INFEREN CE FOR SPARSE CONDITIONAL PRECISION MATRICES

JIALEI WANG

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science at
Department of Computer Science
University of Chicago
Chicago, Illinois

Thesis Committee:
Mladen Kolar (thesis advisor)
Risi Kondor
John Lafferty (thesis advisor)

May 2015
ABSTRACT

Given $n$ i.i.d. observations of a random vector $(X, Z)$, where $X$ is a high-dimensional vector and $Z$ is a low-dimensional index variable, we study the problem of estimating the conditional inverse covariance matrix $\Omega(z) = \left(\mathbb{E}[(X - \mathbb{E}[X | Z])(X - \mathbb{E}[X | Z])^T | Z = z]\right)^{-1}$ under the assumption that the set of non-zero elements is small and does not depend on the index variable. We develop a novel procedure that combines the ideas of the local constant smoothing and the group Lasso for estimating the conditional inverse covariance matrix. A proximal iterative smoothing algorithm is used to solve the corresponding convex optimization problems. We prove that our procedure recovers the conditional independence assumptions of the distribution $X | Z$ with high probability. This result is established by developing a uniform deviation bound for the high-dimensional conditional covariance matrix from its population counterpart, which may be of independent interest. Furthermore, we develop point-wise confidence intervals for individual elements of the conditional inverse covariance matrix. We perform extensive simulation studies, in which we demonstrate that our proposed procedure outperforms sensible competitors. We illustrate our proposal on a S&P 500 stock price data set.

KEYWORDS

conditional covariance selection, undirected graphical models, point-wise confidence bands, proximal iterative smoothing, group lasso, graphical lasso, varying-coefficient models, high-dimensional, network estimation
ACKNOWLEDGMENTS

First I would like to thank my advisors Mladen Kolar and John Lafferty for their encouragement and guidance during the first two years of my PhD study. I learned a lot from John’s great courses (Nonparametric Inference, Topics in Statistical Machine Learning, High-dimensional Statistics). Mladen was the one who guided me every detail on research: from asking research questions to formalizing problems, from scientific writing to presenting ideas. The work presented in this thesis is a collaboration with him.

I would like to thank Nati Srebro at TTI-C with whom I worked during the last two years on exciting research projects, he taught me a lot on optimization and learning, which greatly influence my thinking on statistics and computation. I deeply appreciate Risi Kondor, who helped me a lot at the beginning of my PhD study, and being serving on my thesis committee.

I also thank my other collaborators Samory Kpotufe, Shubhendu Trived and Greg Shakhnarovich who influenced my research, and Ryohei Fujimaki at NEC Research Lab who mentored me last summer. I sincerely thank my fellow students and friends at the university for their help and suggestions on numerous things. Last but not least, I thank my family for their love, support and encouragement.
CONTENTS

1 INTRODUCTION 1
2 METHODOLOGY 5
   2.1 Preliminaries 5
   2.2 Estimation Procedure 6
   2.3 Proximal Iterative Smoothing 7
3 THEORETICAL RESULTS 13
   3.1 Consistent Graph Identification 13
   3.2 Proof of Sparsistency 18
   3.3 Asymptotic Normality 29
4 EMPIRICAL STUDIES 35
   4.1 Simulation Studies 35
   4.2 Illustrative Applications to Real Data 46
5 CONCLUSIONS 49
   5.1 Conclusions and Future work 49

BIBLIOGRAPHY 51
LIST OF FIGURES

Figure 1  Typical convergence behavior of PRISMA and ADMM when solving the optimization problem in (3). The curve was obtained on a problem with \((n, p) = (500, 50)\).  10
Figure 2  Precision-Recall curves for exponential function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.  38
Figure 3  Precision-Recall curves for linear function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.  39
Figure 4  Precision-Recall curves for random sin function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.  40
Figure 5  Average F1-scores of various methods.  41
Figure 6  Case where the underlying precision matrix is static.  41
Figure 7  Average Frobenius norm errors of various methods.  42
Figure 8  Average hamming distance plotted against the re-scaled sample size.  43
Figure 9  Illustration of the estimated graph using various approaches, on a chain graph with linear precision function.  43
Figure 10 Illustration of the constructed confidence bands by the proposed procedure, on a \(n = 500, p = 50\), chain graph problem.  44
Figure 11 Graphical estimation on S&P 500 stock data.  47
Figure 12 Negative likelihood score versus sparsity.  47
Figure 13 (A): Edge weight as time changes: CVS vs Walgreens. (B)(C): Two examples of non-edges.  48
Figure 14 Top row: edge weight as oil price changes: Exxon Mobil Corp with its major neighbors; bottom row: edge weight as EUR/USD changes: Goldman Sachs Group with its major neighbors. 48

**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Confidence interval, for exponential function precision</td>
<td>44</td>
</tr>
<tr>
<td>Table 2</td>
<td>Confidence interval, for linear function precision</td>
<td>44</td>
</tr>
<tr>
<td>Table 3</td>
<td>Confidence interval, for sin function precision</td>
<td>45</td>
</tr>
</tbody>
</table>
In recent years, we have witnessed advancement of large-scale data acquisition technology in numerous engineering, scientific and socio-economical domains. Data collected in these domains are complex, noisy and high-dimensional. Graphical modeling of complex dependencies and associations underlying the system that generated the data has become an important tool for practitioners interested in gaining insights into complicated data generating mechanisms. The central quantity of interest in graphical modeling is the inverse covariance matrix $\Omega = \Sigma^{-1}$, since the pattern of non-zero elements in $\Omega$ characterize the edge set in the graphical representation of the system. In particular, each measured variable is represented by a node and an edge is put between the nodes if the corresponding element in the inverse covariance matrix is non-zero. Unfortunately, the sample covariance matrix $\hat{\Sigma}$ is a poor estimate of the covariance matrix $\Sigma$ in the high-dimensional regime when the number of variables $p$ is comparable to or much larger than the sample size $n$, as is common in contemporary applications. Furthermore, when the number of variables is larger than the sample size, the sample covariance matrix is not invertible, which further complicates the task of estimating the inverse covariance matrix $\Omega$. Due to its importance, rich literature has emerged on the problem of estimating $\Sigma$ and $\Omega$ in both the statistical and machine learning community [see, for example, 43, for a recent overview].

In many scientific domains it is believed that only a few partial correlations between objects in a system are significant, while others are negligible in comparison. This leads to an assumption that the inverse covariance matrix is sparse, which means that only a few elements are non-zero. There is a vast and rich body of work on the problem of covariance selection, which was introduced in Dempster [13], in recent literature [see, for example, 36, 2, 16, 44]. However, most of the current focus and progress is restricted to simple scenarios where the inverse covariance matrix is studied in isolation, without considering effects of other covariates that represent environmental factors. In this paper, we study the problem of conditional covariance selection in a high-dimensional setting.

Let $(X, Z)$ be a pair of random variables where $X \in \mathbb{R}^p$ and $Z \in \mathbb{R}$. The problem of conditional covariance selection can be simply described as the estimation of the conditional inverse covariance matrix $\Omega(z) = (E[(X - E[X | Z])(X - E[X | Z])^T | Z = z])^{-1}$ with zeros. The random variable $Z$ is called the index variable and the matrix $\Omega(z)$
explains associations between different components of the vector $X$ as function of the index variable $Z$. In this paper, we estimate $\Omega(z)$ using a penalized local Gaussian-likelihood approach, which is suitable in a setting where little is known about the functional relationship between the index variable $Z$ and the associations between components of $X$. As we will show later in the paper, our procedure allows for estimation of $\Omega(z)$ without making strong assumptions on the distribution $(X, Z)$ and only smoothness assumption on $z \mapsto \Omega(z)$. The penalty function used in our approach is based on the $\ell_2$ norm, which encourages the estimated sparsity pattern of $\Omega(z)$ to be the same for all values of the index variable while still allowing for the strength of association to change. Fixing the sparsity pattern of $\Omega(z)$, as a function of $z$, reduces the complexity of the model class and allows us to estimate the conditional inverse covariance matrix more reliably by better utilizing the sample. In addition to developing an estimation procedure that works with minimal assumptions, we also focus on statistical properties of the estimator in the high-dimensional setting, where the number of dimensions $p$ is comparable or even larger than the sample size $n$. The high-dimensional aspect of the problem forces us to carefully analyze statistical properties of the estimator, that would otherwise be apparent in a low-dimensional setting. We prove that our procedure recovers the conditional independence assumptions of the distribution $X \mid Z$ with high probability under suitable technical conditions. The key result that allows us to establish sparsistency of our procedure is a uniform deviation bound for the high-dimensional conditional covariance matrix, which may be of independent interest. Finally, we develop point-wise confidence intervals for individual elements of the conditional inverse covariance matrix.

Our paper contributes to the following literature (i) high-dimensional estimation of inverse covariance matrices, (ii) non-parametric estimation of covariance matrices, (iii) high-dimensional estimation of varying-coefficient models, and (iv) high dimensional inference. In the classical problem of covariance selection \cite{13} the inverse covariance matrix does not depend on the index variable, which should be distinguished from our approach. A number of recently developed methods address the problem of covariance selection in high dimensions. These methods are based on maximizing the penalized likelihood \cite{62, 47, 44, 16, 2, 14, 29, 63, 35, 56, 6} or pseudo-likelihood \cite{36, 42, 46, 60, 5, 49, 23}. Many interesting problems cannot be cast in the context of covariance selection from independent and identically distributed data. As a result, a number of authors have proposed ways to estimate sparse inverse covariance matrices and corresponding graphs in more challenging settings. Guo et al. \cite{17}, Varoquaux et al. \cite{52}, Honorio and Samaras \cite{19}, Chiquet et al. \cite{10}, Danaher et al. \cite{12} and Mohan et al. \cite{37} study estimation of multiple related
Gaussian graphical models in a multi-task setting. While our work is related to this body of literature, there are some fundamental differences. Instead of estimating a small, fixed number of inverse covariance matrices that are related in some way, we are estimating a function \( z \mapsto \Omega(z) \), which is an infinite dimensional object and requires a different treatment. Closely related is work on estimation of high-dimensional time-varying graphical models in Zhou et al. [68], Kolar et al. [28], Kolar and Xing [26], Kolar and Xing [24] and Chen et al. [9], where authors focus on estimating \( \Omega(z) \) at a fixed time point.

In contrast, we estimate the whole mapping \( z \mapsto \Omega(z) \) under the assumption that the sparsity pattern does not change with the index variable. Zhou et al. [68] and Kolar et al. [28] allow the sparsity pattern of \( \Omega(z) \) to change with the index variable, however, when data are scarce, it is sensible to restrict the model class within which we perform estimation. Recent literature on conditional covariance estimation includes Yin et al. [57] and Hoff and Niu [18], however, these authors do not consider ultra-high dimensional setting. Kolar et al. [27] uses a pseudolikelihood approach to learn structure of conditional graphical model. Our work is also different from the literature on sparse conditional Gaussian graphical models [see, for example, 58, 30, 8, 7, 11, 69, 59], where it is assumed that \( X \mid Z = z \sim N(\Gamma z, \Sigma) \).

In particular, the covariates \( Z \) only affect the conditional mean, where in our work the covariance matrix depends on the indexing variable \( Z \).

Our work also contributes to the literature on varying-coefficient models [see, for example, 15, 41, for recent survey articles]. Work on variable selection in high-dimensional varying-coefficient models is focused on generalizing the linear regression where the relationship between the response and predictors is linear after conditioning on the index variable [see, for example, 55, 33, 50]. In this paper, after conditioning on the index variable the model is multivariate Normal and the standard techniques cannot be directly applied. Furthermore, we develop point-wise confidence intervals for individual parameters in the model, which has not been done in the context of high-dimensional varying coefficient models and requires different techniques to those used in Zhang and Lee [65]. Only recently have methods for constructing confidence intervals for individual parameters in high-dimensional models been proposed [see, for example, 64, 22, 51]. We construct our point-wise confidence intervals using a similar approach to one described in Jankova and van de Geer [20] (alternative constructions were given in Liu 34 and Ren et al. 45). The basic idea is to construct a de-sparsified estimator, in which the bias introduced by the sparsity inducing penalty is removed, and prove that this estimator is asymptotically Normal. Additional difficulty in our paper is the existence of the bias term that arises due to local smoothing.
In summary, here are the highlights of our paper. Our main contribution is the development of a new procedure for estimation of the conditional inverse covariance matrix under an assumption that the set of non-zero elements of $\Omega(z)$ does not change with the index variable. Our procedure is based on maximizing the penalized local Gaussian likelihood with the $\ell_1/\ell_2$ mixed norm as the penalty and we develop a scalable iterative proximal smoothing algorithm for solving the objective. On the theoretical front, we establish sufficient conditions under which the graph structure can be recovered consistently and show how to construct point-wise confidence intervals for parameters in the model. Our theoretical results are illustrated in simulation studies and the method is used to study associations between a set of stocks in the S&P 500.
In this section, we propose a method for learning a sparse conditional inverse covariance matrix. Section 2.1 formalizes the problem, the estimator is given in Section 3.1, and the numerical procedure is described in Section 2.3.

2.1 Preliminaries

Let \( \{x^i, z^i\}_{i \in [n]} \) be an independent sample from a joint probability distribution \((X,Z)\) over \(\mathbb{R}^p \times [0,1]\). We assume that the conditional distribution of \(X\) given \(Z = z\) is given as

\[
X \mid Z = z \sim N(\mu(z), \Sigma(z)).
\]  

Let \( f(z) \) be the density function of \(Z\). We assume that the density is well behaved as specified later, however, we do not pose any specific distributional assumptions. Furthermore, we do not specify parametric relationships for the conditional mean or variance as functions of \(Z = z\).

