Time Varying Undirected Graphs

Shuheng Zhou, John Lafferty and Larry Wasserman*
Carnegie Mellon University
{szhou, lafferty}@cs.cmu.edu, larry@stat.cmu.edu

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 show how to estimate the sequence of graphs for non-identically distributed data, where the distribution evolves over time.

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 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 p is small, the graph may be estimated from D by testing which partial correlations are not significantly different from zero [DP04]. 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 [MB06, BGd08, FHT07, LF07, BL08, RBLZ07].

All these 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. 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 ℓ_1 regularization method. We 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. We make the following theoretical contributions: (i) nonparametric predictive risk consistency and rate of convergence of the covariance matrices, (ii) consistency and rate of convergence in Frobenius norm of the inverse covariance matrix, (iii) large deviation results for covariance matrices for non-identically distributed observations, and (iv) conditions that guarantee smoothness of the covariances. In addition, we provide simulation evidence that we can recover graphical structure. We believe these are the first such results on time varying undirected graphs.

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 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 $\Sigma(t)^{-1}$. This method can be used to investigate a simple time series model of the form: $W^0 \sim N(0, \Sigma(0))$, and

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

Ultimately, we are interested in the general time series model where the Z^t's 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 iid case, recent work [BGd08, FHT07] has considered ℓ_1-penalized maximum likelihood estimators

*This research was supported in part by NSF grant CCF-0625879. SZ thanks Alan Frieze and Giovanni Leoni for helpful discussions on sparsity and smoothness of functions. We thank J. Friedman, T. Hastie and R. Tibshirani for making GLASSO publicly available, and anonymous reviewers for their constructive comments.

over the entire set of positive definite matrices,

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

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

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

$$\text{where} \quad \widehat{S}_n(t) = \frac{\sum_s w_{st} Z_s Z_s^T}{\sum_s w_{st}} \quad (2)$$

is a weighted covariance matrix, with weights $w_{st} = K\left(\frac{|s-t|}{h_n}\right)$ given by a symmetric nonnegative function kernel over time; in other words, $\widehat{S}_n(t)$ is just the 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 iid 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, $\text{tr}(W)$ the trace of W. Let $\varphi_{\max}(W)$ and $\varphi_{\min}(W)$ be the largest and smallest eigenvalues, respectively. We write $W^{\searrow} = \text{diag}(W)$ for a diagonal matrix with the same diagonal as W, and $W^{\diamond} = W - W^{\searrow}$. 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 $|W|_1 = \|\text{vec}W\|_1 = \sum_i \sum_j |w_{ij}|$, 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. Consider estimates $\widehat{\Sigma}_n(t)$ and $\widehat{G}_n(t) = (V, \widehat{F}_n)$. The first risk function is

$$U(G(t), \widehat{G}_n(t)) = \mathbf{E}L(G(t), \widehat{G}_n(t)) \quad (3)$$

where $L(G(t), \widehat{G}_n(t)) = \left|F(t) \,\Delta\, \widehat{F}_n(t)\right|$, that is, the size of the symmetric difference between two edge sets. We say that $\widehat{G}_n(t)$ is *sparsistent* if $U(G(t), \widehat{G}_n(t)) \xrightarrow{P} 0$ as $n \to \infty$.

The second risk is defined as follows. Let $Z \sim N(0, \Sigma_0)$ and let Σ be a positive definite matrix. Let

$$R(\Sigma) = \text{tr}(\Sigma^{-1}\Sigma_0) + \log|\Sigma|. \quad (4)$$

Note that, up to an additive constant,

$$R(\Sigma) = -2E_0(\log f_\Sigma(Z)),$$

where f_Σ is the density for $N(0, \Sigma)$. We say that $\widehat{G}_n(t)$ is *persistent* [GR04] with respect to a class of positive definite matrices $\mathcal{S}_n$ if $R(\widehat{\Sigma}_n) - \min_{\Sigma \in \mathcal{S}_n} R(\Sigma) \xrightarrow{P} 0$. In the iid case, ℓ_1 regularization yields a persistent estimator, as we now show.

The maximum likelihood estimate minimizes

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

where $\widehat{S}_n$ is the sample covariance matrix. Minimizing $\widehat{R}_n(\Sigma)$ without constraints gives $\widehat{\Sigma}_n = \widehat{S}_n$. We would like to minimize $\widehat{R}_n(\Sigma)$ subject to $\|\Sigma^{-1}\|_0 \leq L$. This would give the "best" sparse graph G, but it is not a convex optimization problem. Hence we estimate $\widehat{\Sigma}_n$ by solving a convex relaxation problem as written in (1) instead. Algorithms for carrying out this optimization are given by [BGd08, FHT07]. Given $L_n, \forall n$, let

$$\mathcal{S}_n = \{\Sigma : \Sigma \succ 0, \left|\Sigma^{-1}\right|_1 \leq L_n\}. \quad (5)$$

We define the oracle estimator and write (1) as (7)

$$\Sigma^*(n) = \arg\min_{\Sigma \in \mathcal{S}_n} R(\Sigma), \quad (6)$$

$$\widehat{\Sigma}_n = \arg\min_{\Sigma \in \mathcal{S}_n} \widehat{R}_n(\Sigma). \quad (7)$$

Note that one can choose to only penalize off-diagonal elements of Σ^{-1} as in [RBLZ07], if desired. We have the following result, whose proof appears in Section 3.2.

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

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

for (5). Then for the sequence of empirical estimators as defined in (7) and $\Sigma^*(n), \forall n$ as in (6),

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

3.1 Risk Consistency for the Non-identical Case

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

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

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

$$\widehat{\Sigma}_n(t) = \arg\min_{\Sigma \in \mathcal{S}_n} \left\{ \text{tr}(\Sigma^{-1}\widehat{S}_n(t)) + \log|\Sigma| \right\} \quad (8)$$

where $\widehat{S}_n(t)$ is given in (2), with $K(\cdot)$ a symmetric nonnegative function with compact support:

A1 *The kernel function K has a bounded support $[-1, 1]$.*

Lemma 2 *Let $\Sigma(t) = [\sigma_{jk}(t)]$. Suppose the following conditions hold:*

1. *There exists $C_0 > 0, C$ 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 $\max_{j,k} |\widehat{S}_n(t,j,k) - \Sigma(t,j,k)| = O_P\left(\frac{\sqrt{\log n}}{n^{1/3}}\right)$ for all $t > 0$.

