5(s+p)<\infty.$$

7.1 Proof of Theorem 5

Let $\|\Theta(x)\|_0 \le s+p$ for all $x\in[0,1]$.

Lemma 14 *Let a function $u:[0,1]\to\mathbf{R}$. Suppose u has a derivative on F (finite or not) with $\mathcal{L}^1(u(F))=0$. Then $u'(x)=0$ for $\mathcal{L}^1$ a.e. $x\in F$.*

Take $F=\{x\in[0,1]:\theta_{ij}(x)=0\}$ and $u=\theta_{ij}$. For $\mathcal{L}^1$ a.e. $x\in F$, that is, except for a set N_{ij} of $\mathcal{L}^1(N_{ij})=0$, $\theta_{ij}'(x)=0$. Let $N=\bigcup_{ij}N_{ij}$. By Lemma 14,

Lemma 15 *If $x\in[0,1]\setminus N$, where $\mathcal{L}^1(N)=0$, if $\theta_{ij}(x)=0$, then $\theta_{ij}'(x)=0$ for all i,j.*

Let $v_{ij}=\theta_{ij}'$. Take $F=\{x\in[0,1]:v_{ij}(x)=0\}$. For $\mathcal{L}^1$ a.e. $x\in F$, that is, except for a set N_{ij}^1 with $\mathcal{L}(N_{ij}^1)=0$, $v_{ij}'(x)=0$. Let $N_1=\bigcup_{ij}N_{ij}^1$. By Lemma 14,

Lemma 16 *If $x\in[0,1]\setminus N_1$, where $\mathcal{L}^1(N_1)=0$, if $\theta_{ij}'(x)=0$, then $\theta_{ij}''(x)=0,\forall i,j$.*

Thus this allows to conclude that

Lemma 17 *If $x\in[0,1]\setminus N\cup N_1$, where $\mathcal{L}^1(N\cup N_1)=0$, if $\theta_{ij}(x)=0$, then $\theta_{ij}'(x)=0$ and $\theta_{ij}''(x)=0,\forall i,j$.*

Thus for all $x\in[0,1]\setminus N\cup N_1$, $\|\Theta''(x)\|_0 \le \|\Theta'(x)\|_0 \le \|\Theta(x)\|_0 \le (s+p)$. □

8 Examples

In this section, we demonstrate the effectiveness of the method in a simulation. Starting at time $t = t_0$, the original graph is as shown at the top of Fig. 1. The graph evolves according to a type of Erdős-Rényi random graph model. Initially we set $\Theta = 0.25 I_{p\times p}$, where $p = 50$. Then, we randomly select 50 edges and update Θ as follows: for each new edge (i, j), a weight $a > 0$ is chosen uniformly at random from $[0.1, 0.3]$; we subtract a from θ_{ij} and θ_{ji}, and increase θ_{ii}, θ_{jj} by a. This keeps Σ positive definite.

When we later delete an existing edge from the graph, we reverse the above procedure with its weight. Weights are assigned to the initial 50 edges, and then we change the graph structure periodically as follows: Every 200 discrete time steps, five existing edges are deleted, and five new edges are added. However, for each of the five new edges, a target weight is chosen, and the weight on the edge is gradually changed over the ensuing 200 time steps in order ensure smoothness. Similarly, for each of the five edges to be deleted, the weight gradually decays to zero over the ensuing 200 time steps. Thus, almost always, there are 55 edges in the graph and 10 edges have weights that are varying smoothly.

8.1 Regularization paths

We increase the sample size from $n = 200$, to 400, 600, and 800 and use a Gaussian kernel with bandwidth $h = \frac{5.848}{n^{1/3}}$. We use the following metrics to evaluate model selection consistency (sparsistency) and predictive risk (3) in Fig. 1 as the ℓ_1 regularization parameter ρ increases. Let $\hat{F}_n$ denote edges in estimated $\hat{\Theta}_n(t_0)$ and F denote edges in $\Theta(t_0)$. Let us define

