Sparse Estimation of Conditional Graphical Models With Application to Gene Networks

Bing LI, Hyonho CHUN, and Hongyu ZHAO

In many applications the graph structure in a network arises from two sources: intrinsic connections and connections due to external effects. We introduce a sparse estimation procedure for graphical models that is capable of isolating the intrinsic connections by removing the external effects. Technically, this is formulated as a *conditional* graphical model, in which the external effects are modeled as predictors, and the graph is determined by the conditional precision matrix. We introduce two sparse estimators of this matrix using the reproduced kernel Hilbert space combined with lasso and adaptive lasso. We establish the sparsity, variable selection consistency, oracle property, and the asymptotic distributions of the proposed estimators. We also develop their convergence rate when the dimension of the conditional precision matrix goes to infinity. The methods are compared with sparse estimators for unconditional graphical models, and with the constrained maximum likelihood estimate that assumes a known graph structure. The methods are applied to a genetic data set to construct a gene network conditioning on single-nucleotide polymorphisms.

KEY WORDS: Conditional random field; Gaussian graphical models; Lasso and adaptive lasso; Oracle property; Reproducing kernel Hilbert space; Sparsity; Sparsistency; von Mises expansion.

1. INTRODUCTION

Sparse estimation of the Gaussian graphical models has undergone intense development during the recent years, partly due to their wide applications in such fields as information retrieval and genomics, and partly due to the increasing maturity of statistical theories and techniques surrounding sparse estimation. See, for example, Meinshausen and Buhlmann (2006), Yuan and Lin (2007), Bickel and Levina (2008), Peng et al. (2009), Guo et al. (2009), and Lam and Fan (2009). The precursor of this line of work is the Gaussian graphical model in which the graph structure is assumed to be known; see Dempster (1972) and Lauritzen (1996).

Let $\mathbf{Y} = (Y^1, \ldots, Y^p)^\mathsf{T}$ be a random vector, $\Gamma = \{1, \ldots, p\}$, and $E \subseteq \Gamma \times \Gamma$. Let $\mathcal{G} = (\Gamma, E)$ be the graph with its vertices in Γ and edges in E. We say that $\mathbf{Y}$ follows a Gaussian graphical model (GGM) with respect to $\mathcal{G}$ (Lauritzen 1996) if $\mathbf{Y}$ has a multivariate normal distribution and

$$Y^i \perp\!\!\!\perp Y^j | Y^{-(i,j)}, \ (i, j) \in E^c, \tag{1}$$

where $A \perp\!\!\!\perp B|C$ means A and B are independent given C, and $Y^{-(i,j)}$ denotes the set $\{Y^1, \ldots, Y^p\} \setminus \{Y^i, Y^j\}$. Let ω_{ij} be the (i, j)th entry of $[\mathrm{var}(\mathbf{Y})]^{-1}$. Then, $Y^i \perp\!\!\!\perp Y^j | Y^{-(i,j)}$ if and only if $\omega_{ij} = 0$. Thus, estimating E is equivalent to estimating the set $\{(i, j) \in \Gamma \times \Gamma : \omega_{ij} = 0\}$. Conventional Gaussian graphical models focus on maximum likelihood estimation of ω_{ij} given the knowledge of E. In sparse estimation of graphical models, however, E is itself estimated by sparse regularization, such as lasso and adaptive lasso (Tibshirani 1996; Zou 2006).

The *conditional* graphical model, with which we are concerned, is motivated by the analysis of gene networks and the regulating effects of DNA markers. Let $X^1, \ldots, X^q$ represent the genetic markers at q locations in a genome, and let $Y^1, \ldots, Y^p$ represent the expression levels of p genes. The objective is to infer how the q genetic markers affect the expression levels of the p genes and how these p genes affect each other. Since some markers may have regulating effects on more than one gene, the connections among the genes are of two kinds: the connections due to shared regulation by the same marker, and the innate connections among the genes aside from their shared regulators. In this setting, we are interested in identifying the network of genes after removing the effects from shared regulations by the markers.

The situation is illustrated by Figure 1, in which X represents a single marker, and Y^1, Y^2, Y^3 represent the expressions of three genes. If we consider the marginal distribution of the random vector (Y^1, Y^2, Y^3), then there are two (undirected) edges in the unconditional graphical model: $1 \leftrightarrow 2$ and $2 \leftrightarrow 3$, as represented by the solid and the dotted line segments. However, if we condition on the marker and consider the conditional distribution of $Y^1, Y^2, Y^3|X$, then there is only one (undirected) edge, $2 \leftrightarrow 3$, in the conditional graphical model, as represented by the solid line segment.