Proof: By the triangle inequality,

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

In Lemma 14 we show that

$$\max_{j,k}\sup_t |\mathbf{E}\widehat{S}_n(t,j,k) - \Sigma(t,j,k)| = O(C_0 h_n).$$

In Lemma 15, we show that

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

for some $c_1 > 0$. Hence,

$$\mathbf{P}\left(\max_{j,k}|\widehat{S}_n(t,j,k) - \mathbf{E}\widehat{S}_n(t,j,k)| > \epsilon\right) \le \exp\left\{-nh_n(C\epsilon^2 - 2\xi\log n/(nh_n))\right\} \quad \text{and} \tag{9}$$

$\max_{j,k}|\widehat{S}_n(t,j,k) - \mathbf{E}\widehat{S}_n(t,j,k)| = O_P\left(\sqrt{\frac{\log n}{nh_n}}\right)$. Hence the result holds for $h_n \asymp n^{-1/3}$. $\square$

With the use of Lemma 2, the proof of the following follows the same lines as that of Theorem 1.

Theorem 3 *Suppose all conditions in Lemma 2 and the following hold:*

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

Then, $\forall t > 0$, for the sequence of estimators as in (8),

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

Remark 4 *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* (3.1)*.*

Remark 5 *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 $\|\Theta\|_0 = \Omega(p)$. Hence the family of graphs that we can guarantee persistency for, although sparse, is likely to include connected graphs, for example, when $\Omega(p)$ edges were formed randomly among p nodes.*

The smoothness condition in Lemma 2 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 a shorthand for $\sigma_{ii}(x)$.

Definition 6 *For a function $u : [0,1] \to \mathbf{R}$, let $\|u\|_\infty = \sup_{x\in[0,1]}|u(x)|$.*

A2 *There exists some constant $S_0 < \infty$ such that*

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

$$\max_{i=1\ldots,p}\|\sigma_i\|_\infty \le S_0. \tag{12}$$

A3 *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*

$$\sup_{t\in[0,1]}\sum_{k=1}^p\sum_{\ell=1}^p\sum_{i=1}^p\sum_{j=1}^p|\theta'_{ki}(t)\theta'_{\ell j}(t)| \le S_1 \tag{13}$$

$$\sup_{t\in[0,1]}\sum_{k=1}^p\sum_{\ell=1}^p|\theta''_{k\ell}(t)| \le S_2, \tag{14}$$

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

The proof is in Section 6. In Section 7, we show some preliminary results on achieving upper bounds on quantities that appear in Condition 1 of Lemma 2 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 $\forall n$, $\sup_{\Sigma\in\mathcal{S}_n}|R(\Sigma) - \widehat{R}_n(\Sigma)| \le$

$$\sum_{j,k}|\Sigma^{-1}_{jk}|\;|\widehat{S}_n(j,k) - \Sigma_0(j,k)| \le \delta_n\;\left|\Sigma^{-1}\right|_1,$$

where it follows from [RBLZ07] that

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

Hence, minimizing over $\mathcal{S}_n$ with $L_n = o\left(\frac{n}{\log p_n}\right)^{1/2}$, $\sup_{\Sigma\in\mathcal{S}_n}|R(\Sigma) - \widehat{R}_n(\Sigma)| = o_P(1)$. By the definitions of $\Sigma^*(n) \in \mathcal{S}_n$ and $\widehat{\Sigma}_n \in \mathcal{S}_n$, we immediately have $R(\Sigma^*(n)) \le R(\widehat{\Sigma}_n)$ and $\widehat{R}_n(\widehat{\Sigma}_n) \le \widehat{R}_n(\Sigma^*(n))$; thus

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

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

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

Thus $\forall\epsilon > 0$, the event $\left\{\left|R(\widehat{\Sigma}_n) - R(\Sigma^*(n))\right| > \epsilon\right\}$ is contained in the event $\left\{\sup_{\Sigma\in\mathcal{S}_n}|R(\Sigma) - \widehat{R}_n(\Sigma)| > \epsilon/2\right\}$. Thus, for $L_n = o((n/\log n)^{1/2})$, and $\forall\epsilon > 0$, as $n \to \infty$,

$$\mathbf{P}\left(\left|R(\widehat{\Sigma}_n) - R(\Sigma^*(n))\right| > \epsilon\right) \le$$

$$\mathbf{P}\left(\sup_{\Sigma\in\mathcal{S}_n}|R(\Sigma) - \widehat{R}_n(\Sigma)| > \epsilon/2\right) \to 0. \quad \square$$

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 Section 6. We first write our time-varying estimator $\widehat{\Theta}_n(t)$ 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,

$$\widehat{\Theta}_n(t) = \arg\min_{\Theta \succ 0} \left\{ \mathrm{tr}(\Theta \widehat{S}_n(t)) - \log|\Theta| + \lambda_n |\Theta|_1 \right\} \quad (15)$$

where λ_n is a non-negative regularization parameter, and $\widehat{S}_n(t)$ 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.

A4 *Let $p+s = o\left(n^{(2/3)}/\log n\right)$ and $\varphi_{\min}(\Sigma_0) \geq \underline{k} > 0$, hence $\varphi_{\max}(\Theta_0) \leq 1/\underline{k}$. For some sufficiently large constant M, let $\varphi_{\min}(\Theta_0) = \Omega\left(2M\sqrt{\frac{(p+s)\log n}{n^{2/3}}}\right)$.*

The proof draws upon techniques from [RBLZ07], with modifications necessary to handle the fact that we penalize $|\Theta|_1$ rather than $|\Theta^{\diamondsuit}|_1$ as in their case.