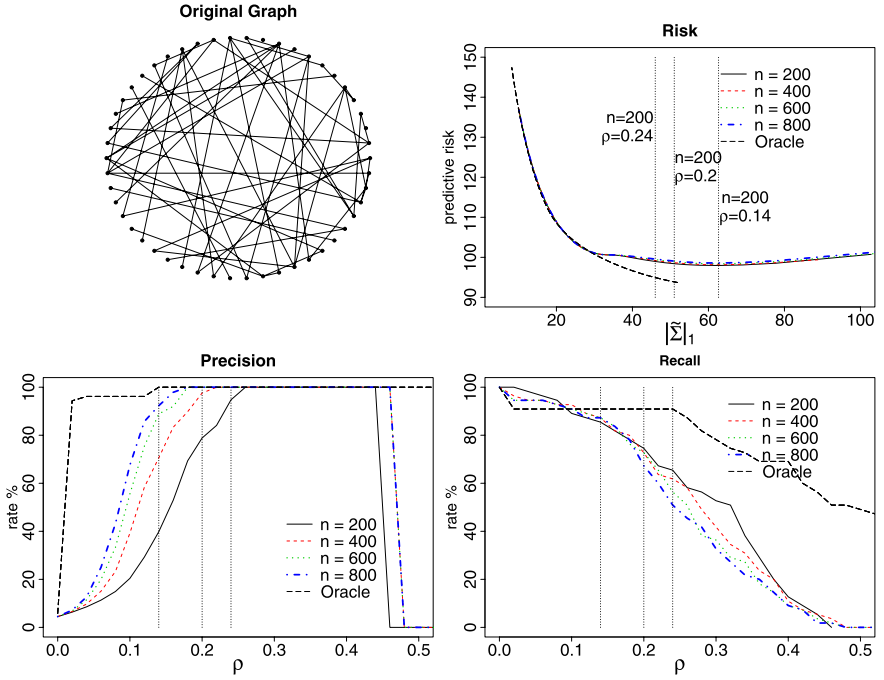

Fig. 1 As the penalization parameter ρ increases, precision goes up, and then down as no edges are predicted in the end. Recall goes down as the estimated graphs are missing more and more edges. The oracle Σ^* performs the best, given the same value for $|\hat{\Sigma}_n(t_0)|_1 = |\Sigma^*|_1, \forall n$

$$\text{precision} = 1 - \frac{\hat{F}_n \setminus F}{\hat{F}_n} = \frac{\hat{F}_n \cap F}{\hat{F}_n},$$
$$\text{recall} = 1 - \frac{F \setminus \hat{F}_n}{F} = \frac{\hat{F}_n \cap F}{F}.$$

Figure 1 shows how they change with ρ.

The predictive risks in (3) are plotted for both the oracle estimator (5) and empirical estimators (6) for each n. They are indexed with the ℓ_1 norm of various estimators vectorized; hence $|\cdot|_1$ for $\hat{\Sigma}_n(t_0)$ and $\Sigma^*(t_0)$ are the same along a vertical line. Note that $|\Sigma^*(t_0)|_1 \leq |\Sigma(t_0)|_1, \forall \rho \geq 0$; for every estimator $\tilde{\Sigma}$ (the oracle or empirical), $|\tilde{\Sigma}|_1$ decreases as ρ increases, as shown in Fig. 1 for $|\hat{\Sigma}_{200}(t_0)|_1$.

Figure 2 shows a subsequence of estimated graphs as ρ increases for sample size $n = 200$. The original graph at t_0 is shown in Fig. 1.