In mathematical terms, a conditional graphical model can be represented by

$$Y^i \perp\!\!\!\perp Y^j | \{Y^{-(i,j)}, \mathbf{X}\} \ \text{ for } (i, j) \notin E, \tag{2}$$

where $\mathbf{X} = (X^1, \ldots, X^q)^\mathsf{T}$. Compared with the unconditional graphical model, here we have an additional random vector $\mathbf{X}$, whose effects we would like to remove when constructing the network for $Y^1, \ldots, Y^p$. The conditional graphical model (2) was introduced by Lafferty, McCallum, and Pereira (2001) under the name "conditional random field." However, in that

Bing Li is Professor of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802 (E-mail: *bing@stat.psu.edu*). Bing Li's research was supported in part by NSF grants DMS-0704621, DMS-0806058, and DMS-1106815. Hyonho Chun is Assistant Professor of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907 (E-mail: *chunh@purdue.edu*). Hyonho Chun's research was supported in part by NSF grant DMS-1107025. Hongyu Zhao is Professor of Biostatistics, Yale University, Suite 503, 300 George Street, New Haven, CT 06510 (E-mail: *hongyu.zhao@yale.edu*). Hongyu Zhao's research was supported in part by NSF grants DMS-0714817 and DMS-1106738 and NIH grants R01 GM59507 and P30 DA018343.

Journal of the American Statistical Association
March 2012, Vol. 107, No. 497, Theory and Methods
DOI: 10.1080/01621459.2011.644498

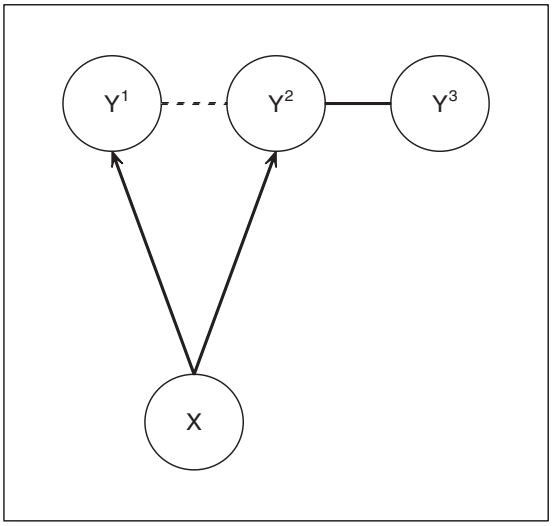

Figure 1. Unconditional and conditional graphical models.

article, the graph $\mathcal{G} = (\Gamma, E)$ was assumed known and maximum likelihood was used to estimate the relevant parameters.

Our strategy for estimating the conditional graphical model (2) is as follows. In the first step, we propose a class of flexible and easy-to-implement initial (nonsparse) estimators for the conditional variance matrix $\mathbf{\Sigma} = \text{var}(\mathbf{Y} \mid \mathbf{X})$, based on a conditional variance operator between the reproducing kernel Hilbert spaces (RKHS) of $\mathbf{X}$ and $\mathbf{Y}$. In the second step, we incorporate the nonsparse estimators with two types of sparse penalties, the lasso and the adaptive lasso, to obtain sparse estimators of the conditional precision matrix $\mathbf{\Sigma}^{-1}$.

We have chosen RKHS as our method to estimate $\mathbf{\Sigma}$ for several reasons. First, it does not require a rigid regression model for $\mathbf{Y}$ versus $\mathbf{X}$. Second, the dimension of $\mathbf{X}$ only appears through a kernel function, so that the dimension of the largest matrix we need to invert is the sample size n, regardless of the dimension of $\mathbf{X}$. This feature is particularly attractive when we deal with a large number of predictors. Finally, RKHS provides a natural mechanism to impose regularization on regression. Here, it is important to realize that our problem involves two kinds of regularization: one for the components of $\text{var}(\mathbf{Y} \mid \mathbf{X})$ and the other for the regression of $\mathbf{Y}$ on $\mathbf{X}$. Considering the nature of our problem, the former must be sparse, but the latter need not be. Indeed, since estimating the regression parameter is not our purpose, it seems more natural to introduce the regularization for regression through RKHS than to try to parameterize the regression and then regularize the parameters.