Our goal is to learn the conditional independence assumptions between the components of the vector \(X\) given \(Z = z\) and strength of associations between the components of the vector \(X\) as a function of the index variable \(Z\). Let \( \Omega(z) = \Sigma(z)^{-1} = (\Omega_{uv}(z))_{u,v \in [p] \times [p]} \) be the inverse conditional covariance matrix. The pattern of non-zero elements of this matrix encodes the conditional independencies between the components of the vector \(X\). In particular

\[
X_u \perp X_v \mid X_{-uv}, Z = z \quad \text{if and only if} \quad \Omega_{uv}(z) = 0
\]

where \( X_{-uv} = \{X_c \mid c \in [p] \setminus \{u,v\}\} \). Denote the set of conditional dependencies given \(Z = z\) as

\[
S(z) = \{(u,v) \mid \Omega_{uv}(z) \neq 0\}.
\]

Using \( S(z) \), we define the set

\[
S = \bigcup_{z \in [0,1]} S(z) = \{(u,v) \mid \Omega_{uv}(z) \neq 0 \text{ for some } z \in [0,1]\}.
\]  

Let \( \bar{S} \) be the complement of \( S \), which denotes pairs of components of \(X\) that are conditionally independent irrespective of the value of \(Z\). Under the sparsity assumption, the cardinality of set \( S \) is going to be much smaller than the sample size. Our results will also depend on the maximum node degree in the graph

\[
d = \max_{u \in [p]} |\{v \neq u \mid (u,v) \in S\}|.
\]
In the next section, we describe our procedure for estimating $\Omega(z)$. Based on the estimator of $\Omega(z)$ we will construct an estimate of the set $S$ by reading of the positions of non-zero elements in the estimate of $\Omega(z)$.

### 2.2 Estimation Procedure

We propose to estimate the conditional sparse precision matrix using the following penalized local likelihood objective

$$
\min_{\{\Omega(z^{i})\geq 0\}_{i\in [n]}} \sum_{i\in [n]} \left( \text{tr}(\hat{\Sigma}^{i} \Omega(z^{i})) - \log \det \Omega(z^{i}) \right) + \lambda \sum_{v\neq u} \sqrt{\sum_{i\in [n]} \Omega_{vu}(z^{i})^2} 
$$

(3)

where

$$
\hat{\Sigma}(z) = \sum_{i\in [n]} w_i(z) (x^i - \hat{m}(z^i))(x^i - \hat{m}(z^i))^T
$$

(4)

and

$$
\hat{m}(z) = \sum_{i\in [n]} w_i(z) x^i
$$

(5)

are the locally weighted estimator of the covariance matrix and mean with weights given as

$$
w_i(z) = \frac{K_h(z^i - z)}{\sum_{i\in [n]} K_h(z^i - z)}.
$$

Here $K_h(x) = h^{-1}K(x/h)$, $K(\cdot)$ is a symmetric kernel function with bounded support, $h$ denotes the bandwidth and $\lambda$ is the penalty parameter. The penalty term

$$
\lambda \sum_{v\neq u} \sqrt{\sum_{j\in [n]} (\Omega_{vu}(z^{j}))^2}
$$

is a group sparsity regularizer [61] that encourages the estimated precision matrices to have the same sparsity patterns. The parameter $\lambda > 0$ determines the strength of the penalty. Given the solution $\{\hat{\Omega}(z^{i})\}_{i\in [n]}$ of the optimization program in (3), we let

$$
\hat{S} = \{(u, v) | \hat{\Omega}_{uv}(z^{i}) \neq 0 \text{ for some } i \in [n] \}
$$

(6)

be the estimator of the set $S$ in (2).

Notice that in (3), we only estimate $\Omega(z)$ at $z \in \{z^{i}\}_{i\in [n]}$. Under the assumption that the marginal density of $Z$ is continuous and bounded away from zero on $[0, 1]$ (see Section 3) the maximal distance between any two neighboring observations is $O_p(\log n/n)$ [21]. This implies
that together with the assumption that elements of $\Omega(z)$ are smooth functions on $z \in [0, 1]$, we can estimate the whole curve $\Omega_{uv}(z)$ from the point estimates $\{\hat{\Omega}_{uv}(z^i)\}_{i \in [n]}$. The resulting approximation is of smaller order than the optimal non-parametric rate of convergence $\mathcal{O}_p(n^{-2/5})$.

**Relationship to related work.** Yin et al. [57] proposed to estimate the parameters in (1), $\mu(z)$ and $\Sigma(z)$, by minimizing the kernel smoothed negative log-likelihood

$$\min_{\Sigma(z), m(z)} \sum_{i \in [n]} w^i(z) \left( (x^i - m(z))\Sigma^{-1}(z)(x^i - m(z))^T - \log \det(\Sigma^{-1}(z)) \right).$$

The solution to the above optimization problem is given in (4). However, these estimates work only in a low-dimensional setting when the sample size is larger than the dimension $p$. Adding the penalty in (3) allows us to estimate the parameters even when $n \ll p$.

The objective in (3) is similar to the one used in Danaher et al. [12]. Their group graphical lasso objective is

$$\min \sum_{k \in [K]} n_k \left( \text{tr} (\hat{\Sigma}_k \Omega_k) - \log \det \Omega_k \right) + \lambda \sum_{v \neq u} \sqrt{\sum_{k \in [K]} \Omega_{vu}^2_k},$$

where $n_k$ denotes the sample size for class $k$ and $\hat{\Sigma}_k$ is the sample covariance of observation in class $k$. Notice, however, that there are fixed number of classes, $K$, and for each class there are multiple observations that are used to estimate the class specific covariance matrix $\Sigma_k$. In contrast, we are estimating $\Omega(z)$ at each $z \in \{z^i\}$ using the local smoothed estimator. Under the assumption that the density of $Z$ is continuous, we will have that all the observed values $\{z^i\}$ are distinct and as a result the optimization problem in (3) becomes more complicated as the sample size grows.

Finally, our proposal is related to the work of Zhou et al. [68] and Kolar and Xing [25] where

$$\min_{\Omega(\tau) > 0} \left( \text{tr} (\hat{\Sigma}(\tau) \Omega(\tau)) - \log \det \Omega(\tau) \right) + \lambda \sum_{v \neq u} |\Omega_{vu}(\tau)|$$

is used to estimate the parameters at a single point $\tau \in [0, 1]$ from independent observations $\{x^i \sim N(\mu(i/n), \Sigma(i/n))\}_{i \in [n]}$. Their model is different from ours in the following aspects: i) the index variable is observed on a grid $z_i = i/n$ (fixed design) and ii) the conditional independence structure can change with the index variable.

### 2.3 Proximal Iterative Smoothing

In this section, we develop a Proximal Iterative Smoothing algorithm, termed PRISMA, for solving the optimization problem in (3). We
build on the work of Orabona et al. [39] who developed a fast algorithm for finding the minimizer of the regular graphical lasso objective [62, 2, 16]. The optimization problem in (3) can also be solved by the Alternating Direction Methods of Multipliers (ADMM) [4, 12]. From our numerical experience, we observed that PRISMA converges faster than ADMM for our task and is as easy as ADMM for implementation (see Figure 1 below).

The proximal iterative smoothing algorithm is suitable for solving a convex optimization that can be written as a sum of three parts: a smooth part, a simple Lipschitz part, and a non-Lipschitz part. The basic idea of PRISMA is to construct a $\beta$-Moreau envelope of the simple Lipschitz function, then perform regular accelerated proximal gradient descent on the smoothed objective. Specifically, we decompose the objective (3) into three parts: (i) the smooth part: $\operatorname{tr} (\hat{\Sigma}(z^j)\Omega(z^j))$, (ii) the non-smooth Lipschitz part: $\lambda \sum_{v \neq u} \sqrt{\sum_{j=1}^n \Omega_{vu}(z^j)^2}$, and (iii) the convex non-continuous part: $-\log \det \Omega(z^j) + \delta_{S^{++}}(\Omega(z^j))$.

Then the non-smooth Lipschitz part is approximated by the following $\beta$-Moreau envelope

$$
\varphi_\beta(\Omega(z)) = \inf_{U(z)} \left\{ \sum_i \|U(z^i) - \Omega(z^i)\|^2_F + \lambda \sum_{v \neq u} \sqrt{\sum_{j=1}^n U_{vu}(z^j)^2} \right\}.
$$

The $\beta$-Moreau envelope given above is $\frac{1}{\beta}$-smooth, which means that it has a Lipschitz continuous gradient with Lipschitz constant $\frac{1}{\beta}$. We can compute the gradient with the following proximal operator

$$
\nabla \varphi_\beta(\Omega(z^j)) = \frac{1}{\beta} \left( \Omega(z^j) - \text{prox}_{\ell_1 \setminus \ell_2}(\Omega(z^j), \lambda \beta) \right),
$$

where

$$
\text{prox}_{\ell_1 \setminus \ell_2}(\Omega(z^j), \lambda \beta) = \arg\min_{\Omega} \left\{ \frac{\|\Omega - \Omega(z^j)\|^2_F}{2\lambda \beta} + \sum_{v \neq u} \sqrt{\sum_{j=1}^n \Omega_{vu}(z^j)^2} \right\}
$$

can be obtained in a closed-form [1] as

$$
\text{prox}_{\ell_1 \setminus \ell_2}(\Omega(z^j), \lambda \beta)_{uv} = \Omega_{uv}(z^j) \left( 1 - \frac{\lambda \beta}{\sqrt{\sum_{j=1}^n \Omega_{vu}(z^j)^2}} \right).
$$

The PRISMA algorithm works by performing an accelerated proximal gradient descent on the smoothed objective

$$
\sum_i \operatorname{tr} (\hat{\Sigma}(z^j)\Omega(z^j)) + \varphi_\beta(\Omega(z)) + \sum_i \left( -\log \det \Omega(z^j) + \delta_{S^{++}}(\Omega(z^j)) \right).
$$
Let $L_f$ denote the Lipschitz constant of the gradient of $\text{tr} (\hat{\Sigma}(z^i)\Omega(z^i))$ and let $L_k = L_f + \frac{1}{\beta}$. Then $\text{tr} (\hat{\Sigma}(z^i)\Omega(z^i)) + \varphi_\beta(\Omega(z))$ is $L_k$-smooth. The proximal gradient descent is performing the following update

$$
\Omega(z^i) \leftarrow \text{prox} \left( -\log \det + \delta_{s^p} \right) \left[ \Omega(z^i) - \frac{1}{L_k} \left( \nabla \text{tr} (\hat{\Sigma}(z^i)\Omega(z^i)) + \nabla \varphi_\beta(\Omega(z)) \right), \frac{1}{L_k} \right].
$$

where $1/L_k$ is the step size and the proximal operator is computing a solution to the following optimization problem

$$
\text{prox} \left( -\log \det + \delta_{s^p} \right) \left( \Lambda, \frac{1}{L_k} \right) = \arg\min_{\Omega} \left\{ L_k \| \Omega - A \|_F^2 - \log \det(\Omega) + \delta_{s^p} \right\}.
$$

The update can be equivalently written as

$$
\Omega(z^i) \leftarrow \text{prox} \left( -\log \det + \delta_{s^p} \right) \left[ \Omega(z^i) - \frac{1}{L_k} \left( \hat{\Sigma}(z^i) + \frac{1}{\beta} \left( \Omega(z^i) - \text{prox}_{\ell_1,\ell_2}(\Omega(z^i), \lambda\beta) \right) \right), \frac{1}{L_k} \right].
$$

Next, we describe how to compute the proximal operator in (8). Observe that (8) is equivalent to the following constrained problem

$$
\min_{\Omega > 0} \frac{L_k \| \Omega - A \|_F^2}{2} - \log \det(\Omega).
$$

From the first-order optimality condition we have that

$$
L_k \Omega - \Omega^{-1} = L_k A \quad \text{and} \quad \Omega > 0.
$$

Let $L_k A = Q\Lambda Q^T$ denote the eigen-decomposition with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. The update will be of the form $\Omega = Q\text{diag}(\omega_1, \ldots, \omega_p)Q^T$ where $\omega_i$ can be obtained as a positive solution to the equation

$$
L_k \omega_i - \frac{1}{\omega_i} = \lambda_i, \quad i \in [p],
$$

that is

$$
\omega_i = \frac{\lambda_i + \sqrt{\lambda_i^2 + 4L_k}}{2L_k}.
$$

In summary, the proximal operator in (8), can be computed as

$$
\text{prox} \left( -\log \det + \delta_{s^p} \right) \left( \Lambda, \frac{1}{L_k} \right) = Q\text{diag}(\omega_1, \ldots, \omega_p)Q^T
$$

where Q contains eigenvectors of A and $\omega_i$ is given in (9).

The above described algorithm can be accelerated by combining two sequences of iterates $\Omega(z)$ and $\Theta(z)$, as in [38]. The details are given in Algorithm 1.

Note that since $\text{tr} (\hat{\Sigma}(z^i)\Omega(z^i))$ is linear, $L_f$ can be arbitrary small. Orabona et al. [39] suggest choosing $\beta = \mathcal{O}(1/k)$ to achieve optimal
Figure 1: Typical convergence behavior of PRISMA and ADMM when solving the optimization problem in (3). The curve was obtained on a problem with \((n, p) = (500, 50)\).
2.3 Proximal Iterative Smoothing

convergence $O(\log k/k)$. Figure 1 shows a typical convergence curve of PRISMA and ADMM for our optimization problem (see Section 4.1 for more details on our simulation studies).

Solving (3) can be slow when the sample size is large, even with the scalable procedure described above. To speed up the estimation procedure, one may consider estimating $\Omega(z)$ at $K \ll n$ index points $(\tilde{z}_i^1)_{i \in [K]}$, which are uniformly spread over the domain of $Z$. In this case the optimization procedure in (5) becomes

$$\min_{\{\Omega(z) > 0\}_{i \in [K]}} \sum_{i \in [K]} \left( \text{tr}(\hat{\Sigma}(\tilde{z}_i^1)\Omega(\tilde{z}_i^1)) - \log \det \Omega(\tilde{z}_i^1) \right) - \lambda \sum_{v \neq u} \sqrt{\sum_{i \in [K]} \Omega_{vu}(\tilde{z}_i^1)^2}.$$  

Choice of $K$ determines trade off between information aggregated across index variables and speed of computation. We report our numerical results in Section 4.1 using $K = 50$.