Theorem 8 *Let $\widehat{\Theta}_n(t)$ be the minimizer defined by (15). Suppose all conditions in Lemma 2 and A 4 hold. If*

$$\lambda_n \asymp \sqrt{\frac{\log n}{n^{2/3}}}, \quad \textit{then}$$

$$\|\widehat{\Theta}_n(t) - \Theta_0\|_F = O_P\left(2M\sqrt{\frac{(p+s)\log n}{n^{2/3}}}\right). \quad (16)$$

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

$$\begin{aligned} Q(\Theta) &= \mathrm{tr}(\Theta\widehat{S}_n(t_0)) - \log|\Theta| + \lambda|\Theta| - \\ &\quad \mathrm{tr}(\Theta_0\widehat{S}_n(t_0)) + \log|\Theta_0| - \lambda|\Theta_0|_1 \\ &= \mathrm{tr}\left((\Theta - \Theta_0)(\widehat{S}_n(t) - \Sigma_0)\right) - \\ &\quad (\log|\Theta| - \log|\Theta_0|) + \mathrm{tr}\left((\Theta - \Theta_0)\Sigma_0\right) \\ &+ \lambda(|\Theta|_1 - |\Theta_0|_1). \end{aligned} \quad (17)$$

$\widehat{\Theta}$ minimizes $Q(\Theta)$, or equivalently $\widehat{\Delta}_n = \widehat{\Theta} - \Theta_0$ minimizes $G(\Delta) \equiv Q(\Theta_0 + \Delta)$. Hence $G(\underline{0}) = 0$ and $G(\widehat{\Theta}_n) \leq G(\underline{0}) = 0$ by definition. Define for some constant C_1, $\delta_n = C_1\sqrt{\frac{\log n}{n^{2/3}}}$. Now, let

$$\lambda_n = \frac{C_1}{\varepsilon}\sqrt{\frac{\log n}{n^{2/3}}} = \frac{\delta_n}{\varepsilon} \quad \text{for some} \quad 0 < \varepsilon < 1. \quad (18)$$

Consider now the set

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

where

$$r_n = \sqrt{\frac{(p+s)\log n}{n^{2/3}}} \asymp \delta_n\sqrt{p+s} \to 0. \quad (19)$$

Claim 9 *Under A 4, for all $\Delta \in \mathcal{T}_n$ such that $\|\Delta\|_F = o(1)$ as in (19), $\Theta_0 + v\Delta \succ 0, \forall v \in I \supset [0,1]$.*

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 10 *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

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

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 $\log|\Theta_0 + \Delta| - \log|\Theta_0| = \frac{d}{dv}\log|\Theta_0 + v\Delta||_{v=0}\Delta + \int_0^1 (1-v)\frac{d^2}{dv^2}\log\det(\Theta_0 + v\Delta)dv = \mathrm{tr}(\Sigma_0\Delta) + A$, where by symmetry, $\mathrm{tr}(\Sigma_0\Delta) = \mathrm{tr}(\Theta - \Theta_0)\Sigma_0$. Hence

$$\begin{aligned} G(\Delta) &= \\ &A + \mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right) + \lambda_n\left(|\Theta_0 + \Delta|_1 - |\Theta_0|_1\right). \end{aligned} \quad (20)$$

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 $\Theta = \Theta^{\searrow} + \Theta_S^{\diamondsuit} + \Theta_{S^c}^{\diamondsuit}, \forall\Theta$ in our notation. Note that we have $\Theta_{0S^c}^{\diamondsuit} = \underline{0}$,

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

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

$$\begin{aligned} &|\Theta_0 + \Delta|_1 - |\Theta_0|_1 = \\ &\quad |\Theta_0^{\diamondsuit} + \Delta^{\diamondsuit}|_1 - |\Theta_0^{\diamondsuit}|_1 + |\Theta_0^{\searrow} + \Delta^{\searrow}|_1 - |\Theta_0^{\searrow}|_1 \\ &\geq \left|\Delta_{S^c}^{\diamondsuit}\right|_1 - \left|\Delta_S^{\diamondsuit}\right|_1 - |\Delta^{\searrow}|_1. \end{aligned} \quad (21)$$

Now, from Lemma 2, $\max_{j,k} |\widehat{S}_n(t,j,k) - \sigma(t,j,k)| = O_P\left(\frac{\sqrt{\log n}}{n^{1/3}}\right) = O_P(\delta_n)$. By (9), with probability $1 - \frac{1}{n^2}$

$$\left|\mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right)\right| \leq \delta_n |\Delta|_1, \quad \text{hence by (21)}$$

$$\begin{aligned}
&\mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right) + \lambda_n \left(|\Theta_0 + \Delta|_1 - |\Theta_0|_1\right) \\
&\geq -\delta_n |\Delta^{\setminus}|_1 - \delta_n \left|\Delta^{\Diamond}_{S^c}\right|_1 - \delta_n \left|\Delta^{\Diamond}_{S}\right|_1 \\
&\quad -\lambda_n |\Delta^{\setminus}|_1 + \lambda_n \left|\Delta^{\Diamond}_{S^c}\right|_1 - \lambda_n \left|\Delta^{\Diamond}_{S}\right|_1 \\
&\geq -(\delta_n + \lambda_n)\left(|\Delta^{\setminus}|_1 + \left|\Delta^{\Diamond}_{S}\right|_1\right) + (\lambda_n - \delta_n)\left|\Delta^{\Diamond}_{S^c}\right|_1 \\
&\geq -(\delta_n + \lambda_n)\left(|\Delta^{\setminus}|_1 + \left|\Delta^{\Diamond}_{S}\right|_1\right), \text{ where} \qquad (22)
\end{aligned}$$

$$\begin{aligned}
&(\delta_n + \lambda_n)\left(|\Delta^{\setminus}|_1 + \left|\Delta^{\Diamond}_{S}\right|_1\right) \\
&\leq (\delta_n + \lambda_n)\left(\sqrt{p}\|\Delta^{\setminus}\|_F + \sqrt{s}\|\Delta^{\Diamond}_{S}\|_F\right) \\
&\leq (\delta_n + \lambda_n)\left(\sqrt{p}\|\Delta^{\setminus}\|_F + \sqrt{s}\|\Delta^{\Diamond}\|_F\right) \\
&\leq (\delta_n + \lambda_n)\max\{\sqrt{p}, \sqrt{s}\}\left(\|\Delta^{\setminus}\|_F + \|\Delta^{\Diamond}\|_F\right) \\
&\leq (\delta_n + \lambda_n)\max\{\sqrt{p}, \sqrt{s}\}\sqrt{2}\|\Delta\|_F \\
&\leq \delta_n \frac{1+\varepsilon}{\varepsilon}\sqrt{p+s}\sqrt{2}\|\Delta\|_F. \qquad (23)
\end{aligned}$$