The rest of the article is organized as follows. In Section 2, we introduce a conditional variance operator in RKHS and describe its relation with the conditional gaussian graphical model. In Section 3, we derive two RKHS-based estimators of conditional variance $\mathbf{\Sigma}$. In Section 4, we subject the RKHS estimators to sparse penalties to estimate the conditional precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. In Sections 5, 6, and 7, we establish the asymptotic properties of the sparse estimators for a fixed dimension p. In Section 8, we derive the convergence rate of the sparse estimators when p goes to infinity with the sample size. In Section 9, we discuss some issues involved in implementation. In Sections 10 and 11, we investigate and explore the performance of the proposed methods through simulation and data analysis.

2. CONDITIONAL VARIANCE OPERATOR IN RKHS

We begin with a formal definition of the conditional Gaussian graphical model. Throughout this article, we use $\mathbb{E}$ to denote expectation to avoid confusion with the edge set E. We use $\mathbb{P}$ to denote probability measures.

Definition 1. We say that $(\mathbf{X}, \mathbf{Y})$ follows a conditional Gaussian graphical model (CGGM) with respect to a graph $\mathcal{G} = (\Gamma, E)$ if

1. relation (2) holds;
2. $\mathbf{Y} \mid \mathbf{X} \sim N(\mathbb{E}(\mathbf{Y} \mid \mathbf{X}), \mathbf{\Sigma})$ for some nonrandom, positive-definite matrix $\mathbf{\Sigma}$.

Note that we do not assume a regression model for $\mathbf{Y}$ versus $\mathbf{X}$. However, we do require that $\mathbf{Y} \mid \mathbf{X}$ is multivariate normal with a constant conditional variance, which is satisfied if $\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}$, and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma})$ for an arbitrary f.

Let $\Omega_X \subseteq \mathbb{R}^q$ and $\Omega_Y \subseteq \mathbb{R}^p$ be the support of $\mathbf{X}$ and $\mathbf{Y}$. Let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$, $\kappa_Y : \Omega_Y \times \Omega_Y \to \mathbb{R}$ be positive-definite kernels and $\mathcal{H}_X$ and $\mathcal{H}_Y$ be their corresponding RKHS's. For further information about RKHS and the choices of kernels, see Aronszajn (1950) and Vapnik (1998). For two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and a bounded bilinear form $b : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$, there uniquely exist bounded linear operators $A : \mathcal{H}_1 \to \mathcal{H}_2$ and $B : \mathcal{H}_2 \to \mathcal{H}_1$ such that $\langle f, Bg\rangle_{\mathcal{H}_1} = \langle Af, g\rangle_{\mathcal{H}_2} = b(f, g)$ for any $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Applying this fact to the bounded bilinear forms

$$\text{cov}[f_1(\mathbf{X}), f_2(\mathbf{X})], \ \text{cov}[g_1(\mathbf{Y}), g_2(\mathbf{Y})], \ \text{cov}[f(\mathbf{X}), g(\mathbf{Y})],$$

we obtain three bounded linear operators

$$\begin{aligned} &\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X, \quad \Sigma_{YY} : \mathcal{H}_Y \to \mathcal{H}_Y, \\ &\Sigma_{XY} : \mathcal{H}_Y \to \mathcal{H}_X. \end{aligned} \tag{3}$$

Furthermore, Σ_{XY} can be factorized as $\Sigma_{XX}^{\frac{1}{2}} R_{XY} \Sigma_{YY}^{\frac{1}{2}}$, where $R_{XY} : \mathcal{H}_Y \to \mathcal{H}_X$ is a uniquely defined bounded linear operator. The conditional variance operator of $\mathbf{Y}$, given $\mathbf{X}$, is then defined as the bounded operator from $\mathcal{H}_Y$ to $\mathcal{H}_Y$:

$$\Sigma_{YY \mid X} = \Sigma_{YY} - \Sigma_{YY}^{\frac{1}{2}} R_{YX} R_{XY} \Sigma_{YY}^{\frac{1}{2}}.$$

This construction is due to Fukumizu, Bach, and Jordan (2009).