The optimization procedure can be further accelerated by adopting the covariance screening trick $[56, 12]$ in which one identifies the connected components of the resulting graph directly from the sample covariance matrix. The optimization program in (3) is then solved for each connected component separately, which dramatically improves the running time, while still obtaining exactly the same solution as solving the whole big problem. See Section 4 in Danaher et al. [12] for more details.
3.1 CONSISTENT GRAPH IDENTIFICATION

In this section we give a result that the set \( S \) in (2) can be recovered with high-probability. We start by describing the assumption under which this result will be established. Let \( A(x) = (A_{uv}(x))_{u \in [k_1], v \in [k_2]} \) with \( A_{uv}(x) : \mathcal{X} \rightarrow \mathbb{R} \) be a matrix of functions. Without loss of generality, let \( \mathcal{X} = [0, 1] \) (otherwise we can normalize any interval \([a, b]\) to \([0, 1])\). Given a collection of index points \( n = \{x^i\}_{i \in [n]} \), let \( \mathcal{H}^n(A(x)) = \mathcal{H}^n(A(x); n) \in \mathbb{R}^{k_1 \times k_2} \) be defined as

\[
\mathcal{H}_{uv}^n(A(x)) = \frac{1}{\sqrt{n}} \sqrt{\sum_{x^i \in n} A_{uv}^2(x^i), \ u \in [k_1], v \in [k_2]}
\]

As we can see, element \((u, v)\) of \( \mathcal{H}^n(A(x)) \) is the quadratic mean of \( A_{uv}(x^1), \ldots, A_{uv}(x^n) \). The following lemma summarizes some basic facts about \( \mathcal{H}^n(\cdot) \).

**Lemma 1.** Let \( A(x) \) and \( B(x) \) be symmetric matrices for all \( x \in \mathcal{X} \). Given \( n \) index points in \( \mathcal{X} \), \( \{x^i\}_{i \in [n]} \), we have

(a) \( \| \mathcal{H}^n(A(x)) \|_{\infty, \infty} \leq \max_{i \in [n]} \| A(x^i) \|_{\infty, \infty} \)

(b) \( \| \mathcal{H}^n(A(x)) \|_{\infty} \leq \max_{i \in [n]} \| A(x^i) \|_{\infty} \)

(c) \( \| \mathcal{H}^n(A(x)B(x)) \|_{\infty} \leq \| \mathcal{H}^n(A(x)) \|_{\infty} \| \mathcal{H}^n(B(x)) \|_{\infty, \infty} \)

(d) \( \| \mathcal{H}^n(A(x) + B(x)) \|_{\infty} \leq \| \mathcal{H}^n(A(x)) + \mathcal{H}^n(B(x)) \|_{\infty} \)

**Proof.** (a)

\[
\| \mathcal{H}^n(A(x)) \|_{\infty, \infty} = \max_u \sum_v \left( \frac{\sum_{i \in [n]} A_{uv}^2(x^i)}{n} \right)^{1/2} 
\leq \max_u \max_{i \in [n]} \sum_v A_{uv}(x^i) 
= \max_{i \in [n]} \| A(x^i) \|_{\infty, \infty}.
\]

(b) \( \forall u, v \), we have

\[
\mathcal{H}_{uv}^n(A(x)) = \sqrt{\sum_{i \in [n]} A_{uv}^2(x^i) / n} 
\leq \max_{i \in [n]} A_{uv}(x^i) 
= \max_{i \in [n]} \| A(x^i) \|_{\infty}.
\]
thus \( \|\mathcal{I}_n(A(x))\|_\infty \leq \max_{i \in [n]} \|A(x^i)\|_\infty \).

(c) \( \forall u, v \), we have

\[
\mathcal{I}_{uv}^n(A(x)B(x)) = \sum_{i \in [n]} \left( \frac{\sum_{j \in [n]} A_{uj}(x^j)B_{jk}(x^j)}{n} \right) ^2 \]

\[
\leq \sum_{k} \left( \frac{\sum_{i \in [n]} A_{uk}^2(x^i)}{n} \right) \left( \frac{\sum_{i \in [n]} B_{kv}^2(x^i)}{n} \right) \]

\[
\leq \max_k \left( \frac{\sum_{i \in [n]} A_{uk}^2(x^i)}{n} \right) \sum_k \left( \frac{\sum_{i \in [n]} B_{kv}^2(x^i)}{n} \right) \]

\[
= \|\mathcal{I}_n(A(x))\|_\infty \|\mathcal{I}_n(B(x))\|_\infty. \]

Thus \( \|\mathcal{I}_n(A(x)B(x))\|_\infty \leq \|\mathcal{I}_n(A(x))\|_\infty \|\mathcal{I}_n(B(x))\|_\infty. \)

(d) \( \forall u, v \), we have

\[
\mathcal{I}_{uv}^n(A(x) + B(x)) = \sum_{i \in [n]} \left( \frac{A_{uv}(x^i) + B_{uv}(x^i)}{n} \right)^2 \]

\[
\leq \sum_{i \in [n]} \left( \frac{A_{uv}(x^i)}{n} \right)^2 + \sum_{i \in [n]} \left( \frac{B_{uv}(x^i)}{n} \right)^2 \]

\[
= \mathcal{I}_{uv}^n(A(x)) + \mathcal{I}_{uv}^n(B(x)) \]

Thus \( \|\mathcal{I}_n(A(x) + B(x))\|_\infty \leq \|\mathcal{I}_n(A(x))\|_\infty + \|\mathcal{I}_n(B(x))\|_\infty. \)

\[
\square \]

We will use the following assumptions to develop our theoretical results.

**A1 (Distribution).** We assume the data \( \{(x^i, z^i)\}_{i \in [n]} \) are drawn i.i.d. from the following model

\[
z^i \sim f(z) \]

\[
x^i | z^i \sim N(0, \Sigma^*(z^i)) \quad (12) \]

Marginal density \( f(z) \) of \( Z \) is lower bounded on the support \([0, 1] \), that is, there is a constant \( C_f \) such that

\[
\inf_{z \in [0,1]} f(z) \geq C_f > 0. \]

Furthermore, the density function is \( C_d \)-smooth, that is, the first derivative is \( C_d \)-Lipschitz

\[
|f'(z) - f'(z')| \leq C_d |z - z'|, \quad \forall z, z' \in [0, 1]. \]

**A2 (Kernel).** The kernel function \( K(\cdot) : \mathbb{R} \to \mathbb{R} \) is a symmetric, prob-
ability density function supported on $[-1, 1]$. There exists a constant $C_K < \infty$, such that

$$\sup_x |K(x)| \leq C_K \quad \text{and} \quad \sup_x K(x)^2 \leq C_K.$$  

Furthermore, the kernel function is $C_L$-Lipschitz on the support $[-1, 1]$, that is

$$|K(x) - K(x')| \leq C_L |x - x'|, \quad \forall x, x' \in [-1, 1].$$

A3 (Covariance). There exist constants $\Lambda_\infty, C_\infty < \infty$, such that

$$\frac{1}{\Lambda_\infty} \leq \inf_z \sigma_{\min}(\Sigma^*(z)) \leq \sup_z \sigma_{\max}(\Sigma^*(z)) \leq \Lambda_\infty$$

and

$$\sup_z ||| \Sigma^*(z) |||_{\infty, \infty} \leq C_\infty.$$  

A4 (Smoothness). There exists a constant $C_\Sigma$, such that

$$\max_{u,v} \sup_z \left| \frac{d}{dz} \Sigma^*_{uv}(z) \right| \leq C_\Sigma \quad \text{and} \quad \max_{u,v} \sup_z \left| \frac{d^2}{dz^2} \Sigma^*_{uv}(z) \right| \leq C_\Sigma$$

$$\max_u \sup_z \left| \frac{d}{dz} \mu^*_u(z) \right| \leq C_\Sigma \quad \text{and} \quad \max_u \sup_z \left| \frac{d^2}{dz^2} \mu^*_u(z) \right| \leq C_\Sigma.$$  

A5 (Irrepresentable Condition). Let

$$\mathcal{I}(z) = \Omega^*(z)^{-1} \otimes \Omega^*(z)^{-1}.$$  

The sub-matrix $\mathcal{I}_{SS}(z)$ is invertible for all $z \in [0, 1]$. There exist constants $\alpha \in (0, 1]$ and $C_2$ such that

$$\sup_z ||| \mathcal{I}_{SS}(z) (\mathcal{I}_{SS}(z))^{-1} |||_{\infty, \infty} \leq 1 - \alpha \quad \text{and} \quad \sup_z \| \mathcal{I}_{SS}(z)^{-1} \|_{\infty, \infty} \leq C_2.$$  

Assumption A1 on the density function $f(z)$ is rather standard [see, for example, 32]. In (12), we assume that the conditional mean $\mu(z) = E[X \mid Z = z]$ is equal to zero. This assumption is not necessary, however, it makes for simpler exposition and derivation of results. Alternatively, we could impose a smoothness condition on the conditional mean, for example, that each component of $m(\cdot)$ has two continuous derivatives. This alternative assumption would not change results presented in this and the following section as the rate of convergence will be dominated by how fast we can estimate elements of $\Omega^*(z)$. The assumption that $X \mid Z$ follows a multivariate Normal distribution is quite strong. This could be changed to a sub-Gaussian distribution and our results would not change except for
constants. However, establishing our results under moment conditions would require different proof techniques. Assumption A2 is a standard technical condition that is used in literature on kernel smoothing [see, for example, 32]. A number of commonly used kernel functions satisfy A2, such as the Box car kernel \( K(x) = \frac{1}{2} I(|x| \leq 1) \), the Epanechnikov kernel \( K(x) = \frac{3}{4} (1 - x^2) I(|x| \leq 1) \) and the Tricube kernel \( K(x) = \frac{70}{81} (1 - |x|^3) I(|x| \leq 1) \). Assumption A3 ensures that the model is identifiable in the population setting. Our assumption could be relaxed by allowing \( \Lambda_\infty \) and \( C_\infty \) to depend on the sample size \( n \) as in [44], however, we choose not do keep track of this dependence to simplify the exposition. Assumption A4 ensures sufficient smoothness of the mean and covariance functions. These conditions are necessary [57]. Assumption A5 is a version of the irrepresentable condition [66, 54, 36] commonly used in the high-dimensional literature to ensure exact model selection. We use an assumption on the Fisher information matrix previously imposed in Ravikumar et al. [44], however, we assume that it holds uniformly for all values of \( z \in [0, 1] \).

Remark 2. It follows directly from Lemma 1 that given the index points \( \{z_i\}_{i \in [n]} \), we have

\[
\| \mathcal{T}^n(\Sigma^*(z)) \|_{\infty, \infty} \leq \sup_z | \Sigma^*(z) |_{\infty, \infty} \leq C_{\infty}, \\
| \mathcal{T}^n(J_{SS}(z)(J_{SS}(z))^{-1}) |_{\infty, \infty} \leq \sup_z | J_{SS}(z)(J_{SS}(z))^{-1} |_{\infty, \infty} \leq 1 - \alpha, \\
| \mathcal{T}^n(J_{SS}(z)) |_{\infty, \infty} \leq \sup_z | J_{SS}(z) |_{\infty, \infty} \leq C_3.
\]

Our first result provides conditions needed for correct recovery of the sparsity pattern of the conditional inverse covariance matrix.

Theorem 3. Suppose that assumptions A1-A5 hold. Let \( \{x^i, z^i\}_{i \in [n]} \) be independent observations from the model in (12). Suppose there are two constants \( C_1 \) and \( C_2 \) such that

\[
\Omega_{\min} = \min_{(u,v) \in S} \sqrt{\sum_{i \in [n]} (\Omega^*_uv(z^i))^2} \geq C_1 n^{-2/5} \sqrt{\log p}
\]

and the sample size satisfies

\[
n \geq C_2 d^{5/2} (\log p)^{5/4}.
\]

If the bandwidth is chosen as \( h \asymp n^{-1/5} \) and the penalty parameter as \( \lambda \asymp n^{1/10} \sqrt{\log p} \), then

\[
\text{pr}(\hat{S} = S) = 1 - O(p^{-1}) \xrightarrow{n \to \infty} 1,
\]

where \( \hat{S} \) is the estimated support given in (6) and \( S \) is given in (2).
The proof of Theorem 3 is given in Section 3.2 and, at a high-level, is based on the primal-dual witness proof described in [44]. The main technical difficulty lies in establishing a bound on 
\[ \sup_{z \in [0,1]} \max_{u,v} \left| \hat{S}_{uv}(z) - S_{uv}(z) \right|, \]
which is provided in Lemma 6. This result is new and may be of independent interest.

Our results should be compared to those in Kolar and Xing [25], where \( \hat{S}(z) \) was estimated using the optimization problem in (7). Kolar and Xing [25] required the sample size \( n \geq Cd^3(\log p)^{3/2} \), while our procedure requires the sample size to satisfy \( n \geq Cd^{5/2}(\log p)^{5/4} \). This difference primarily arises from more careful treatment of the smoothness of the component functions of \( \Omega^*(z) \). Next, we compare requirements on the minimum signal strength (commonly referred to as the \( \beta_{\text{min}} \) condition). Kolar and Xing [25] require that \( \min_{(u,v)\in S(z)} |\Omega^*(u,v)(z)| \geq Cn^{-1/3} \sqrt{\log p} \), while our procedure relaxes this condition to \( \min_{(u,v)\in S(z)} \sqrt{n^{-1} \sum_{i\in[n]} (\Omega^*(u,v)(z))^2} \geq Cn^{-2/5} \sqrt{\log p} \).