Combining (20), (22), and (23), we have with probability $1 - \frac{1}{n^c}$, for all $\Delta \in \mathcal{T}_n$,

$$\begin{aligned}
G(\Delta) &\geq A - (\delta_n + \lambda_n)\left(|\Delta^{\setminus}|_1 + \left|\Delta^{\Diamond}_{S}\right|_1\right) \\
&\geq \frac{\underline{k}^2}{2+\tau}\|\Delta\|_F^2 - \delta_n \frac{1+\varepsilon}{\varepsilon}\sqrt{p+s}\sqrt{2}\|\Delta\|_F \\
&= \|\Delta\|_F^2\left(\frac{\underline{k}^2}{2+\tau} - \delta_n \frac{\sqrt{2}(1+\varepsilon)}{\varepsilon\|\Delta\|_F}\sqrt{p+s}\right) \\
&= \|\Delta\|_F^2\left(\frac{\underline{k}^2}{2+\tau} - \frac{\delta_n\sqrt{2}(1+\varepsilon)}{\varepsilon M r_n}\sqrt{p+s}\right) > 0
\end{aligned}$$

for M sufficiently large, where the bound on A comes from Lemma 11 by [RBLZ07]. □

Lemma 11 *([RBLZ07]) For some $\tau = o(1)$, under A 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 claim.

Claim 12 *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 (19)$\}$. 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: Now by contradiction, 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 Claim 13. 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 Claim 12 and the fact that $G(\widehat{\Delta}_n) \leq G(0) = 0$, we have the following: If $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $\widehat{\Delta}_n \notin (\mathcal{T}_n \cup \mathcal{V}_n)$, that is, $\|\widehat{\Delta}_n\|_F < M r_n$, given that $\widehat{\Delta}_n = \widehat{\Theta}_n - \Theta_0$, where $\widehat{\Theta}_n, \Theta_0 \succ 0$. Therefore

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

We thus establish that $\|\widehat{\Delta}_n\|_F \leq O_P(M r_n)$. □

Claim 13 *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 $v \in (0, 1)$, where $1 - v > 0$; $\forall x \in \mathbf{R}^{\mathbf{P}}$, by $B \succ 0$ and $B + D \succ 0$, $x^T B x > 0$ and $x^T(B + D)x > 0$; hence $x^T D x > -x^T B x$. Thus $x^T(B + vD)x = x^T B x + v x^T D x > (1-v)x^T B x > 0$. □

5 Large Deviation Inequalities

Before we go on, we explain the notation that we follow throughout this section. We switch notation from t to x and form a regression problem for non-iid 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

$$Z_k = (Z_{k1}, \ldots, Z_{kp})^T \sim N(0, \Sigma_k), \ \forall k. \qquad (24)$$

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

$$P^n(|\widehat{\Sigma}_{ij} - \Sigma_{ij}(x_0)| > \epsilon) < e^{-cn\epsilon^2} \qquad (25)$$

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 $\widehat{\Sigma}_{ij}$ 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 Section 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 A 1. This kernel has the property that:

$$2\int_{-1}^0 vK(v)dv \leq 2\int_{-1}^0 K(v)dv = 1 \qquad (26)$$

$$2\int_{-1}^0 v^2 K(v)dv \leq 1. \qquad (27)$$

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

$$\ell_i(x_0) = \frac{2}{nh} K\left(\frac{x_i - x_0}{h}\right) \approx \frac{K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \quad (28)$$

where the approximation is due to replacing the sum with the 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 $\sigma_{ij}(x_k) = \text{cov}(Z_{ki}, Z_{kj}) = \rho_{ij}(x_k)\sigma_i(x_k)\sigma_j(x_k)$ 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

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

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

We thus decompose and bound for point of interest x_0

$$\begin{aligned}
\left|\sum_{k=1}^n \ell_k(x_0) Z_{ki}Z_{kj} - \sigma_{ij}(x_0)\right| &\leq \\
&\left|\mathbf{E}\sum_{k=1}^n \ell_k(x_0) Z_{ki}Z_{kj} - \sigma_{ij}(x_0)\right| + \\
&\left|\sum_{k=1}^n \ell_k(x_0) Z_{ki}Z_{kj} - \mathbf{E}\sum_{k=1}^n \ell_k(x_0) Z_{ki}Z_{kj}\right| \quad (29)\\
= &\left|\sum_{k=1}^n \ell_k(x_0) Z_{ki}Z_{kj} - \Phi_1(i,j)\right| + |\Phi_1(i,j) - \sigma_{ij}(x_0)|.
\end{aligned}$$

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

Lemma 14 *Suppose there exists $C > 0$ such that* $\max_{i,j}\sup_t |\sigma''(t,i,j)| \leq C$. *Then*

$$\forall t \in [0,1], \quad \max_{i,j} |\mathbf{E}\widehat{S}_n(t,i,j) - \sigma_{ij}(t)| = O(h).$$

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

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

We now use Taylor's Formula to replace $\sigma_{ij}(x_0 + hv)$ and obtain $2\int_{-1/h}^0 K(v)\sigma_{ij}(x_0 + hv)dv =$
$2\int_{-1}^0 K(v)\left(\sigma_{ij}(x_0) + hv\sigma'_{ij}(x_0) + \frac{\sigma''_{ij}(y(v))(hv)^2}{2}\right)dv$
$= \sigma_{ij}(x_0) + 2\int_{-1}^0 K(v)\left(hv\sigma'_{ij}(x_0) + \frac{C(hv)^2}{2}\right)dv,$

$$\begin{aligned}
\text{where} \quad & 2\int_{-1}^0 K(v)\left(hv\sigma'_{ij}(x_0) + \frac{C(hv)^2}{2}\right)dv\\
= \quad & 2h\sigma'_{ij}(x_0)\int_{-1}^0 vK(v)dv + \frac{Ch^2}{2}\int_{-1}^0 v^2 K(v)dv\\
\leq \quad & h\sigma'_{ij}(x_0) + \frac{Ch^2}{4}, \text{ where } y(v) - x_0 < hv.
\end{aligned}$$