Let $L_2(\mathbb{P}_X)$ denote the class of functions of $\mathbf{X}$ that are squared integrable with respect to $\mathbb{P}_X$, $\mathcal{H}_X + \mathbb{R}$ denote the set of functions $\{h + c : h \in \mathcal{H}_X, c \in \mathbb{R}\}$, and $\text{cl}(\cdot)$ denote the closure of a set in $L_2(\mathbb{P}_X)$. The next theorem describes how the conditional operator $\Sigma_{YY \mid X}$ uniquely determines the CGGM, and suggests a way to estimate $\mathbf{\Sigma}$.

Theorem 1. Suppose

1. $\mathcal{H}_X \subseteq L_2(\mathbb{P}_X)$ and, for each $i = 1, \ldots, p$, $\mathbb{E}(Y^i \mid X) \in \text{cl}(\mathcal{H}_X + \mathbb{R})$;
2. κ_Y is the linear kernel: $\kappa_Y(\mathbf{a}, \mathbf{b}) = 1 + \mathbf{a}^{\mathsf{T}}\mathbf{b}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$.

Then,

$$\mathbf{\Sigma} = \{\langle y^r, \Sigma_{YY \mid X} y^s\rangle_{\mathcal{H}_Y} : r, s = 1, \ldots, p\}. \tag{4}$$

The assumption $\mathcal{H}_X \subseteq L_2(\mathbb{P}_X)$ is satisfied if the function $\kappa_X(\mathbf{x}, \mathbf{x})$ belongs to $L_2(\mathbb{P}_X)$, which is a mild requirement.

Proof. By assumption 2, any member of $\mathcal{H}_Y$ can be written as $\boldsymbol{\alpha}^\mathsf{T}\mathbf{y}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^p$. Hence, by Proposition 2 of Fukumizu et al. (2009),

$$\begin{aligned}\langle \boldsymbol{\alpha}^\mathsf{T}\mathbf{y}, \Sigma_{YY\,|\,X}(\boldsymbol{\alpha}^\mathsf{T}\mathbf{y})\rangle_{\mathcal{H}_Y} &= \inf_{f\in\mathcal{H}_X+\mathbb{R}} \operatorname{var}[\boldsymbol{\alpha}^\mathsf{T}\mathbf{Y} - f(\mathbf{X})] \\ &= \mathbb{E}[\operatorname{var}(\boldsymbol{\alpha}^\mathsf{T}\mathbf{Y}|\mathbf{X})] + \inf_{f\in\mathcal{H}_X+\mathbb{R}} \operatorname{var}\{\mathbb{E}[(\boldsymbol{\alpha}^\mathsf{T}\mathbf{Y}|\mathbf{X}) - f(\mathbf{X})]\}.\end{aligned}$$

By assumption 1, for any $\epsilon > 0$, there is an $f \in \mathcal{H}_X + \mathbb{R}$ such that $\operatorname{var}\{\mathbb{E}[(\boldsymbol{\alpha}^\mathsf{T}\mathbf{Y}\,|\,\mathbf{X}) - f(\mathbf{X})]\} < \epsilon$. So, the second term on the right-hand side above is 0, and we have

$$\langle \boldsymbol{\alpha}^\mathsf{T} y, \Sigma_{YY\,|\,X}(\boldsymbol{\alpha}^\mathsf{T} y)\rangle_{\mathcal{H}_Y} = \mathbb{E}[\operatorname{var}(\boldsymbol{\alpha}^\mathsf{T}\mathbf{Y}|\mathbf{X})]. \tag{5}$$

In the meantime, we note that

$$\begin{aligned}&\langle y^i + y^j, \Sigma_{YY\,|\,X}(y^i + y^j)\rangle_{\mathcal{H}_Y} - \langle y^i - y^j, \Sigma_{YY\,|\,X}(y^i - y^j)\rangle_{\mathcal{H}_Y} \\ &\quad = 4\langle y^i, \Sigma_{YY\,|\,X} y^j\rangle_{\mathcal{H}_Y}, \mathbb{E}[\operatorname{var}(Y^i + Y^j|\mathbf{X})] \\ &\qquad - \mathbb{E}[\operatorname{var}(Y^i - Y^j|\mathbf{X})] = 4\mathbb{E}[\operatorname{cov}(Y^i, Y^j|\mathbf{X})].\end{aligned}$$

Applying Equation (5) to the left-hand sides of the above equations, we obtain

$$\langle y^i, \Sigma_{YY\,|\,X} y^j\rangle_{\mathcal{H}_Y} = \mathbb{E}[\operatorname{cov}(Y^i, Y^j|\mathbf{X})] = \operatorname{cov}(Y^i, Y^j|\mathbf{X}),$$

as desired. ∎

Condition 1 in the theorem can be replaced by the stronger assumption that $\mathcal{H}_X + \mathbb{R}$ is a dense subset of $L_2(P_X)$, which is satisfied by some well known kernels, such as the Gaussian radial kernel. See Fukumizu et al. (2009).