Furthermore, we only require the following functionals of the covariance and Fisher information matrix to be upper bounded
\[ |\mathcal{H}^n(\Sigma^*(z))|_{\infty,\infty}, \quad |\mathcal{H}^n(J_{S\cdot S}(z)(J_{SS}(z))^{-1})|_{\infty,\infty}, \quad \text{and} \quad |\mathcal{H}^n(J_{SS}(z))|_{\infty,\infty}, \]
while Kolar and Xing [25] required the conditions to hold uniformly in \( z \), which is a more stringent condition as shown in Remark 2.

**Remark 4.** As discussed in Section 2.3, the set \( \hat{S} \) can be estimated using the optimization program in (10) instead of the one in (3). The statement of Theorem 3 remains to hold if
\[ \min_{(u,v)\in S(z)} \sqrt{n^{-1} \sum_{i\in[K]} (\Omega^*(u,v)(z))^2} \geq C_1 n^{-2/5} \sqrt{\log p}. \]

The following proposition gives a consistency result on the estimation of \( \Omega^*(z) \) in Frobenius norm.

**Proposition 5.** Under the assumption of Theorem 3, there exists a constant \( C \), such that
\[ n^{-1} \sum_{i\in[n]} \|\hat{\Omega}(z^i) - \Omega^*(z^i)\|_F^2 \leq Cn^{-4/5}(|S| + p) \log p \]
with probability \( 1 - o(p^{-1}) \).

**Proof.** Recall that
\[ \frac{1}{n} \sum_{i=1}^{n} \|\hat{\Omega}(z^i) - \Omega^*(z^i)\|_F^2 = \|\mathcal{H}^n(\hat{\Omega}(z) - \Omega^*(z))\|_F^2. \]
Under the assumption of Theorem 3, the pattern of non-zero elements is recovered correctly. Therefore, there are $|S| + p$ non-zero entries in the matrix $\mathcal{H}^n(\hat{\Omega}(z) - \Omega^*(z))$. We have

$$
\|\mathcal{H}^n(\hat{\Omega}(z) - \Omega^*(z))\|_F \leq \sqrt{|S| + p}\|\mathcal{H}^n(\hat{\Omega}(z) - \Omega^*(z))\|_\infty
$$

and the proof follows from an application of Lemma 8 to the right hand side.

Zhou et al. [68] obtained $O_p(n^{-2/3}(|S| + p)\log n)$ as the rate of convergence in the Frobenius norm. The improvement in the rate, from $n^{-2/3}$ to $n^{-4/5}$ is due to a more careful treatment of the smoothness of the precision matrix functions.

### 3.2 Proof of Sparsistency

In this section we provide a proof of Theorem 3. The proof is based on the primal-dual witness strategy described in Wainwright [54] and Ravikumar et al. [44]. In the course of the proof we will construct a solution $(\hat{\Omega}(z^i))_{i \in [n]}$ to the optimization problem in (3) and show that it recovers the set $S$ correctly.

We will need the following restricted optimization problem

$$
\min_{\{\bar{\Omega}(z^i)\}_{i \in [n]} \atop \Omega(\hat{z}^i) = \hat{\Omega}(z^i) \in \Omega S \in [n]} \sum_{i \in [n]} \left(\text{tr}(\hat{\Sigma}(z^i)\Omega(z^i)) - \log \det \Omega(z^i)\right) + \lambda \sum_{v \neq u} \sqrt{\sum_{i \in [n]} \Omega_{uv}(z^i)^2}.
$$

(13)

Let $(\hat{\Omega}(z^i))_{i \in [n]}$ be the unique solution to the restricted problem. Our strategy is to show the following

1. $\hat{\Omega}_{uv}(z^i) = \hat{\Omega}_{uv}(z^i)$ and $\sum_{i \in [n]} (\hat{\Omega}_{uv}(z^i))^2 \neq 0$ for $(u, v) \in S$ and $i \in [n]$;

2. $\hat{\Omega}_{uv}(z^i) = 0$ for $(u, v) \in \hat{S}$ and $i \in [n]$,

hold with high probability. We will achieve these goals through the following steps

**Step 1:** Solve the restricted problem in (13) to obtain $(\hat{\Omega}(z^i))_{i \in [n]}$.

**Step 2:** Compute the sub-differential $\{\hat{Z}(z^i))_{i \in [n]}$ of the penalty term

$$
\sum_{v \neq u} \sqrt{\sum_{i \in [n]} (\Omega_{uv}(z^i))^2} \text{ at } (\hat{\Omega}(z^i))_{i \in [n]}.
$$

which can be obtained in a closed form as

$$
\hat{Z}_{uv}(z^i) = \frac{\hat{\Omega}_{uv}(z^i)}{\sqrt{\sum_{i \in [n]} (\hat{\Omega}_{uv}(z^i))^2}}.
$$

**Step 3:** For each $(u, v) \in S^c$, set

$$
\hat{Z}_{uv}(z^i) = \frac{1}{\lambda} \left(-\hat{\Sigma}_{uv}(z^i) + (\hat{\Omega}(z)^{-1})_{uv}\right).
$$
3.2 Proof of Sparsistency

STEP 4: Verify that $\|\hat{\mathcal{H}}^n(\tilde{Z}_S(z))\|_\infty < n^{-1/2}$ and $\sum_{i \in [n]} (\Omega_{uv}(z^1))^2 \neq 0$.

In the first three steps, we construct the solution to the restricted optimization problem $\{\hat{\Omega}(z^1)\}_{i \in [n]}$ and the sub-differential $\{\tilde{Z}(z^1)\}_{i \in [n]}$. In the forth step, we check whether the construction satisfies first order optimality conditions for the optimization program in (3) and that no elements of $S$ are excluded. The sub-differential of the penalty term in (3) evaluated at the solution $\{\hat{\Omega}(z^1)\}_{i \in [n]}$ satisfies

$$\hat{Z}_{uv}(z^1) = \begin{cases} 0 & \text{if } u = v, \\ \frac{\hat{\Omega}_{uv}(z^1)}{\sqrt{\sum_{i \in [n]} \hat{\Omega}_{uv}^2(z^1)}} & \text{if } u \neq v \text{ and } (u, v) \in \text{support}(\hat{\Omega}(z)), \\ g_{uv}(z^1) \text{ s.t. } \sqrt{\sum_{i \in [n]} g_{uv}^2(z^1)} \leq 1 & \text{if } u \neq v \text{ and } (u, v) \not\in \text{support}(\hat{\Omega}(z)). \end{cases}$$

By showing that

$$\sqrt{\sum_{i \in [n]} \hat{Z}_{uv}^2(z^1)} < 1$$

holds with high probability, we establish that $\{\hat{\Omega}(z^1)\}_{i \in [n]}$ is the solution to the unrestricted optimization program (3) with high probability.

Lemma 6-9, given below, provide technical details. Let

$$\Delta(z) = \tilde{\Omega}(z) - \Omega^*(z)$$

and

$$R(\Delta(z)) = \tilde{\Omega}^{-1}(z) - \Omega^*^{-1}(z) + \Omega^*^{-1}(z)\Delta(z)\Omega^*^{-1}(z).$$

Lemma 6 shows that with high probability

$$\|\mathcal{I}^n(\tilde{\Sigma}(z) - \Sigma^*(z))\|_\infty \leq Cn^{-2/5} \sqrt{\log p}.$$ 

This is a main technical result that allows us to establish the desired rate of convergence. Using Lemma 6 we establish in Lemma 8 that

$$\|\mathcal{I}^n(\Delta(z))\|_\infty \lesssim \|\mathcal{I}^n(\tilde{\Sigma}(z) - \Sigma^*(z))\|_\infty + n^{-1/2}\lambda.$$ 

Lemma 7 shows that

$$\|\mathcal{I}^n(R(\Delta(z)))\|_\infty \lesssim d\|\mathcal{I}^n(\Delta(z))\|_\infty$$

when $\|\mathcal{I}^n(\Delta(z))\|_\infty$ is sufficiently small. Lemma 9 obtains the strict dual feasibility and, finally, by applying Lemma 8 again we obtain that the procedure can identify the true support $S$ if

$$\sqrt{\sum_{i \in \mathbb{N}} \Omega_{uv}^*(z^1)^2} \sim n^{-2/5} \sqrt{\log p}, \quad \forall (u, v) \in S.$$
Lemma 6. Suppose that assumptions A1–A4 hold. If the bandwidth parameter satisfies \( h = n^{-1/5} \), then there exists a constant \( C \), which depends only on \( C_K, C_\Sigma \), and \( \Lambda_\infty \), such that
\[
\sup_z \| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty \leq C n^{-2/5} \sqrt{\log p}
\]
and
\[
\| \mathcal{R}_n(\hat{\Sigma}(z) - \Sigma^*(z)) \|_\infty \leq C n^{-2/5} \sqrt{\log p}
\]
with probability at least \( 1 - \mathcal{O}(p^{-1}) \).

Proof. Lemma follows from the bound
\[
\| \mathcal{R}_n(A(z)) \|_\infty \leq \sup_{z \in [0,1]} \| A(z) \|_\infty
\]
once we find a bound for \( \sup_z \| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty \).

Consider the following decomposition
\[
\sup_z \| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty \leq \sup_z \| \hat{\Sigma}(z) - E \hat{\Sigma}(z) \|_\infty + \sup_z \| E \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty.
\]

For the bias term, \( \sup_z \| E \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty \), we have
\[
E \left[ \hat{\Sigma}_{uv}(z) \right] = E_Z \left[ E_{X|Z} \left[ \frac{\sum_{i=1}^n h(z^i - z)X_i^u X_i^v}{\sum_{i=1}^n K_h(z^i - z)} \right] \right]
\]
\[
= E_Z \left[ \frac{\sum_{i=1}^n K_h(z^i - z)\Sigma^*_u(z^i)}{\sum_{i=1}^n K_h(z^i - z)} \right]
\]
\[
= E_Z \left[ \frac{\sum_{i=1}^n K_h(z^i - z)\Sigma^*_u(z^i)}{\sum_{i=1}^n K_h(z^i - z)} \right] + \mathcal{O} \left( \frac{1}{nh} \right)
\]
where the last step follows from Pagan and Ullah [40, Page 101, (3.56)]. Since
\[
E_Z \left[ K_h(z^i - z)\Sigma^*_u(z^i) \right] = \frac{1}{h} \int h \frac{z^i - z}{h} \Sigma^*_u(z^i) f(z^i) dz^i
\]
\[
= \int_{-1}^1 K(a) \Sigma^*_u(z + ah) f(z + ah) da
\]
\[
= \Sigma^*_u(z) f(z) + \mathcal{O}(h^2)
\]
and, similarly,
\[
E_Z \left[ K_h(z^i - z) \right] = f(z) + \mathcal{O}(h^2),
\]
we obtain that
\[
E \left[ \hat{\Sigma}_{uv}(z) \right] = \Sigma^*_u(z) + \mathcal{O}(h^2) + \mathcal{O} \left( \frac{1}{nh} \right)
\]
for any \( u, v \in [p] \) and \( z \in [0, 1] \). Therefore,
\[
\sup_z \| E \left[ \hat{\Sigma}_{uv}(z) \right] - \Sigma^*_u(z) \|_\infty = \mathcal{O}(h^2) + \mathcal{O} \left( \frac{1}{nh} \right).
\]
For the variance term, sup\( z \) \( \| \hat{\Sigma}(z) - E\hat{\Sigma}(z) \|_\infty \), we divide the interval \([0, 1]\) into \( K \) segments \( \{ [(k - 1)/K, k/K] \mid k \in [K]\} \), where \( K \) will be chosen later. We have
\[
\operatorname{sup}_{z} \| \hat{\Sigma}(z) - E\hat{\Sigma}(z) \|_\infty = \max_{k \in [K]} \operatorname{sup}_{z \in [(k - 1)/K, k/K]} \| \hat{\Sigma}(z) - E\hat{\Sigma}(z) \|_\infty \\
\leq \max_{k \in [K]} \left\| \hat{\Sigma} \left( \frac{2k - 1}{2K} \right) - E\hat{\Sigma} \left( \frac{2k - 1}{2K} \right) \right\|_\infty \\
+ \max_{k \in [K]} \operatorname{sup}_{z \in [(k - 1)/K, k/K]} \left\| \hat{\Sigma}(z) - E\hat{\Sigma}(z) - \left( \hat{\Sigma} \left( \frac{2k - 1}{2K} \right) - E\hat{\Sigma} \left( \frac{2k - 1}{2K} \right) \right) \right\|_\infty.
\]

(15)

We start with the first term in (15). We will use some standard results for the kernel density estimator
\[
\hat{f}(z) = \frac{1}{nh} \sum_{i \in [n]} K \left( \frac{z - x_i}{h} \right).
\]
In particular, we have that
\[
\operatorname{sup}_{z} |E[\hat{f}(z)] - f(z)| = O(h^2)
\]
and
\[
\operatorname{sup}_{z} |\hat{f}(z) - E[\hat{f}(z)]| = O_P \left( \sqrt{\frac{\log(1/h)}{nh}} \right).
\]
See, for example, Pagan and Ullah [40] for the first result and Silverman [48] for the second one. From these results, we have that
\[
\operatorname{sup}_{z} \frac{\hat{f}(z)}{f(z)} = 1 + O(h^2) + O_P \left( \sqrt{\frac{\log(1/h)}{nh}} \right).
\]
(16)