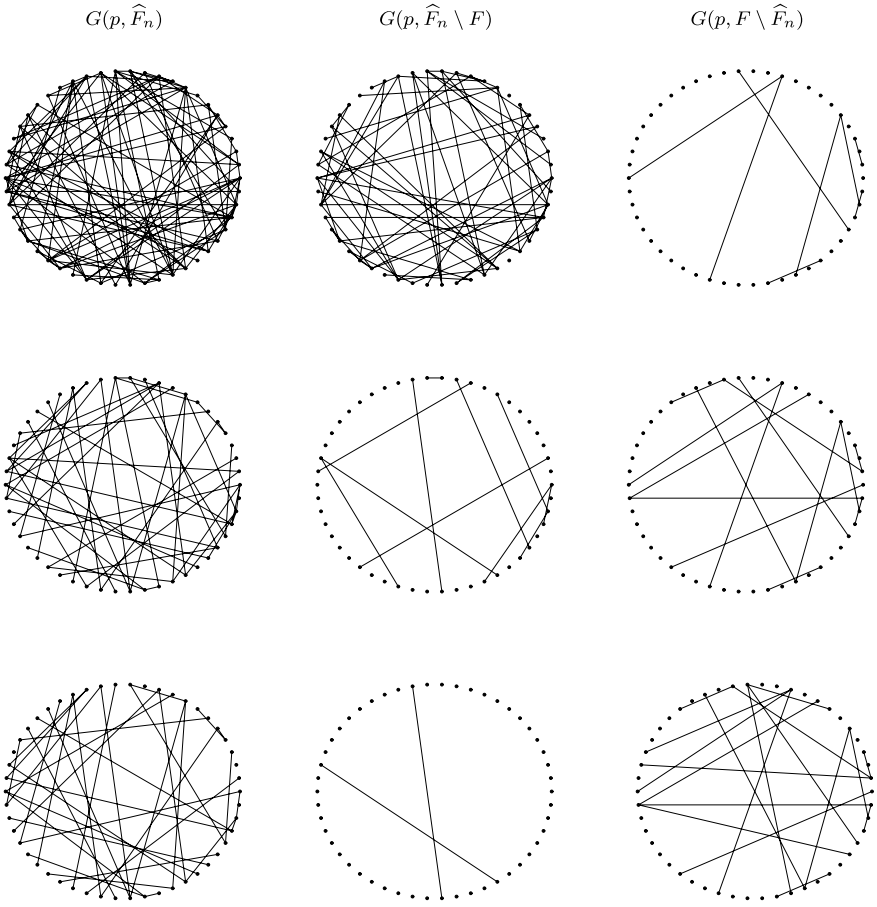

Fig. 2 $n = 200$ and $h = 1$ with $\rho = 0.14, 0.2, 0.24$ indexing each row. The three columns show sets of edges in $\hat{F}_n$, extra edges, and missing edges with respect to the true graph $G(p, F)$. This array of plots show that ℓ_1 regularization is effective in selecting the subset of edges in the true model $\Theta(t_0)$, even when the samples before t_0 were from graphs that evolved over time

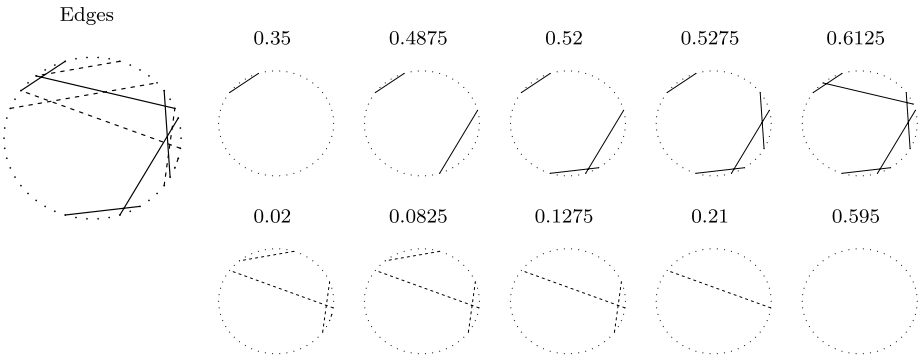

Fig. 3 There are 400 discrete steps in [0, 1] such that the edge set $F(t)$ remains unchanged before or after $t = 0.5$. This sequence of plots shows the times at which each of the new edges added at $t = 0$ appears in the estimated graph (*top row*), and the times at which each of the old edges being replaced is removed from the estimated graph (*bottom row*), where the weight decreases from a positive value in [0.1, 0.3] to zero during the time interval [0, 0.5]. *Solid* and *dashed lines* denote new and old edges respectively

8.2 Chasing the changes

Finally, we show how quickly the smoothed estimator using GLASSO (Friedman et al. 2008) can include the edges that are being added in the beginning of the interval [0, 1], and get rid of edges being replaced, whose weights start to decrease at $x = 0$ and become 0 at $x = 0.5$ in Fig. 3.

9 Conclusions and extensions

We have shown that if the covariance changes smoothly over time, then minimizing an ℓ_1-penalized kernel risk function leads to good estimates of the covariance matrix. This, in turn, allows estimation of time varying graphical structure. The method is easy to apply and is feasible in high dimensions.