Thus $\Phi_1(i,j) - \sigma_{ij}(x_0) = O(h)$. □

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

Lemma 15 *For* $\epsilon < \frac{C_1\left(\sigma_i^2(x_0)\sigma_j^2(x_0) + \sigma_{ij}^2(x_0)\right)}{\max_{k=1,\ldots,n}\left(2K\left(\frac{x_k - x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)}$, *where C_1 is defined in Claim 18, for some $C > 0$,*

$$\mathbf{P}\left(|\widehat{S}_n(t,i,j) - \mathbf{E}\widehat{S}_n(t,i,j)| > \epsilon\right) \leq \exp\left\{-Cnh\epsilon^2\right\}.$$

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

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

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

$$\begin{aligned}
&\mathbf{P}\left(\sum_{k=1}^n n\ell_k(x_0)A_k > n\epsilon\right)\\
&= \mathbf{P}\left(e^{t\sum_{k=1}^n \frac{2}{h}K\left(\frac{x_i - x_0}{h}\right)A_k} > e^{nt\epsilon}\right)\\
&\leq \frac{\mathbf{E}e^{t\sum_{k=1}^n \frac{2}{h}K\left(\frac{x_i - x_0}{h}\right)A_k}}{e^{nt\epsilon}}. \quad (30)
\end{aligned}$$

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

- $a_k = \frac{2t}{h}K\left(\frac{x_k - x_0}{h}\right)(\sigma_i(x_k)\sigma_j(x_k) + \sigma_{ij}(x_k))$
- $b_k = \frac{2t}{h}K\left(\frac{x_k - x_0}{h}\right)(\sigma_i(x_k)\sigma_j(x_k) - \sigma_{ij}(x_k))$ thus
- $\Phi_1 = \frac{1}{n}\sum_{k=1}^n \frac{a_k - b_k}{2t}$, $\Phi_2 = \frac{1}{n}\sum_{k=1}^n \frac{a_k^2 + b_k^2}{4t^2}$
- $\Phi_3 = \frac{1}{n}\sum_{k=1}^n \frac{a_k^3 - b_k^3}{6t^3}$, $\Phi_4 = \frac{1}{n}\sum_{k=1}^n \frac{a_k^4 + b_k^4}{8t^4}$
- $M = \max_{k=1,\ldots,n}\left(\frac{2}{h}K\left(\frac{x_k - x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)$

We now establish some convenient comparisons; see Section B.1 and B.2 for their proofs.

Claim 16 $\frac{\Phi_3}{\Phi_2} \leq \frac{4M}{3}$ *and* $\frac{\Phi_4}{\Phi_2} \leq 2M^2$, *where both equalities are established at $\rho_{ij}(x_k) = 1, \forall k$.*

Lemma 17 *For $b_k \leq a_k \leq \frac{1}{2}, \forall k$, $\frac{1}{2}\sum_{k=1}^n \ln \frac{1}{(1-a_k)(1+b_k)}$ $\leq nt\Phi_1 + nt^2\Phi_2 + nt^3\Phi_3 + \frac{9}{5}nt^4\Phi_4$.*

To show the following, we first replace the sum with a Riemann integral, and then use Taylor's Formula 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 Section B.3 for details.

Claim 18 *For $h = n^{-\epsilon}$ for some $1 > \epsilon > 0$, there exists some constant $C_1 > 0$ such that*

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

Lemma19 computes the moment generating function for $\frac{2}{h}K\left(\frac{x_k - x_0}{h}\right)Z_{ki}\cdot Z_{kj}$. The proof proceeds exactly as that of Lemma 21 after substituting t with $\frac{2t}{h}K\left(\frac{x_k-x_0}{h}\right)$ everywhere.

Lemma 19 *Let $\frac{2t}{h}K\left(\frac{x_k-x_0}{h}\right)(1+\rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k)$ $< 1, \forall k$. For $b_k \leq a_k < 1$.*

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

Remark 20 *Thus when we set $t = \frac{\epsilon}{4\Phi_2}$, the bound on ϵ implies that $b_k \leq a_k \leq 1/2, \forall k$:*

$$\begin{aligned} a_k &= t(1+\rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k) \\ &\leq 2t\sigma_i(x_k)\sigma_j(x_k) = \frac{\epsilon\sigma_i(x_k)\sigma_j(x_k)}{2\Phi_2} \leq \frac{1}{2}. \end{aligned}$$

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

$$\begin{aligned} \mathbf{E}e^{t\sum_{k=1}^n \frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)A_k} &= \prod_{k=1}^n \mathbf{E}e^{\frac{2t}{h}K\left(\frac{x_1-x_0}{h}\right)A_k} \\ &= \prod_{k=1}^n \exp\left(-\frac{2t}{h}K\left(\frac{x_k-x_0}{h}\right)\sigma_{ij}(x_k)\right)\cdot \\ &\quad \prod_{k=1}^n \mathbf{E}e^{\frac{2t}{h}K\left(\frac{x_k-x_0}{h}\right)Z_{ki}Z_{kj}} \qquad (31) \end{aligned}$$

By (30), (31), Lemma 19, for $t \leq \frac{\epsilon}{4\Phi_2}$,

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

where the last step is due to Remark 20 and Lemma 17. Now let us consider taking t that minimizes $\exp\left(-nt\epsilon + nt^2\Phi_2 + nt^3\Phi_3 + \frac{9}{5}nt^4\Phi_4\right)$; Let $t = \frac{\epsilon}{4\Phi_2}$: $\frac{d}{dt}\left(-nt\epsilon + nt^2\Phi_2 + nt^3\Phi_3 + \frac{9}{5}nt^4\Phi_4\right) \leq -\frac{\epsilon}{40}$; Now given that $\frac{\epsilon^2}{\Phi_2} < \frac{1}{M}$, Claim 16 and 18:

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

Finally, let's check the requirement on $\epsilon \leq \frac{\Phi_2}{M}$,

$$\begin{aligned} \epsilon &\leq \frac{\left(C_1(1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0)\right)/h}{\max_{k=1,\ldots,n}\left(\frac{2}{h}K\left(\frac{x_k-x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)} \\ &= \frac{\left(C_1(1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0)\right)}{\max_{k=1,\ldots,n}\left(2K\left(\frac{x_k-x_0}{h}\right)\sigma_i(x_k)\sigma_j(x_k)\right)}. \end{aligned}$$

□

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

Lemma 21 *Let $t(1+\rho_{ij}(x_k))\sigma_i(x_k)\sigma_j(x_k) < 1, \forall k$, so that $b_k \leq a_k < 1$, 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 $j = 2$.