Thus
\[
\hat{\Sigma}_{uv} \left( \frac{2k - 1}{2K} \right) = \frac{n^{-1} \sum_{i} K_h \left( z^i - \frac{2k - 1}{2K} \right) x^i_u x^i_v}{\frac{n^{-1} \sum_{i} K_h \left( z^i - \frac{2k - 1}{2K} \right)}{}} \\
= \frac{1}{nf \left( \frac{2k - 1}{2K} \right)} \sum_{i} K_h \left( z^i - \frac{2k - 1}{2K} \right) x^i_u x^i_v \\
+ O(h^2) + O_P \left( \sqrt{\frac{\log(1/h)}{nh}} \right).
\]
For a fixed \( k \in [K] \), we have
\[
\left\| \hat{\Sigma} \left( \frac{2k - 1}{2K} \right) - E \left[ \hat{\Sigma} \left( \frac{2k - 1}{2K} \right) \right] \right\|_{\infty}
= \max_{u,v} \left| E \left[ \hat{\Sigma}_{uv} \left( \frac{2k - 1}{2K} \right) \right] - \hat{\Sigma}_{uv} \left( \frac{2k - 1}{2K} \right) \right|
= \max_{u,v} \left\{ \frac{1}{nf} \left( \frac{2k - 1}{2K} \right) \right\}
\sum_{i} K_{h} \left( z^{i} - \frac{2k - 1}{2K} \right) x_{i}^{u} x_{i}^{v} - E \left[ K_{h} \left( z^{i} - \frac{2k - 1}{2K} \right) x_{i}^{u} x_{i}^{v} \right] \right\}
\quad + O(h^{2}) + O_{P} \left( \sqrt{\log \left( \frac{1}{h} \right)} \right). \]

Conditionally of \( z^{i} \), we have
\[
\| x_{i}^{u} x_{i}^{v} \|_{\Psi_{1}} \leq 2 \| e_{i}^{T} x_{i} \|_{\Psi_{2}} \| e_{i}^{T} x_{i} \|_{\Psi_{2}} = 2 \Sigma_{uv}^{1/2} (z^{i}) \Sigma_{uv}^{1/2} (z^{i})
\]
and
\[
\| x_{i}^{u} x_{i}^{v} - E \left[ x_{i}^{u} x_{i}^{v} \mid z^{i} \right] \|_{\Psi_{1}} \leq 2 \| x_{i}^{u} x_{i}^{v} \|_{\Psi_{1}} \leq 4 \Lambda_{\infty}. \]

Conditionally on \( \{ z^{i} \}_{i \in [n]} \), Proposition 5.16 in [53] gives us that
\[
\left| \sum_{i \in [n]} K \left( \frac{z^{i} - z}{h} \right) \left( x_{i}^{u} x_{i}^{v} - E \left[ x_{i}^{u} x_{i}^{v} \mid z^{i} \right] \right) \right|
\leq M_{\sigma} \left( \sqrt{\sum_{i \in [n]} K \left( \frac{z^{i} - z}{h} \right)^{2} \log(2/\delta)} \right) \left( \max_{i \in [n]} K \left( \frac{z^{i} - z}{h} \right) \right) \log(2/\delta) \right) \right)
\]
with probability at least \( 1 - \delta \). We have
\[
E \left[ K \left( \frac{z^{i} - z}{h} \right)^{2} \right] = \int K^{2} \left( \frac{x - z}{h} \right) f_{Z}(x) dx = O(h) \]
and, similarly,
\[
\text{var} \left[ K \left( \frac{z^{i} - z}{h} \right)^{2} - E \left[ K \left( \frac{z^{i} - z}{h} \right)^{2} \right] \right] \leq E \left[ K \left( \frac{z^{i} - z}{h} \right)^{4} \right] = O(h). \]

Since
\[
\left| K \left( \frac{z^{i} - z}{h} \right)^{2} - E \left[ K \left( \frac{z^{i} - z}{h} \right)^{2} \right] \right| \leq 2 C_{K}^{2} \text{ a.s.,} \]
we can use Bernstein’s inequality to obtain
\[
\sum_{i \in [n]} K \left( \frac{z^{i} - z}{h} \right)^{2} \approx nh + C \sqrt{nh \log(1/\delta)} \]
with probability at least $1 - \delta$. Therefore, for a fixed $u, v$ and $k$, we obtain
\[
\frac{1}{nf(\frac{2k-1}{2K})} \sum_i \left[ K_h \left( z^i - \frac{2k-1}{2K} \right) x^i_u x^i_v - E \left[ K_h \left( z^i - \frac{2k-1}{2K} \right) x^i_u x^i_v \right] \right]
\lesssim \sqrt{\frac{\log(1/\delta)}{nh}}
\] with probability $1 - 2\delta$. An application of the union bound gives us
\[
\Pr \left( \max_{k \in [K]} \left\| \hat{\Sigma} \left( \frac{2k-1}{2K} \right) - E \hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty \geq \epsilon \right) \leq C_1 \exp \left( -C_2 \left( nh^2 + \log(pK) \right) \right).
\] (17)

For the second term in (15), we have
\[
\left\| \hat{\Sigma}(z) - E\hat{\Sigma}(z) - \left( \hat{\Sigma} \left( \frac{2k-1}{2K} \right) - E \hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right) \right\|_\infty
\leq \left\| E\hat{\Sigma}(z) - E\hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty + \left\| \hat{\Sigma}(z) - \hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty.
\] (18)

We further have
\[
\sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left\| E\hat{\Sigma}(z) - E\hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty
\leq \sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left\| E\hat{\Sigma}(z) - \Sigma^*(z) \right\|_\infty + \sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left\| \Sigma^*(z) - \Sigma^* \left( \frac{2k-1}{2K} \right) \right\|_\infty
\leq \left\| \Sigma^* \left( \frac{2k-1}{2K} \right) - E\hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty
= O(h^2) + o \left( \frac{1}{nh} \right) + O(K^{-2}).
\]

using (14) and the smoothness of $\Sigma^*(z)$. We also have
\[
\sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left\| \hat{\Sigma}(z) - \hat{\Sigma} \left( \frac{2k-1}{2K} \right) \right\|_\infty
= \max_{u, v} \sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left| \frac{n^{-1} \sum_{i} K_h \left( z^i - z \right) x^i_u x^i_v}{f(z)} - \frac{n^{-1} \sum_{i} K_h \left( z^i - \frac{2k-1}{2K} \right) x^i_u x^i_v}{f \left( \frac{2k-1}{2K} \right)} \right|
\lesssim \sqrt{\frac{\log(1/\delta)}{nh}} + O(h^2) + O_P \left( \sqrt{\frac{\log(1/h)}{nh}} \right)
\]
using the property of $\hat{f}(z)$ in (16). Using smoothness of $f(z)$,
\[
\sup_{z \in \left[ \frac{k-1}{K}, \frac{k}{K} \right]} \left| (f(z))^{-1} - \left( f \left( \frac{2k-1}{2K} \right) \right)^{-1} \right| = O(K^{-2})
\]
and
\[
\sup_{z \in [\frac{1}{K^2}, K]} \left\| \tilde{\Sigma}(z) - \tilde{\Sigma} \left( \frac{2k - 1}{2K} \right) \right\|_\infty \\
= \max_{u, v} \sup_{z \in [\frac{1}{K^2}, K]} \frac{1}{nh(z)} \left| \sum_i K_h \left( z^i - z \right) x_u^i x_v^i - K_h \left( z^i - \frac{2k - 1}{2K} \right) x_u^i x_v^i \right|
\]
\[
+ O_p(h^2) + O_p \left( \sqrt{\frac{\log(1/h)}{nh}} \right) + O_p(K^{-2})
\]
\[
\leq \max_{u, v} \sup_{z \in [\frac{1}{K^2}, K]} \frac{1}{nh(z)} \left| \sum_i K_h \left( z^i - z \right) x_u^i x_v^i \right|
\]
\[
+ \max_{u, v} \sup_{z \in [\frac{1}{K^2}, K]} \frac{1}{nh(z)} \left| \sum_i K_h \left( z^i - z \right) x_u^i x_v^i \right|
\]
\[
+ \max_{u, v} \sup_{z \in [\frac{1}{K^2}, K]} \frac{1}{nh(z)} \left| \sum_i K_h \left( z^i - z \right) x_u^i x_v^i \right|
\]
\[
+ O(h^2) + O_p \left( \sqrt{\frac{\log(1/h)}{nh}} \right) + O(K^{-2})
\]
\[
= O(h^2) + O_p \left( \sqrt{\frac{\log(1/h)}{nh}} \right) + O(K^{-2}) + O_p ((Kh)^{-1}).
\]

Choosing \( K \propto n \) and inserting back into (18), we obtain
\[
\max_{k \in [K]} \sup_{z \in [1/k^2, k]} \left\| \tilde{\Sigma}(z) - \tilde{\Sigma} \left( \frac{2k - 1}{2K} \right) - \tilde{\Sigma} \left( \frac{2k - 1}{2K} \right) \right\|_\infty
\]
\[
= O(h^2) + O_p \left( \sqrt{\frac{\log(1/h)}{nh}} \right) + O_p ((nh)^{-1}).
\]

(19)

Combining (14), (15), (17), and (19), together with \( K \propto n \) and \( h \propto n^{-1/5} \), we get
\[
\text{pr}(\sup_z \| \tilde{\Sigma}(z) - \Sigma^*(z) \|_\infty \geq \epsilon) \leq C_1 \exp(-C_2 n^{4/5} \epsilon + \log p).
\]

Choosing \( \epsilon \propto n^{-2/5} \sqrt{\log p} \) completes the proof. \( \square \)

**Lemma 7.** Under conditions of Theorem 3, if
\[
\sup_z \| \Delta(z) \|_\infty \leq \frac{1}{3C_{\infty} d'}
\]

then
\[
\| \mathcal{C}^n(R(\Delta(z))) \|_\infty \leq \frac{3d}{2} C_{\infty}^{3} \sup_z \| \Delta(z) \|_\infty^2.
\]
Proof. The remainder term can be re-written as

\[ R(\Delta(z)) = (\Omega^*(z) + \Delta(z))^{-1} - \Omega^*(z) + \Omega^{*-1}(z)\Delta(z)\Omega^{*-1}(z). \]

We have \( \forall \mathbf{z} \in [0, 1], \)

\[ |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty} \leq |(\Omega^{*-1}(z)\Delta(z))J\Omega^{*-1}(z)|_{\infty} \]

\[ \leq |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty}d\|\Delta(z)\|_\infty \]

\[ \leq \frac{1}{3}. \]

where the first step uses facts in Lemma 1, and the second step is based on the fact that \( \Delta(z) \) has at most \( d \) non-zero entries per row/column. Based on convergent matrix expansion

\[ (\Omega^*(z) + \Delta(z))^{-1} = \sum_{k=0}^{\infty} (-1)^k(\Omega^{*-1}(z)\Delta(z))J\Omega^{*-1}(z) \]

\[ = \Omega^{*-1}(z) - \Omega^{*-1}(z)\Delta(z)\Omega^{*-1}(z) \]

\[ + \Omega^{*-1}(z)\Delta(z)\Omega^{*-1}(z)\Delta(z)J\Omega^{*-1}(z) \]

where \( J(z) = \sum_{k=0}^{\infty} (-1)^k(\Omega^{*-1}(z)\Delta(z))J \). Thus

\[ \|R(\Delta(z))\|_\infty = \|\Omega^{*-1}(z)\Delta(z)\Omega^{*-1}(z)\Delta(z)J\Omega^{*-1}(z)\|_\infty \]

\[ \leq |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty}\|\Omega^{*-1}(z)\Delta(z)J\|_{\infty} \]

\[ \leq |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty}^3\|\Delta(z)\|_{\infty}\|J(z)\|_{\infty,\infty} \]

\[ \leq dC^2_\infty(\|\Delta(z)\|_{\infty})^2\|J(z)\|_{\infty,\infty} \]

Where above steps use the fact that \( \Delta(z) \) has at most \( d \) non-zero entries per row/column, and \( A_2 \). Furthermore, we have a bound on \( \|J(z)\|_{\infty,\infty} \) as

\[ \|J(z)\|_{\infty,\infty} \leq \sum_{k=0}^{\infty} (-1)^k(\Omega^{*-1}(z)\Delta(z))J\|_{\infty,\infty} \]

\[ \leq \sum_{k=0}^{\infty} |(\Omega^{*-1}(z)\Delta(z))J|_{\infty,\infty} \]

\[ \leq \sum_{k=0}^{\infty} |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty}^k \]

\[ \leq \frac{1}{1 - |\Omega^{*-1}(z)\Delta(z)|_{\infty,\infty}} \]

\[ \leq \frac{2}{3}. \]

Combining the results

\[ \|\mathcal{O}^n(R(\Delta(z)))\|_\infty \leq \sup_z \|R(\Delta(z))\|_\infty \leq \frac{3d}{2}C^3_\infty \sup_z \|\Delta(z)\|_{\infty}^2 \]

as claimed. \( \Box \)
Lemma 8. Under conditions of Theorem 3, if
\[
 r := 2C_3 (\sup_z \| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty + n^{-1/2} \lambda) \leq \min \left( \frac{1}{3C_\infty d'}, \frac{1}{3C_3 C_\infty^3 d} \right),
\]
then
\[
 \sup_z \| \Delta(z) \|_\infty \leq r.
\]