Acknowledgements We thank Alan Frieze and Giovanni Leoni for helpful discussions, and J. Friedman, T. Hastie and R. Tibshirani for making the GLASSO code publicly available. We thank the two anonymous reviewers for their careful reading of this paper and for their helpful comments and for pointing out the connection of our results to sparsistency results in Ravikumar et al. (2008). This research was supported in part by NSF grant CCF-0625879.

Appendix: Large deviation inequalities for boxcar kernel function

In this section, we prove the following lemma, which implies the i.i.d. case as in the corollary.

Lemma 18 *Let $Z_k \sim N(0, \Sigma(k)), k = 1, \ldots, n$, be independently but not identically distributed. Let $\hat{S}_n(t)$ be defined as in (2) with $w_{st} = 1/n, \forall s, \forall t$. For ϵ small enough, for some $c_2 > 0$, we have*

$$\mathbf{P}(|\hat{S}_n(t, i, j) - \mathbf{E}\hat{S}_n(t, i, j)| > \epsilon) \leq \exp\{-c_2 n\epsilon^2\}.$$

Corollary 3 *For the i.i.d. case, for some* $c_3 > 0$,

$$\mathbf{P}(|\hat{S}_n(i,j) - \mathbf{E}\hat{S}_n(i,j)| > \epsilon) \leq \exp\{-c_3 n\epsilon^2\}.$$

Lemma 18 is implied by Lemma 19 for diagonal entries, and Lemma 20 for non-diagonal entries.

A.1 Inequalities for squared sum of independent normals with changing variances

Throughout this section, we use σ_i^2 as a shorthand for σ_{ii} as before. Hence $\sigma_i^2(x_k) = \mathrm{Var}(Z_{k,i}) = \sigma_{ii}(x_k), \forall k = 1, \ldots, n$. Ignoring the bias term as in (18), we wish to show that each of the diagonal entries of $\hat{\Sigma}_{ii}$ is close to $\sigma_i^2(x_0), \forall i = 1, \ldots, p$. The following lemma might be of its independent interest; hence we include it here. We omit the proof due to its similarity to that of Lemma 6.

Lemma 19 *We let* $z_1, \ldots, z_n$ *represent a sequence of independent Gaussian random variables such that* $z_k \sim N(0, \sigma^2(x_k))$. *Let* $\sigma^2 = \frac{1}{n}\sum_{k=1}^n \sigma^2(x_k)$. *Then* $\forall \epsilon < c\sigma^2$, *for some* $c \geq 2$, *we have*

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{k=1}^n z_k^2 - \sigma^2\right| > \epsilon\right) \leq \exp\left\{\frac{-(3c-5)n\epsilon^2}{3c^2\sigma^2\sigma_{\max}^2}\right\},$$

where $\sigma_{\max}^2 = \max_{k=1,\ldots,n}\{\sigma^2(x_k)\}$.

A.2 Inequalities for independent sum of products of correlated normals

The proof of Lemma 20 follows that of Lemma 6.

Lemma 20 *Let* $\Psi_2 = \frac{1}{n}\sum_{k=1}^n \frac{(\sigma_i^2(x_k)\sigma_j^2(x_k) + \sigma_{ij}^2(x_k))}{2}$ *and* $c_4 = \frac{3}{20\Psi_2}$. *Let* $\hat{S}_n(t)$ *be defined as in* (2) *with* $w_{st} = 1/n, \forall s, \forall t$. *Then for* $\epsilon \leq \frac{\Psi_2}{\max_k(\sigma_i(x_k)\sigma_j(x_k))}$, *we have*

$$\mathbf{P}\big(|\hat{S}_n(t,i,j) - \mathbf{E}\hat{S}_n(t,i,j)| > \epsilon\big) \leq \exp\{-c_4 n\epsilon^2\}.$$

References

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, *9*, 485–516.

Bickel, P., & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, *36*(1), 199–227.

Drton, M., & Perlman, M. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, *91*(3), 591–602.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, *9*(3), 432–441. doi:10.1093/biostatistics/kxm045.

Greenshtein, E., & Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, *10*, 971–988.

Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, *37*(6B), 4254–4278.

Meinshausen, N., & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*(3), 1436–1462.

Ravikumar, P., Wainwright, M., Raskutti, G., & Yu, B. (2008) *High-dimensional covariance estimation by minimizing* ℓ_1*-penalized log-determinant divergence* (Tech. Rep. 767). UC Berkeley, Department of Statistics.

Rothman, A., Bickel, P., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, *2*, 494–515.