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

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 $t(1+\rho_{12})\sigma_1\sigma_2 \leq 1$, which implies that $t\sigma_1\sigma_2 \leq 1 - t\rho_{12}\sigma_1\sigma_2$ and hence $t^2\sigma_1^2\sigma_2^2 \leq (1-t\rho_{12}\sigma_1\sigma_2)^2$, the lemma holds. □

6 Smoothness and Sparsity of Σ_t via Σ_t^{-1}

In this section 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 A 3, then the smoothness conditions of Lemma 2 are satisfied. The following is a standard result in matrix analysis.

Lemma 22 *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 23 *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*

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

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

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 22,

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

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 23,

$$\sigma'_{ij}(x) = -\Sigma_i^T(x)\Theta'(x)\Sigma_j(x), \quad (32)$$

$$\sigma''_{ij}(x) = \Sigma_i^T(x)D(x)\Sigma_j(x), \quad (33)$$

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

Lemma 24 *Given A 2 and A 3, $\forall x \in [0, 1]$,*

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

Proof: $|\sigma'_{ij}(x)| = |\Sigma_i^T(x)\Theta'(x)\Sigma_j(x)|$

$$\leq \max_{i=1\ldots,p} |\sigma_i^2(x)| \sum_{k=1}^p \sum_{\ell=1}^p |\theta'_{k\ell}(x)| \leq S_0^2\sqrt{S_1}.$$

□

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

Theorem 25 *Given A 2 and A 3, $\forall i, j, \forall x \in [0, 1]$,*

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

Proof: By (33) and the triangle inequality,

$$\begin{aligned}
&\left|\sigma''_{ij}(x)\right| = \left|\Sigma_i^T(x)D(x)\Sigma_j(x)\right| \\
&\leq \max_{i=1\ldots,p} \left|\sigma_i^2(x)\right| \sum_{k=1}^p \sum_{\ell=1}^p |D_{k\ell}(x)| \\
&\leq S_0^2 \sum_{k=1}^p \sum_{\ell=1}^p 2|\theta_k'^T(x)\Sigma(x)\theta'_\ell(x)| + |\theta''_{k\ell}(x)| \\
&= 2S_0^3S_1 + S_0^2S_2,
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{k=1}^p \sum_{\ell=1}^p \left|\theta_k'^T(x)\Sigma(x)\theta'_\ell(x)\right| \\
&= \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p \left|\theta'_{ki}(x)\theta'_{\ell j}(x)\sigma_{ij}(x)\right| \\
&\leq \max_{i=1\ldots,p} |\sigma_i(x)| \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p \left|\theta'_{ki}(x)\theta'_{\ell j}(x)\right| \\
&\leq S_0S_1. \quad \square
\end{aligned}$$

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 A 3, but only for $\mathcal{L}^1$ a.e. $x \in [0, 1]$, based on a single sparsity assumption on Θ as in A 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 $\|\Theta(x)\|_0 \leq s + p$ for all $x \in [0, 1]$, we immediately obtain Theorem 26, whose proof appears in Section 7.1. We like to point out that although we apply Theorem 26 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 $\forall x$.

A5 *Assume that $\|\Theta(x)\|_0 \leq s + p \; \forall x \in [0, 1]$.*

A6 $\exists S_4, S_5 < \infty$ *such that*

$$S_4 = \max_{ij} \left\|\theta'_{ij}\right\|_\infty^2 \quad \textit{and} \quad S_5 = \max_{ij} \left\|\theta''_{ij}\right\|_\infty. \quad (34)$$

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

Theorem 26 *Under A 5, we have $\|\Theta''(x)\|_0 \leq \|\Theta'(x)\|_0 \leq \|\Theta(x)\|_0 \leq s + p$ for $\mathcal{L}^1$ a.e. $x \in [0, 1]$.*

Corollary 27 *Given A 2 and A 5, for $\mathcal{L}^1$ a.e. $x \in [0,1]$*

$$|\sigma'_{ij}(x)| \le S_0^2 \sqrt{S_4}(s+p) < \infty. \quad (35)$$

Proof: By proof of Lemma 24,
$|\sigma'_{ij}(x)| \le \max_{i=1\dots,p} \|\sigma_i^2\|_\infty \sum_{k=1}^p \sum_{\ell=1}^p |\theta'_{k\ell}(x)|$.
Hence by Theorem 26, for $\mathcal{L}^1$ a.e. $x \in [0,1]$, $|\sigma'_{ij}(x)| \le \max_{i=1\dots,p} \|\sigma_i^2\|_\infty \sum_{k=1}^p \sum_{\ell=1}^p |\theta'_{k\ell}(x)|$
$\le S_0^2 \max_{k,\ell} \|\theta'_{k\ell}\|_\infty \|\Theta'(x)\|_0 \le S_0^2\sqrt{S_4}(s+p)$. □

Lemma 28 *Under A 5 and 6, for $\mathcal{L}^1$ a.e. $x \in [0,1]$,*

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

$$\sum_{k=1}^p \sum_{\ell=1}^p \theta''_{k\ell} \le (s+p) \max_{ij} \left\|\theta''_{ij}\right\|_\infty, \quad \textit{hence}$$

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

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

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

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

$$\begin{aligned}
\sum_{k=1}^p \sum_{\ell=1}^p |\theta_k'^T \Sigma \theta'_\ell| &\le \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p |\theta'_{ki}\theta'_{\ell j}\sigma_{ij}| \\
&\le \max_{i=1\dots,p} \|\sigma_i\|_\infty \sum_{k=1}^p \sum_{\ell=1}^p \sum_{i=1}^p \sum_{j=1}^p |\theta'_{ki}\theta'_{\ell j}| \\
&\le S_0(s+p)^2 S_4
\end{aligned}$$

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 $\sum_k \sum_i \sum_\ell \sum_j \left|\theta'_{ki}(x)\theta'_{\ell j}(x)\right|$ 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 26. 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 29 *For the bad set $E \subset [0,1]$ with $\mathcal{L}^1(E) = 0$, $\sigma'_{ij}(x)$ is well defined as shown in Lemma 22, 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 28, using the Lebesgue integral, we can derive the following corollary.