Proof. \( \forall z \in [0, 1] \), define the gradient mapping function as
\[
 G(\Omega_S(z)) = -\left( \Omega^{-1}(z) \right)_S + \hat{\Sigma}(z) + \lambda \tilde{Z}_S(z).
\]
Construct the following function \( F : \mathbb{R}^{|\mathcal{S}| \times n} \to \mathbb{R}^{|\mathcal{S}| \times n} \) via
\[
 F(\hat{\Delta}_S(z)) := -(J_{SS}(z))^{-1} G(\Omega^*_S(z) + \hat{\Delta}_S(z)) + \hat{\Delta}_S(z)
\]
By construction, \( F(\Delta_S(z)) = \Delta_S(z) \) since \( G(\Omega_S(z)) = (\Omega_S(z) + \Delta_S(z)) = 0 \) based on the zero-gradient condition. Define the ball \( \mathcal{B}(r) \) as
\[
 \mathcal{B}(r) := \{ A : \| A \|_\infty \leq r \}.
\]
The next thing we need to shows is \( F(\mathcal{B}(r)) \subseteq \mathcal{B}(r) \), thus by fix point theorem, we can conclude \( \sup_z \| \Delta(z) \|_\infty \leq r \). For any \( \hat{\Delta}_S(z) \in \mathcal{B}(r) \), first decompose \( F(\hat{\Delta}_S(z)) \) as
\[
 F(\hat{\Delta}_S(z)) = (J_{SS}(z))^{-1} R(\hat{\Delta}_S(z))_S - (J_{SS}(z))^{-1} (\hat{\Sigma}_S(z) - \Sigma^*_S(z) + \lambda \tilde{Z}_S(z)),
\]
and bound the two terms of above equation separately: for the first term, we have
\[
 \|(J_{SS}(z))^{-1} R(\hat{\Delta}_S(z))_S\|_\infty \leq \|(J_{SS}(z))^{-1}\|_\infty \| R(\Delta(z))_S \|_\infty \leq C_3 \| R(\hat{\Delta}(z)) \|_\infty.
\]
Since we assume \( \hat{\Delta}_S(z) \in \mathcal{B}(r) \), applying Lemma 7, we get
\[
 \|(J_{SS}(z))^{-1} R(\hat{\Delta}_S(z))_S\|_\infty \leq \|(J_{SS}(z))^{-1}\|_{\infty, \infty} \| R(\hat{\Delta}(z)) \|_\infty \leq C_3 \| R(\hat{\Delta}(z)) \|_\infty
\]
\[
 \leq C_3 \frac{3d}{2} C_\infty^2 \| \hat{\Delta}(z) \|_\infty^2
\]
\[
 \leq C_3 \frac{3d}{2} C_\infty^3 r^2
\]
\[
 \leq C_3 \frac{3d}{2} C_\infty^3 \frac{1}{3C_3 C_\infty^3} d^r
\]
\[
 \leq \frac{r}{2}.
\]
For the second term, we have
\[
\| (J_{SS}(z))^{-1} (\hat{\Sigma}_S(z) - \Sigma^*_S(z) + \lambda \tilde{Z}_S(z)) \|_\infty \\
\leq C_3 \| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty + C_3 \lambda \| \tilde{Z}_S(z) \|_\infty \\
\leq C_3 (\| \hat{\Sigma}(z) - \Sigma^*(z) \|_\infty + \pi^{-1/2} \lambda) \\
= \frac{r}{\hat{2}}.
\]

Substituting (21) and (22) into (20) we get \( F(\hat{\Delta}_S(z)) \in B(r) \), which concludes the proof.

\[ \square \]

**Lemma 9.** Under conditions of Theorem 3, if
\[
\| \mathcal{H}^n(\hat{\Sigma}(z) - \Sigma^*(z)) \|_\infty \leq \frac{n^{-1/2} \alpha \lambda}{8} \quad \text{and} \quad \| \mathcal{H}^n(R(\Delta(z))) \|_\infty \leq \frac{n^{-1/2} \alpha \lambda}{8},
\]
then \( \tilde{Z}_{sc} \) constructed in Step 3 of the primal-dual witness process satisfies
\[
\| \mathcal{H}^n(\tilde{Z}_{sc}(z)) \|_\infty \leq n^{-1/2}.
\]

**Proof.** Based on the construction \( \forall i \in [n] \), \( \tilde{Z}_{uv}(z^i) = \frac{1}{n} (-\hat{\Sigma}(z^i))_{uv} + \hat{\Omega}_{uv}^{-1}(z^i) \). Substituting with the definition of \( \Delta(z) \) and \( R(\Delta(z)) \), we have \( \forall i \in [n] \)
\[
\Omega^*(z^i)^{-1} \Delta(z^i) \Omega^*(z^i)^{-1} + \hat{\Sigma}(z^i) - \Omega^*(z^i)^{-1} - R(\Delta(z^i)) + \lambda \tilde{Z}(z^i) = 0
\]

(23)

Since
\[
\vec(\Omega^*(z^i)^{-1} \Delta(z^i) \Omega^*(z^i)^{-1}) = (\Omega^*(z^i)^{-1} \otimes \Omega^*(z^i)^{-1}) \Delta(z^i) = \mathcal{I}(z^i)\Delta(z^i)
\]

thus we can re-write Equation (23) as two blocks of linear equations
\[
J_{SS}(z^i)\hat{\Delta}(z^i) + \tilde{\Sigma}(z^i) - \hat{\Omega}(z^i)_{SS}^{-1} - \hat{R}(\Delta(z^i))_{SS} + \lambda \tilde{Z}(z^i) = 0
\]

(24)

\[
J_{sc}(z^i)\hat{\Delta}(z^i) + \tilde{\Sigma}(z^i) - \hat{\Omega}(z^i)_{sc}^{-1} - \hat{R}(\Delta(z^i))_{sc} + \lambda \tilde{Z}(z^i) = 0
\]

(25)

Solving \( \hat{\Delta}(z^i) \) using Equation (24), we get
\[
\hat{\Delta}(z^i) = J_{SS}^{-1}(z^i)[R(\Delta(z^i))_{SS} - \tilde{\Sigma}(z^i) + \hat{\Omega}(z^i)_{SS}^{-1} - \lambda \tilde{Z}(z^i)]
\]

substituting above to Equation (25), we can solve \( \tilde{Z}_{sc}(z^i) \) as
\[
\tilde{Z}_{sc}(z^i) = -\frac{1}{\lambda} J_{sc}(z^i)\hat{\Delta}(z^i) + \frac{1}{\lambda} \hat{R}(\Delta(z^i))_{sc} - \frac{1}{\lambda} (\hat{\Sigma}_{sc}(z^i) - \hat{\Omega}(z^i)_{sc}^{-1})
\]
\[
= -\frac{1}{\lambda} J_{sc}(z^i)J_{SS}^{-1}(z^i)[R(\Delta(z^i))_{SS} - \tilde{\Sigma}(z^i) + \hat{\Omega}(z^i)_{SS}^{-1} - \lambda \tilde{Z}(z^i)]
\]
\[
+ \frac{1}{\lambda} \hat{R}(\Delta(z^i))_{sc} - \frac{1}{\lambda} (\hat{\Sigma}_{sc}(z^i) - \hat{\Omega}(z^i)_{sc}^{-1})
\]
Thus

\[
\|\mathcal{G}^n(\bar{Z}_{Sc}(z))\|_\infty \leq \frac{1}{\lambda} \|\mathcal{G}^n(J_{ScS}(z)J_{SS}^{-1}(z)[R(\Delta(z))_S - \hat{\Sigma}_S(z) + \Omega^*(z)\hat{S}^{-1} - \lambda\bar{Z}_S(z)])[I\tilde{S} - \hat{\Sigma}_S(z) - \hat{\Sigma}_S^{-1} + \Omega^* - \lambda\bar{Z}_S(z)]\|_\infty
\]
\[
+ \frac{1}{\lambda} \|R(\Delta(z))_S\|_\infty + \frac{1}{\lambda} \|\mathcal{G}^n(\bar{Z}_{Sc}(z^*) - \Omega^*(z)\hat{S}^{-1})\|_\infty
\]
\[
\leq \frac{1}{\lambda} \|\mathcal{G}^n(J_{ScS}(z)(J_{SS}(z))^{-1})\|_{\infty,\infty}
\cdot (\|\mathcal{G}^n(\hat{\Sigma}(z) - \Sigma^*(z))\|_\infty + \|\mathcal{G}^n(R(\Delta(z)))\|_\infty)
\]
\[
+ \|\mathcal{G}^n(J_{ScS}(z)(J_{SS}(z))^{-1})\|_{\infty,\infty}\|\mathcal{G}^n(\bar{Z}_S(z))\|_\infty
\]
\[
+ \frac{1}{\lambda} \frac{n^{-1/2} \alpha \lambda}{8} + \frac{1}{\lambda} \frac{n^{-1/2} \alpha \lambda}{8}
\]
\[
\leq \frac{(1 - \alpha) n^{-1/2} \alpha \lambda}{4} + n^{-1/2}(1 - \alpha) + \frac{n^{-1/2} \alpha}{4}
\]
\[
\leq \frac{n^{-1/2} \alpha}{2} + n^{-1/2}(1 - \alpha)
\]
\[
< n^{-1/2}.
\]
3.3 ASYMPTOTIC NORMALITY

In this section we describe how to construct point-wise confidence intervals for individual elements of $\Omega^*(z)$. Our approach is based on the idea presented in [20] where confidence intervals are constructed for individual elements of the inverse covariance matrix based on the de-sparsified estimator and its asymptotic Normality.

Let $\{\hat{\Omega}(z^i)\}_{i \in [n]}$ be the minimizer of (3). Based on this minimizer, we construct the following de-sparsified estimator

$$
\hat{\Omega}(z) = \text{vec} \left( \hat{\Omega}(z) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \hat{\Sigma}(z) - \hat{\Omega}(z)^{-1} \right) \right), \quad z \in [0, 1]. \tag{26}
$$

The second term in (26) is to remove the bias in $\hat{\Omega}(z)$ introduced by the sparsity inducing penalty. We have the following theorem about $\hat{\Omega}(z)$.

**Theorem 10.** Suppose that the conditions of Theorem 3 are satisfied. For a fixed $z \in [0, 1]$, we have

$$
\frac{n^{2/5} \sqrt{\hat{f}(z) \left( \hat{T}_{uv}(z) - \Omega^*_{uv}(z) \right)}}{\sqrt{\left( \hat{\Omega}^2_{uv}(z) + \hat{\Omega}_{uu}(z) \hat{\Omega}_{vv}(z) \right) \int_{-\infty}^{\infty} K^2(t) dt}} = N_{uv}(z) + \text{bias} + o_p(1)
$$

where $N_{uv}(z)$ converges in distribution to $N(0, 1)$, and $\text{bias} = O(1)$.

**Proof.** We start the proof by stating the following relationship

$$
\hat{T}_{uv}(z) = \Omega^*_{uv}(z) - e_{uv}^T \Omega^*(z) \otimes \Omega^*(z) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^*(z) \right) + o_p(n^{-3/8})
$$

which holds for all $z \in \{z^i : i \in [n]\}$ and $u, v \in [p]$. This relationship is established in Lemma 11 at the end of the section. Furthermore

$$
n^{3/8} \left( \Omega^*(z) \otimes \Omega^*(z) \right) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^*(z) \right) = n^{3/8} \left( \Omega^*(z) \otimes \Omega^*(z) \right) \text{vec} \left( \hat{\Sigma}(z) - E\hat{\Sigma}(z) \right) + n^{3/8} \left( \Omega^*(z) \otimes \Omega^*(z) \right) \text{vec} \left( E\hat{\Sigma}(z) - \Sigma^*(z) \right)
$$

From (14) we have $\|E\hat{\Sigma}(z) - \Sigma^*(z)\|_\infty = O(h^2) = O(n^{-1/2})$. Combined with the fact that $\Omega^*(z)$ is sparse, we have

$$
\left\| n^{3/8} \left( \Omega^*(z) \otimes \Omega^*(z) \right) \text{vec} \left( E\hat{\Sigma}(z) - \Sigma^*(z) \right) \right\|_\infty = o(1).
$$

Putting these results together, we have

$$
n^{3/8} \left( \hat{T}_{uv}(z) - \Omega^*_{uv}(z) \right) = -e_{uv}^T \Omega^*(z) \otimes \Omega^*(z) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^*(z) \right) + o_p(1).
$$

Define the variance normalization function $V(z) : [0, 1] \rightarrow R^{p \times p}$ as

$$
V_{uv}(z) = \frac{\hat{f}(z)}{\sqrt{\left( \hat{\Omega}^2_{uv}(z) + \hat{\Omega}_{uu}(z) \hat{\Omega}_{vv}(z) \right) \int_{-\infty}^{\infty} K^2(t) dt}}.
$$
We proceed to show that
\[ n^{3/8} V_{uv}(z) e_{uv}^T (\Omega^*(z) \otimes \Omega^*(z)) \text{vec} \left( \hat{\Sigma}(z) - E \hat{\Sigma}(z) \right) \]
converges in distribution to \( N(0,1) \). The result will then follow from an application of Lemma 11 given below.