3. TWO RKHS ESTIMATORS OF CONDITIONAL COVARIANCE

To construct a sample estimate of Equation (4), we need to represent operators as matrices in a finite-dimensional space. To this end, we first introduce a coordinate notation system, adopted from Horn and Johnson (1985, p. 31) with slight modifications. Let $\mathcal{H}$ be a finite-dimensional Hilbert space and $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\} \subseteq \mathcal{H}$ be a set of functions that span $\mathcal{H}$ but may not be linearly independent. We refer to $\mathcal{B}$ as a spanning system (as opposed to a basis). Any $f \in \mathcal{H}$ can be written as $\alpha_1\mathbf{b}_1 + \cdots + \alpha_m\mathbf{b}_m$ for some $\alpha = (\alpha_1, \ldots, \alpha_m)^\mathsf{T} \in \mathbb{R}^m$. This vector is denoted by $[f]_\mathcal{B}$, and is called a $\mathcal{B}$-coordinate of f. Note that $[f]_\mathcal{B}$ is not unique unless $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are linearly independent. However, this does not matter because $\sum_{i=1}^m ([f]_\mathcal{B})_i b_i$ is unique regardless of the form of $[f]_\mathcal{B}$. The same reasoning also applies to the nonuniqueness of coordinates below.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be finite-dimensional Hilbert spaces with spanning systems $\mathcal{B}_1 = \{\mathrm{b}_{11}, \ldots, \mathrm{b}_{1m_1}\}$ and $\mathcal{B}_2 = \{\mathbf{b}_{21}, \ldots, \mathbf{b}_{2m_2}\}$. Let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator. The $(\mathcal{B}_1, \mathcal{B}_2)$-representation of T, denoted by ${}_{\mathcal{B}_2}[T]_{\mathcal{B}_1}$, is the $m_2 \times m_1$ matrix

$$\{([T\mathrm{b}_{1i}]_{\mathcal{B}_2})_j : j = 1, \ldots, m_2, i = 1, \ldots, m_1\}.$$

Then, $({}_{\mathcal{B}_2}[T]_{\mathcal{B}_1})[f]_{\mathcal{B}_1}$ is a $\mathcal{B}_2$-coordinate of Tf. We write this relation as

$$[Tf]_{\mathcal{B}_2} = ({}_{\mathcal{B}_2}[T]_{\mathcal{B}_1})[f]_{\mathcal{B}_1}. \tag{6}$$

Let $\mathcal{H}_3$ be another finite-dimensional Hilbert space with a spanning system $\mathcal{B}_3$. Let $T_1 : \mathcal{H}_1 \to \mathcal{H}_2$ and $T_2 : \mathcal{H}_2 \to \mathcal{H}_3$ be linear operators. Then,

$${}_{\mathcal{B}_3}[T_2T_1]_{\mathcal{B}_1} = ({}_{\mathcal{B}_3}[T_2]_{\mathcal{B}_2})({}_{\mathcal{B}_2}[T_1]_{\mathcal{B}_1}). \tag{7}$$

The equality means the right-hand side is a $(\mathcal{B}_1, \mathcal{B}_3)$-representation of T_2T_1. Finally, if $\mathcal{H}$ is a finite-dimensional Hilbert space with a spanning system $\mathcal{B}$ and $T : \mathcal{H} \to \mathcal{H}$ is a self-adjoint and positive-semidefinite linear operator, then, for any $c > 0$,

$${}_{\mathcal{B}}[T^c]_{\mathcal{B}} = ({}_{\mathcal{B}}[T]_{\mathcal{B}})^c. \tag{8}$$