Corollary 30 *Under A 2, A 5, and A 6,*

$$\int_0^1 \left(\sigma''_{ij}(x)\right)^2 dx \le 2S_0^3 S_4 s + p^2 + S_0^2 S_5(s+p) < \infty.$$

7.1 Proof of Theorem 26.

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

Lemma 31 *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 31,

Lemma 32 *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^1_{ij} with $\mathcal{L}(N^1_{ij}) = 0$, $v'_{ij}(x) = 0$. Let $N_1 = \bigcup_{ij} N^1_{ij}$. By Lemma 31,

Lemma 33 *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 34 *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 Figure 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.

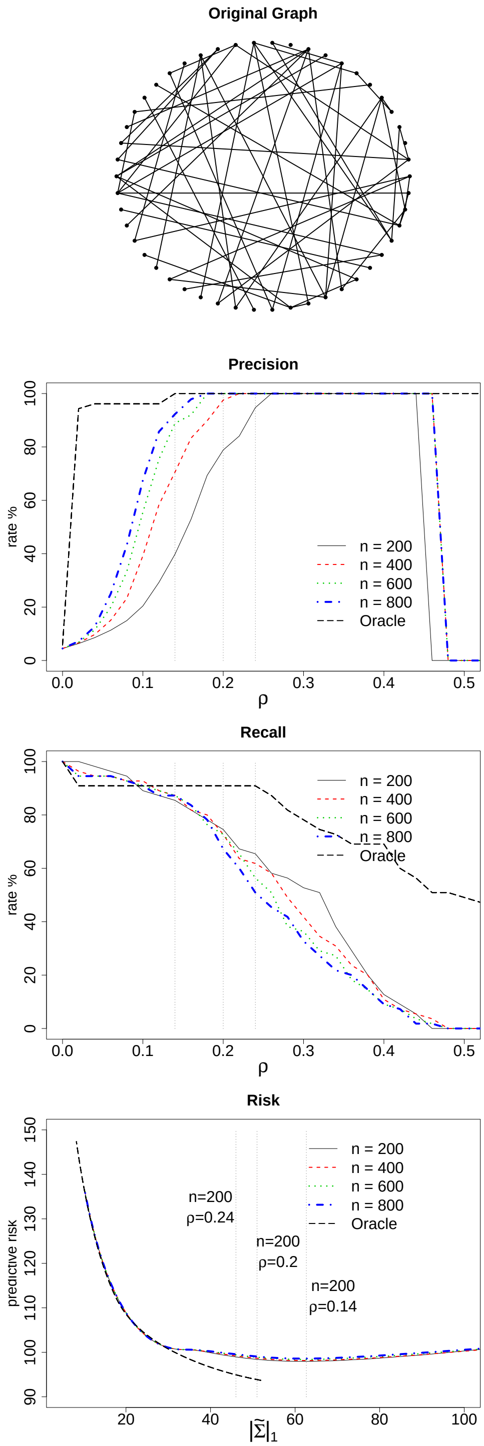

Figure 1: Plots from top to bottom show that as the penalization parameter ρ increases, precision goes up, and then down as no edges are predicted in the end. Recall goes down as the estimated graphs are missing more and more edges. The oracle Σ^* performs the best, given the same value for $|\widehat{\Sigma}_n(t_0)|_1 = |\Sigma^*|_1, \forall n$.

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 consistency risk for (3) and predictive risk (4) in Figure 1 as the ℓ_1 regularization parameter ρ increases.

- Let $\widehat{F}_n$ denote edges in estimated $\widehat{\Theta}_n(t_0)$ and F denote edges in $\Theta(t_0)$. Let us define

$$\text{precision} = 1 - \frac{\widehat{F}_n \setminus F}{\widehat{F}_n} = \frac{\widehat{F}_n \cap F}{\widehat{F}_n},$$

$$\text{recall} = 1 - \frac{F \setminus \widehat{F}_n}{F} = \frac{\widehat{F}_n \cap F}{F}.$$

  Figure 1 shows how they change with ρ.

- Predictive risks in (4) are plotted for both the oracle estimator (6) and empirical estimators (7) for each n. They are indexed with the ℓ_1 norm of various estimators vectorized; hence $|\cdot|_1$ for $\widehat{\Sigma}_n(t_0)$ 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 $\widetilde{\Sigma}$ (the oracle or empirical), $|\widetilde{\Sigma}|_1$ decreases as ρ increases, as shown in Figure 1 for $|\widehat{\Sigma}_{200}(t_0)|_1$.

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 Figure 1.

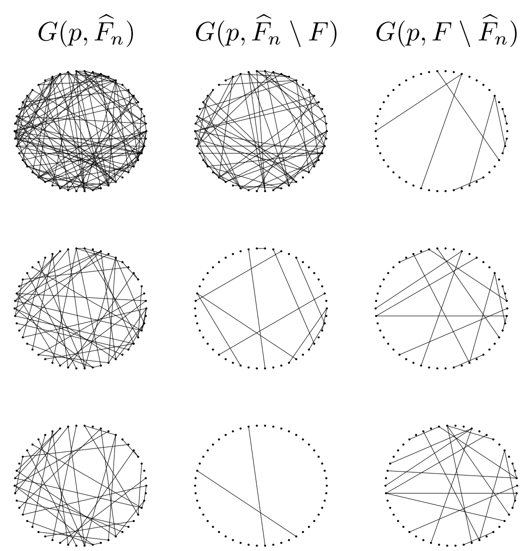

Figure 2: $n = 200$ and $h = 1$ with $\rho = 0.14, 0.2, 0.24$ indexing each row. The three columns show sets of edges in $\widehat{F}_n$, extra edges, and missing edges with respect to the true graph $G(p, F)$. This array of plots show that ℓ_1 regularization is effective in selecting the subset of edges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