We first consider the elements of \( n^{3/8} (\Omega^*(z) \otimes \Omega^*(z)) \text{vec} \left( \hat{\Sigma}(z) - E \hat{\Sigma}(z) \right) \) and show that they are asymptotically Normal with bounded variance. Fix \( u, v, \) and \( z \), and define the random variable \( J_{uv}(z) \) as
\[ J_{uv}(z) = \sum_{i \in [n]} L^i_{uv}(z), \]
where
\[ L^i_{uv}(z) = (\Omega^*(z) (w^i(z) x^i(z)^T - E w^i(z) x^i(z)^T) \Omega^*(z))_{uv}. \]
We have that \( E [L^i_{uv}(z)] = 0 \) and \( \text{var} [L^i_{uv}(z)] < \infty \). Let
\[ s_n^2 = \sum_i \text{var} [L^i_{uv}(z)] = n^{3/4} \text{var} [J_{uv}(z)]. \]

In order to establish asymptotic Normality, we proceed to show that the Lindeberg condition is satisfied, that is, for all \( \varepsilon > 0 \)
\[ \lim_{n \to \infty} \frac{1}{s_n^2} \sum_i E \left[ (L^i_{uv}(z))^2 I(|L^i_{uv}(z)| > \varepsilon s_n) \right] = 0. \]  
(27)

Since \( L^i_{uv}(z) \) is a sub-exponential random variable, (27) is easily verified. We only need to show
\[ \lim_{n \to \infty} E \left[ (L^i_{uv}(z))^2 I(|L^i_{uv}(z)| > \varepsilon s_n) \right] = 0. \]

Using Fubini’s theorem, we have
\[ E \left[ (L^i_{uv}(z))^2 I(|L^i_{uv}(z)| > \varepsilon s_n) \right] = 2 \int_{\varepsilon s_n}^{\infty} t \text{pr} [L^i_{uv}(z) > t] \, dt + \varepsilon^2 s_n^2 \text{pr} [L^i_{uv}(z) > \varepsilon s_n]. \]
Therefore, we will prove that
\[ \lim_{n \to \infty} \varepsilon^2 s_n^2 \text{pr} [L^i_{uv}(z) > \varepsilon s_n] = 0 \]
and
\[ \lim_{n \to \infty} \int_{\varepsilon s_n}^{\infty} t \text{pr} [L^i_{uv}(z) > t] \, dt = 0. \]

Since \( |L^i_{uv}(z)| \) is a sub-exponential random variable, we have
\[ \text{pr} [L^i_{uv}(z) > \varepsilon s_n] \leq c_1 \exp(-c_2 \varepsilon s_n) \]
for some constant $c_1, c_2$. Note that $\text{var}\left[ L_{uv}^i(z) \right]$ is both upper and lower bounded, thus $s_n \propto n^{3/8}$. We have

$$\lim_{n \to \infty} \varepsilon^2 s_n \Pr \left[ \left| L_{uv}^i(z) \right| > \varepsilon s_n \right] \leq \lim_{n \to \infty} \varepsilon^2 n^{3/4} c_1 \exp(-c_2 \varepsilon n^{3/8}) = 0$$

and

$$\lim_{n \to \infty} \int_{\varepsilon s_n}^{\infty} \Pr \left[ \left| L_{uv}^i(z) \right| > u \right] du \leq \lim_{n \to \infty} \int_{\varepsilon n^{3/8}}^{\infty} uc_1 \exp(-c_2 \varepsilon n^{3/8}) du = 0.$$ 

This verifies the Lindenberg condition in (27) and we have that

$$\frac{J_{uv}(z)}{\sqrt{\text{var}[J_{uv}(z)]}} \xrightarrow{D} N(0,1).$$

The proof will be complete once we establish

$$nhV_{uv}^2(z) \text{var}[J_{uv}(z)] \xrightarrow{p} 1.$$ 

Note that for every $i \in [n]$ such that $K(z^i - z) \neq 0$ we have $\|\Omega^i(z) - \Omega^i(z^i)\|_\infty = 0(h^2)$ and $\|\Omega^i(z) - \Omega^i(z^i)\|_\infty = 0(h^2)$. Therefore,

$$\text{var}[J_{uv}(z)] = \text{var} \left[ \sum_{i \in [n]} e_u^T \Omega^i(z) \left( w^i(z)x^i x^iT - E \left[ w^i(z)x^i x^iT \right] \right) \Omega^i(z)e_v \right]$$

$$= \frac{1}{(nhf(z))^2} \sum_{i \in [n]} \left( E \left[ \left( K(h^{-1}(z^i - z))(\Omega^i(z)x^i)_u(\Omega^i(z)x^i)_v \right) \right. 
\left. - E \left[ K(h^{-1}(z^i - z))(\Omega^i(z)x^i)_u(\Omega^i(z)x^i)_v \right] \right)^2 \right] + o(1)$$

$$= \frac{\Omega_{uu}^i(z)\Omega_{vv}^i(z) + \Omega_{uv}^i(z)^2}{nhf(z)} \int_{-\infty}^{\infty} K^2(t)dt + o(1).$$

From here, the proof follows since $\hat{\Omega}_{uu}^i(z)\hat{\Omega}_{vv}^i(z) = \Omega_{uu}^i(z)\Omega_{vv}^i(z) + o_P(1)$, $\hat{\Omega}_{uv}^i(z) = \Omega_{uv}^i + o_P(1)$ and $\hat{f}(z) = f(z) + o_P(1)$.

**Lemma 11.** Assume that the conditions of Theorem 10 are satisfied. Furthermore, suppose that the bandwidth parameter $h \propto n^{-1/4}$. We have for all $z \in \{z^i : i \in [n]\}$ that

$$\left\| n^{3/8} \left( \hat{T}(z) - \text{vec}(\Omega^i(z)) + (\Omega^i(z) \otimes \Omega^i(z)) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^i(z) \right) \right) \right\|_\infty = o_P(1).$$

**Proof.** First by re-arranging the terms, we have

$$\hat{T}(z) - \text{vec}(\Omega^i(z)) + (\Omega^i(z) \otimes \Omega^i(z)) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^i(z) \right)$$

$$= \text{vec} \left( \hat{\Omega}(z) - \Omega^i(z) \right) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \Sigma^i(z) - \hat{\Omega}(z)^{-1} \right)$$

$$+ \left( \Omega^i(z) \otimes \Omega^i(z) - \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^i(z) \right)$$

$$= v^T \hat{h}(z) + o_P(1).$$

where $v^T \hat{h}(z) = \text{vec} \left( \hat{\Omega}(z) - \Omega^i(z) \right) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \Sigma^i(z) - \hat{\Omega}(z)^{-1} \right)$. We have

$$\text{vec} \left( \hat{\Omega}(z) - \Omega^i(z) \right) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \Sigma^i(z) - \hat{\Omega}(z)^{-1} \right)$$

$$= o_P(1)$$

since $\hat{\Omega}(z)$ is a consistent estimator of $\Omega^i(z)$. Therefore, we have

$$\left\| n^{3/8} \left( \hat{T}(z) - \text{vec}(\Omega^i(z)) + (\Omega^i(z) \otimes \Omega^i(z)) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^i(z) \right) \right) \right\|_\infty = o_P(1).$$

3.3 ASYMPOTIC NORMALITY

31
Let $\Delta(z) = \hat{\Omega}(z) - \Omega^*(z)$, the first term of right hand side in (28) can be bounded by

$$
\left\| \text{vec} \left( \hat{\Omega}(z) - \Omega^*(z) \right) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \Sigma^*(z) - \hat{\Omega}(z)^{-1} \right) \right\|_{\infty}
$$

$$
\leq \left\| \Delta(z) \Omega^*(z)^{-1} \Delta(z) \right\|_{\infty}
$$

(29)

and the second term of right hand side in (28) can be bounded by

$$
\left\| \Omega^*(z) \otimes \Omega^*(z) - \left( \hat{\Omega}(z) \otimes \hat{\Omega}(z) \right) \text{vec} \left( \hat{\Sigma}(z) - \Sigma^*(z) \right) \right\|_{\infty}
$$

$$
\leq \left\| \Omega^*(z) \otimes \Delta(z) + \Delta(z) \otimes \Omega^*(z) + \Delta(z) \otimes \Delta(z) \right\|_{\infty} \left\| \hat{\Sigma}(z) - \Sigma^*(z) \right\|_{\infty}
$$

$$
\leq \left( 2 \left\| \Omega^*(z) \otimes \Delta(z) \right\|_{\infty} + \left\| \Delta(z) \otimes \Delta(z) \right\|_{\infty} \right) \left\| \hat{\Sigma}(z) - \Sigma^*(z) \right\|_{\infty}
$$

(30)

Combining (29) and (30), together with $\|\Delta(z)\|_{\infty} = O(\|z\|^{-3/8}\sqrt{\log p})$ and $\|\hat{\Sigma}(z) - \Sigma^*(z)\|_{\infty} = O(\|z\|^{-3/8}\sqrt{\log p})$ we obtain the conclusion since $[\Omega^*(z)]_{\infty,\infty}$ and $[\Omega^*(z)^{-1}]_{\infty,\infty}$ are upper bounded.

An explicit form for the bias term can be found in the proof. The bias term can be removed by under-smoothing, that is, using a smaller bandwidth. We have the following corollary.
Corollary 12. Suppose the bandwidth satisfies $h \asymp n^{-1/4}$, the penalty $\lambda \asymp n^{1/8} \sqrt{\log p}$, and sample size satisfies $n \gtrsim d^{8/3} (\log p)^{4/3}$. If

$$
\min_{u,v \in S} \sqrt{n^{-1} \sum_{i \in [n]} \Omega_{uv}^*(z_i)} \geq C n^{-3/8} \sqrt{\log p}
$$

for some constant $C$, then

$$
\frac{n^{3/8} \sqrt{\hat{f}(z) \left( \hat{T}_{uv}(z) - \Omega_{uv}^*(z) \right)}}{\sqrt{\left( \hat{\Omega}^2_{uv}(z) + \hat{\Omega}_{uu}(z) \hat{\Omega}_{vv}(z) \right) \int_{-\infty}^{\infty} K^2(t) dt}} = N_{uv}(z) + o_p(1)
$$

where $N_{uv}(z)$ converges in distribution to $N(0, 1)$.

Based on the theorem and its corollary, we can construct $100(1 - \alpha)$% asymptotic confidence intervals $B_{uv}(z)$ for $\Omega_{uv}^*(z)$ as

$$
B_{uv}(z) = \left[ \hat{T}_{uv}(z) - \delta(z, \alpha, n), \hat{T}_{uv}(z) + \delta(z, \alpha, n) \right]
$$

where

$$
\delta(z, \alpha, n) = \Phi^{-1}(1 - \alpha/2) n^{-3/8} \sqrt{\frac{\hat{\Omega}^2_{uv}(z) + \hat{\Omega}_{uu}(z) \hat{\Omega}_{vv}(z)}{\hat{f}(z) \int_{-\infty}^{\infty} K^2(t) dt}}.
$$

We will illustrate finite sample properties of the confidence interval in Section 4.1.
4.1 SIMULATION STUDIES

In this section we present extensive simulation results. First, we report results on recovery of graph structure and precision matrix. Next, we report coverage properties and average length of the confidence intervals. These simulation studies serve to illustrate finite sample performance of our method. We compare against two other methods: 1) a method that ignores the index variable $Z$ and uses the glasso $[62]$ to estimate the underlying conditional independence structure, and 2) a method from Zhou et al. $[68]$ which does not exploit the common structure across the indexing variable. We denote the first procedure as the glasso, the second one as the locally smoothed glasso. For the locally smoothed glasso and our procedure, we use the Epanechnikov kernel defined as

$$K_h(x) = \frac{3}{4} \left(1 - \left(\frac{x}{h}\right)^2\right) 1(|x| \leq h).$$

For our procedure, we solve the optimization problem in (10) at index points $\{0, 0.02, \ldots, 1\}$.

We first explain the experimental setup. Data are generated according to the model in (12), for which we need to specify the conditional independence structure given a graph and the edge values as functions of the index variable. We generate four kinds of graphs, described below.

- Chain graph is a random graph formed by permuting the nodes and connecting them in succession.

- Nearest Neighbor graph is a random graph generated by the procedure described in Li and Gui $[31]$. For each node, a point is drawn uniformly at random on a unit square. Then pairwise distances between nodes are computed and each node is connected to $d = 3$ closest neighbors.

- Erdös-Rényi graph is a random graph where each pair of nodes is connected independently of other pairs. We generated a random graph with $2p$ edges and maximum degree of the nodes is set to 5 [see, for example, 36].

- Scale-free graph is a random graph created according to the procedure described in [3]. The network begins with an initial 3 nodes clique. New nodes are added to the graph one at a time.
and each new node is connected to one of the existing nodes with a probability proportional to the node’s current degree. Formally, the probability $p_i$ that the new node is connected to an existing node $i$ is $p_i = \frac{d_i}{\sum d_i}$ where $d_i$ is the degree of node $i$.

Given the graph, we generate a precision matrix with the sparsity pattern adhering to the graph structure, that is, $\Omega^*_uv(z) \equiv 0$ if $(u,v) \notin S$. Non-zero components of $\Omega^*(z)$ were generated in one of the following ways:

- **Exponential function.** We generate $\Omega^*_uv(z) = \exp(z) - \exp(c_{uv})$ where $c_{uv} \sim \text{Uniform}(0,1)$.

- **Linear function.** We generate $\Omega^*_uv(z) = 2z - c_{uv}$ where $c_{uv} \sim \text{Uniform}(0,1)$ is a random offset.

- **Sin function.** We generate $\Omega^*_uv(z) = \sin(2\pi z + c_{uv})$ where $c_{uv} \sim \text{Uniform}(0,1)$ is a random offset.

To ensure the positive-definiteness of $\Omega^*(z)$ we add $\delta(z)I$ to $\Omega^*(z)$, so that the minimum eigenvalue of $\Omega^*(z)$ is at least 0.5.

In the simulation studies, we change the sample size $n$ and the number of variables $p$ as:i) $n = 500, p = 20$, ii) $n = 500, p = 50$, and iii) $n = 500, p = 100$. Once the model is fixed, we generate 10 independent data sets according to the model and report the average performance over these 10 runs. Implemented in MATLAB, run on a laptop with 2.2GHz CPU, the algorithm took 146.56 seconds to get a solution path (40 $\lambda$s) when the number of knots is set to 50 when $n = 500, p = 50$ (On a Chain graph with sin functions), took 461.47 seconds when $n = 500, p = 100$. Thus the average CPU running time for $p = 50$ is 3.66s, and 11.53s for $p = 100$.