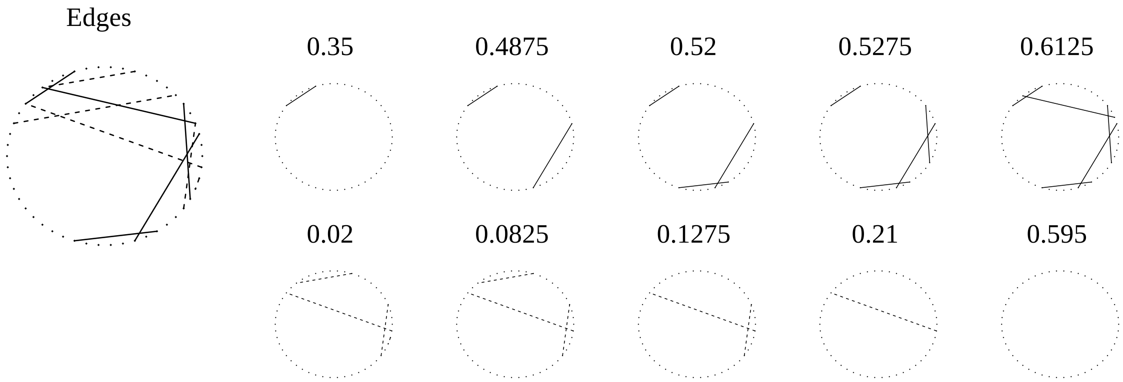

Figure 3: There are 400 discrete steps in $[0, 1]$ such that the edge set $F(t)$ remains unchanged before or after $t = 0.5$. This sequence of plots shows the times at which each of the new edges added at $t = 0$ appears in the estimated graph (top row), and the times at which each of the old edges being 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 [FHT07] can include the edges that are being added in the beginning of 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 Figure 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.

We are currently addressing several extensions to this work. First, with stronger conditions we expect that we can establish *sparsistency*, that is, we recover the edges with probability approaching one. Second, we can relax the smoothness assumption using nonparametric changepoint methods [GH02] which allow for jumps. Third, we used a very simple time series model; extensions to more general time series models are certainly feasible.

References

[BGd08] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9:485–516, March 2008.

[BL08] P.J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 2008. To appear.

[DP04] M. Drton and M.D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

[FHT07] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 2007.

[GH02] G. Grégoire and Z. Hamrouni. Change point estimation by local linear smoothing. *J. Multivariate Anal.*, 83:56–83, 2002.

[GR04] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli*, 10:971–988, 2004.

[LF07] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrices estimation, 2007. arXiv:0711.3933v1.

[MB06] N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[RBLZ07] A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation, 2007. Technical report 467, Dept. of Statistics, Univ. of Michigan.

A 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 35 *Using a boxcar kernel that weighs uniformly over n samples $Z_k \sim N(0, \Sigma(k)), k = 1, \ldots, n$, that are independently but not identically distributed, we have for ϵ small enough, for some $c_2 > 0$,*

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

Corollary 36 *For the i.i.d. case, for some $c_3 > 0$,*

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

Lemma 35 is implied by Lemma 37 for diagonal entries, and Lemma 38 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 (29), we wish to show that each of the diagonal entries of $\widehat{\Sigma}_{ii}$ is close to $\sigma_i^2(x_0), \forall i = 1, \ldots, p$. For a boxcar kernel that weighs uniformly over n samples, we mean strictly $\ell_k(x_0) = \frac{1}{n}, \forall k = 1, \ldots, n$, and $h = 1$ for (28) in this context. We omit the mention of i or t in all symbols from here on. 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 15.

Lemma 37 *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 $\sigma^2 = \frac{1}{n}\sum_{k=1}^n \sigma^2(x_k)$. Using a boxcar kernel that weighs uniformly over n samples, $\forall \epsilon < c\sigma^2$, 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 38 follows that of Lemma 15.

Lemma 38 *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 $c_4 = \frac{3}{20\Psi_2}$. Using a boxcar kernel that weighs uniformly over n samples, for $\epsilon \leq \frac{\Psi_2}{\max_k(\sigma_i(x_k)\sigma_j(x_k))}$,*

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

B Proofs for Large Deviation Inequalities

B.1 Proof of Claim 16

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

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

B.2 Proof of Lemma 17

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,

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

for $a_k < 1/2$; Similarly,

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

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

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

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

B.3 Proof of Claim 18

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 $\sigma_{ij}(x_k)$,

$$\begin{aligned}
\Phi_2(i,j) &= \frac{1}{n}\sum_{k=1}^n \frac{2}{h^2}K^2\left(\frac{x_k - x_0}{h}\right)\left(\sigma_i^2(x_k)\sigma_j^2(x_k) + \sigma_{ij}^2(x_k)\right) \\
&\approx \int_{x_n}^{x_0}\frac{2}{h^2}K^2\left(\frac{u - x_0}{h}\right)\left(\sigma_i^2(u)\sigma_j^2(u) + \sigma_{ij}^2(u)\right)du \\
&= \frac{2}{h}\int_{-\frac{1}{h}}^{0}K^2(v)\left(\sigma_i^2(x_0+hv)\sigma_j^2(x_0+hv) + \sigma_{ij}^2(x_0+hv)\right)dv \\
&= \frac{2}{h}\int_{-1}^{0}K^2(v)\left(\sigma_i(x_0) + hv\sigma_i'(x_0) + \frac{\sigma_i''(y_1)(hv)^2}{2}\right)^2 \\
&\quad \left(\sigma_j(x_0) + hv\sigma_j'(x_0) + \frac{\sigma_j''(y_2)(hv)^2}{2}\right)^2 + \\
&\quad \left(\sigma_{ij}(x_0) + hv\sigma_{ij}'(x_0) + \frac{\sigma_{ij}''(y_3)(hv)^2}{2}\right)^2 dv \\
&= \frac{2}{h}\int_{-1}^{0}K^2(v)\left((1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_k^2(x_0)\right)dv + \\
&\quad C_2\int_{-1}^{0}vK^2(v)dv + O(h) \\
&= \frac{C_1(1+\rho_{ij}^2(x_0))\sigma_i^2(x_0)\sigma_j^2(x_0)}{h}
\end{aligned}$$

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. $\square$