Figure 2 shows the FP-TP (false positive-true positive) curves for different estimation methods when the non-zero components of $\Omega^*(z)$ are generated as exponential functions. For point smoothing glasso, we report the results for graph estimated using $z = 0.5$. The number of false positive and true positive edges selected are defined as

$$\text{FP}(\hat{S}, S) = |\hat{S} \setminus S| \quad \text{and} \quad \text{TP}(\hat{S}, S) = |\hat{S} \cap S|,$$

where $\hat{S}$ is the set of estimated edges. Each row in the figure corresponds to a different graph type, while columns correspond to different sample and problem sizes. Figures 3 and 4 similarly show Precision-Recall curves for linear and sin component functions. Results are further summarized in Figure 5 which shows $F_1$ score for the graph estimation task, given as the harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$
where the precision and recall of the estimated graph are given as

\[
\text{Precision}(\hat{S}, S) = 1 - \frac{|\hat{S} \setminus S|}{|S|} \quad \text{and} \quad \text{Recall}(\hat{S}, S) = 1 - \frac{|S \setminus \hat{S}|}{|S|}.
\]

Finally, Figure 7 shows the performance of different procedures in terms of Frobenius norm error

\[
\sum_{i \in [n]} \| \hat{\Omega}(z^i) - \Omega^*(z^i) \|_F.
\]

From these results, several observations could be drawn:

- For precision matrix functions: exponential, linear and sin, our procedure performs significantly better than naive graphical lasso in terms of FP-TP curve, \(F_1\), and Frobenius norm error. For example, under the sin precision function case, for the chain and NN graph, our procedure can recover the graph perfectly while the glasso totally fails. When comparing the glasso and point smoothing glasso, we found that glasso performs better for exponential function, while point smoothing glasso performs better for sin function.

- When comparing results for different types of graphs, we found that the chain and NN graphs are relatively easy to recover, and Scale-free networks are the hardest to recover due to presence of high-degree nodes. Recall that for chain and NN graph the maximum degree is small, for Erdös-Rényi graph the maximum degree is larger (\(d \approx 5\)), and for scale-free graph the maximum degree is the largest.

As the theory suggests, \(\Omega(d^{5/2}(\log p)^{5/4})\) sample is needed for consistent graph recovery. To empirically study this behavior, we vary the sample size \(n\) as \(n = C \cdot d^{5/2}(\log p)^{5/4}\) by varying \(C\), and plot the hamming distance between the true and estimated graph against \(C\), and the theory suggest that the hamming distance curves are going to reach zero for different \(p\) at a same point. Figure 8 shows the plot, and we can see that generally this is true for all the graphs and precision matrix function we studied here.

In the next set of experiments our purpose is to verify the asymptotic normality of our de-biased estimator. We follow the simulations done in [20]. Using Theorem 10 we construct \(100(1 - \alpha)\)% confidence intervals as

\[
B_{uv}(z^{(i)}) := [\hat{T}_{uv}(z^{(i)}) - \delta(\alpha, n), \hat{T}_{uv}(z^{(i)}) + \delta(z^{(i)}, \alpha, n)],
\]

where

\[
n^{-3/8}\delta(z^{(i)}, \alpha, n) := \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\hat{\Omega}^2(z^{(i)}) + \hat{\Omega}_{uv}(z^{(i)})\hat{\Omega}_{mv}(z^{(i)})}{f(z)}} \int_{-\infty}^{\infty} K^2(t) dt
\]

For the simulation studies, we follow the same graph and precision construction described earlier, set \(\alpha = 0.025\), and set the bandwidth
Figure 2: Precision-Recall curves for exponential function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.
Figure 3: Precision-Recall curves for linear function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.
Figure 4: Precision-Recall curves for random sin function, from top to bottom: chain graph, NN graph, Erdős-Rényi graph, Scale-free graph.
Figure 5: Average F1-scores of various methods.

Figure 6: Case where the underlying precision matrix is static.
Figure 7: Average Frobenius norm errors of various methods.
Figure 8: Average hamming distance plotted against the re-scaled sample size.

Figure 9: Illustration of the estimated graph using various approaches, on a chain graph with linear precision function.
Table 1: Confidence interval, for exponential function precision

<table>
<thead>
<tr>
<th>Graph(s,p)</th>
<th>Exp</th>
<th>Sin</th>
<th>Avg</th>
<th>Avglength</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>S*</td>
<td>(u,v) ∈ S</td>
<td>(u,v) ∈ S*</td>
</tr>
<tr>
<td>Chain(100,20)</td>
<td>0.944</td>
<td>0.973</td>
<td>0.964</td>
<td>0.968</td>
</tr>
<tr>
<td>Chain(100,50)</td>
<td>0.935</td>
<td>0.974</td>
<td>0.956</td>
<td>0.976</td>
</tr>
<tr>
<td>Chain(100,100)</td>
<td>0.947</td>
<td>0.981</td>
<td>0.949</td>
<td>1.000</td>
</tr>
<tr>
<td>NN(500,20)</td>
<td>0.943</td>
<td>0.974</td>
<td>0.958</td>
<td>0.975</td>
</tr>
<tr>
<td>NN(500,50)</td>
<td>0.942</td>
<td>0.979</td>
<td>0.950</td>
<td>0.989</td>
</tr>
<tr>
<td>NN(500,100)</td>
<td>0.944</td>
<td>0.983</td>
<td>0.947</td>
<td>0.988</td>
</tr>
<tr>
<td>Erdős-Rényi(500,20)</td>
<td>0.954</td>
<td>0.977</td>
<td>0.960</td>
<td>0.988</td>
</tr>
<tr>
<td>Erdős-Rényi(500,50)</td>
<td>0.949</td>
<td>0.983</td>
<td>0.946</td>
<td>0.992</td>
</tr>
<tr>
<td>Erdős-Rényi(500,100)</td>
<td>0.949</td>
<td>0.987</td>
<td>0.985</td>
<td>0.985</td>
</tr>
<tr>
<td>Scale-free(500,20)</td>
<td>0.945</td>
<td>0.980</td>
<td>0.954</td>
<td>0.916</td>
</tr>
<tr>
<td>Scale-free(500,50)</td>
<td>0.956</td>
<td>0.988</td>
<td>0.930</td>
<td>0.963</td>
</tr>
<tr>
<td>Scale-free(500,100)</td>
<td>0.968</td>
<td>0.990</td>
<td>0.986</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 2: Confidence interval, for linear function precision

<table>
<thead>
<tr>
<th>Graph(s,p)</th>
<th>Lin</th>
<th>Sin</th>
<th>Avg</th>
<th>Avglength</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>S*</td>
<td>(u,v) ∈ S</td>
<td>(u,v) ∈ S*</td>
</tr>
<tr>
<td>Chain(100,20)</td>
<td>0.922</td>
<td>0.966</td>
<td>0.957</td>
<td>0.993</td>
</tr>
<tr>
<td>Chain(100,50)</td>
<td>0.931</td>
<td>0.986</td>
<td>0.940</td>
<td>0.975</td>
</tr>
<tr>
<td>Chain(100,100)</td>
<td>0.930</td>
<td>0.987</td>
<td>0.976</td>
<td>0.989</td>
</tr>
<tr>
<td>NN(500,20)</td>
<td>0.943</td>
<td>0.966</td>
<td>0.931</td>
<td>0.956</td>
</tr>
<tr>
<td>NN(500,50)</td>
<td>0.922</td>
<td>0.986</td>
<td>0.956</td>
<td>0.972</td>
</tr>
<tr>
<td>NN(500,100)</td>
<td>0.910</td>
<td>0.990</td>
<td>0.959</td>
<td>1.000</td>
</tr>
<tr>
<td>Erdős-Rényi(500,20)</td>
<td>0.909</td>
<td>0.952</td>
<td>0.904</td>
<td>0.948</td>
</tr>
<tr>
<td>Erdős-Rényi(500,50)</td>
<td>0.930</td>
<td>0.988</td>
<td>0.945</td>
<td>0.981</td>
</tr>
<tr>
<td>Erdős-Rényi(500,100)</td>
<td>0.931</td>
<td>0.982</td>
<td>0.973</td>
<td>0.995</td>
</tr>
<tr>
<td>Scale-free(500,20)</td>
<td>0.945</td>
<td>0.988</td>
<td>0.922</td>
<td>0.998</td>
</tr>
<tr>
<td>Scale-free(500,50)</td>
<td>0.949</td>
<td>0.991</td>
<td>0.895</td>
<td>0.931</td>
</tr>
<tr>
<td>Scale-free(500,100)</td>
<td>0.966</td>
<td>0.991</td>
<td>0.901</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Figure 10: Illustration of the constructed confidence bands by the proposed procedure, on a n = 500, p = 50, chain graph problem.
\( h = n^{-1/4} \) and the regularization parameter \( \lambda = n^{-3/8} \sqrt{\log(p)} \) as the theory suggests. We fix the model and generate different data sets 100 times, test the empirical frequency that \( \Omega^*(z^1) \) is covered by the constructed confidence intervals:

\[
\hat{\alpha}_{uv} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{100} \sum_{l=1}^{100} \mathbb{I}_{\{\Omega^*_{uv}(z^{(l)}) \in B_{uv}(z^{(l)})\}}
\]

where \( B_{uv}(z^{(l)}) \) is the confidence intervals built based on t-th data set. And the average coverage on support \( S \) is defined as

\[
\text{Avgcov}_S = \frac{1}{|S|} \sum_{(u,v) \in S} \hat{\alpha}_{uv}.
\]

Likewise, we can define the \( \text{Avgcov}_{Sc} \), as well the average length of the confidence intervals \( \text{Avglength}_S \), \( \text{Avglength}_{Sc} \).

Table 1, 2, 3 shows coverage properties for exponential, linear, and sin component functions, respectively. We make the following observations:

- When looking at the Avgcov, confidence intervals cover \( \Omega^*(z^1) \) reasonably well, no matter what kinds of graph, for what kind of precision matrix function, which verifies the theories of asymptotic normality.

- When looking at the Avglength, and comparing different types of graph structures, we can see that the Avglength for chain and NN graphs are much shorter than the that of Erdös-Rényi and scale-free graph, because generally the later are harder to estimate.

Figure 10 illustrates confidence bands constructed by our procedure on a \( n = 500, p = 50 \) chain graph problem.
4.2 ILLUSTRATIVE APPLICATIONS TO REAL DATA

In this section we apply our procedure on the stock price data, which is available in the “huge” package [67]. It contains 1258 closing prices of 452 stocks between January 1, 2003 and January 1, 2008. Instead of using the raw closing prices, we use logarithm of the ratio of prices at times $t$ and $t-1$. We further normalize the data for each stock to have mean 0 and variance 1. The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Telecommunications Services, Materials, and Utilities, each sector has 6 to 70 stocks.

We apply the glasso and our procedure to estimate associations between the stocks. In the estimated graph each node corresponds to a stock. An edge is put between two stocks if the corresponding element in the estimated inverse covariance matrix is not zero. When applying our procedure we consider the following as index variables: i) date, normalized to $[0, 1]$, ii) the Crude Oil price¹ (WTI - Cushing, Oklahoma), and iii) USD/EUR exchange rate². Figure 11 shows the estimated graphs with colors representing different sectors. We can see that the graphs learned by CCS $\ell_1/\ell_2$ give more clear cluster structure than the regular glasso graph.

We also consider a quantitative measure of the estimation performance. Since there is no true graph for this data, we use the Kullback-Leibler loss, as in Yuan and Lin [62]. First we divide the whole data set $[n]$ into $K$ folds: $\{D_1, D_2, ..., D_K\}$, then the $K$-fold cross-validation negative log-likelihood is

$$\sum_{k \in [K]} \sum_{i \in D_k} -\log \det(\hat{\Omega}^{[n]/D_k}(\tilde{z}^i)) + \text{tr}(\hat{\Omega}^{[n]/D_k}(\tilde{z}^i)(\tilde{x}^i)^T),$$

where $\hat{\Omega}^{[n]/D_k}(z)$ is learned from the training data set $[n]/D_k$ and $\{\tilde{x}^i, \tilde{z}^i\}_{i \in D_k}$ is a test set. For the regular glasso procedure, since it only estimate a single precision matrix, the above defined loss reduces to

$$- \sum_{k \in [K]} |D_k| \log \det(\hat{\Omega}^{[n]/D_k}) + \text{tr}(\hat{\Omega}^{[n]/D_k} \hat{\Sigma}_{D_k})$$

where $\hat{\Omega}^{[n]/D_k}$ is learned from the training data $[n]/D_k$, $\hat{\Sigma}_{D_k}$ is the sample covariance matrix from the testing data $D_k$.

Figure 12 reports 5-fold cross-validation score. We can see that for all index variable we considered here, our procedure improves over the regular glasso.

We also study the estimated edge function and the constructed confidence intervals. Figure 13 demonstrated the constructed interval about the edge: CVS versus Walgreens. We could observe that

¹ http://www.eia.gov/dnav/pet/pet_pri_spt_s1_d.htm
² http://www.oanda.com/currency/historical-rates/
4.2 ILLUSTRATIVE APPLICATIONS TO REAL DATA

Figure 11: Graphical estimation on S&P 500 stock data.

Figure 12: Negative likelihood score versus sparsity.
there are two jumps in the function: one in early 2004 and one in mid 2006. These might be the reaction of two major events about these two companies: in April 2004, CVS purchased Eckerd 3; in July 2006, Walgreens changed its CEO 4. Figure 14 showed more illustration.

Figure 13: (A): Edge weight as time changes: CVS vs Walgreens. (B)(C): Two examples of non-edges.

Figure 14: Top row: edge weight as oil price changes: Exxon Mobil Corp with its major neighbors; bottom row: edge weight as EUR/USD changes: Goldman Sachs Group with its major neighbors.

3 http://en.wikipedia.org/wiki/CVS_Pharmacy
4 http://en.wikipedia.org/wiki/Walgreens
CONCLUSIONS

5.1 CONCLUSIONS AND FUTURE WORK

We develop a new nonparametric method for high-dimensional conditional covariance selection. The method is developed under the assumption that the graph structure does not change with sample size. This assumption restricts the class of models and allows for better estimation when the sample size is small. Our theoretical results and simulation studies demonstrate this point. Furthermore, we develop point-wise confidence intervals for the individual elements in the model. These are useful for practitioners interested in making inferences and testing hypothesis about complex systems under investigation.