Let $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be independent copies of $(\mathbf{X}, \mathbf{Y})$, $\mathbb{P}_n$ be the empirical measure based on this sample, and $\mathbb{E}_n$ be the integral with respect to $\mathbb{P}_n$. Let

$$\begin{aligned}\hat{\mathcal{H}}_X &= \operatorname{span}\{\kappa_X(\cdot, \mathbf{X}_i) - \mathbb{E}_n\kappa_X(\cdot, \mathbf{X}) : i = 1, \ldots, n\}, \\ \hat{\mathcal{H}}_Y &= \operatorname{span}\{\kappa_Y(\cdot, \mathbf{Y}_i) - \mathbb{E}_n\kappa_Y(\cdot, \mathbf{Y}) : i = 1, \ldots, n\},\end{aligned} \tag{9}$$

where, for example, $\mathbb{E}_n\kappa_X(\cdot, \mathbf{X})$ stands for the function $\boldsymbol{x} \mapsto \mathbb{E}_n\kappa_X(\mathbf{x}, \mathbf{X})$. We center the functions $\kappa_X(\cdot, \mathbf{X}_i)$ because constants do not play a role in our development. Let $\hat{\Sigma}_{XX}$, $\hat{\Sigma}_{YY}$, and $\hat{\Sigma}_{XY}$ be as defined in the last section but with $\mathcal{H}_X$, $\mathcal{H}_Y$ replaced by $\hat{\mathcal{H}}_X$, $\hat{\mathcal{H}}_Y$, and cov replaced by the sample covariance. Let $\mathcal{B}_X$ and $\mathcal{B}_Y$ denote the spanning systems in Equation (9). Let $\mathbf{K}_X$ and $\mathbf{K}_Y$ represent the $n \times n$ kernel matrices $\{\kappa_X(\mathbf{X}_i, \mathbf{X}_j)\}$ and $\{\kappa_Y(\mathbf{Y}_i, \mathbf{Y}_j)\}$. For a symmetric matrix, $\mathbf{A}$, let $\mathbf{A}^\dagger$ represent its Moore-Penrose inverse. Let $\mathbf{Q}_n = \mathbf{I}_n - n^{-1}\mathbf{J}_n$, $\mathbf{I}_n$ is the $n \times n$ identity matrix, and $\mathbf{J}_n$ be the $n \times n$ matrix whose entries are 1. To simplify notation, we abbreviate $\mathbf{Q}_n$ by $\mathbf{Q}$ throughout the rest of the article. The following lemma crystallizes some known results, which can be proved by coordinate manipulation via formulas (6), (7), and (8).

Lemma 1. The following relations hold:

1. ${}_{\mathcal{B}_X}[\hat{\Sigma}_{XX}]_{\mathcal{B}_X} = n^{-1}\mathbf{Q}\mathbf{K}_X\mathbf{Q}$, ${}_{\mathcal{B}_Y}[\hat{\Sigma}_{YY}]_{\mathcal{B}_Y} = n^{-1}\mathbf{Q}\mathbf{K}_Y\mathbf{Q}$;
2. ${}_{\mathcal{B}_Y}[\hat{\Sigma}_{YX}]_{\mathcal{B}_X} = n^{-1}\mathbf{Q}\mathbf{K}_X\mathbf{Q}$, ${}_{\mathcal{B}_X}[\hat{\Sigma}_{XY}]_{\mathcal{B}_Y} = n^{-1}\mathbf{Q}\mathbf{K}_Y\mathbf{Q}$;
3. ${}_{\mathcal{B}_Y}[\hat{\Sigma}_{YY\,|\,X}]_{\mathcal{B}_Y} = n^{-1}[\mathbf{Q}\mathbf{K}_Y\mathbf{Q} - (\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger (\mathbf{Q}\mathbf{K}_Y\mathbf{Q})]$.

When the dimension q of $\mathbf{X}$ is large relative to n, it is beneficial to use regularized version of $(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger$. Here, we employ two types of regularization: the principal-component (PC) regularization and the ridge-regression (RR) regularization. Let $\mathbf{A}$ be a positive-semidefinite matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigen-vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Let $\epsilon \geq 0$. We call the matrix $(\mathbf{A})^\dagger_\epsilon = \sum_{i=1}^n \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^\mathsf{T} I(\lambda_i > \epsilon)$ the PC-inverse of $\mathbf{A}$. Note that $(\mathbf{A})^\dagger_0 = \mathbf{A}^\dagger$. We call the matrix $(\mathbf{A})^\ddagger_\epsilon = (\mathbf{A} + \epsilon\mathbf{I}_n)^{-1}$ the RR-inverse of $\mathbf{A}$. The following result can be verified by simple calculation.

Lemma 2. For any $f_1, f_2 \in \hat{\mathcal{H}}_Y$, we have $\langle f_1, f_2\rangle_{\hat{\mathcal{H}}_Y} = ([f_1]_{\mathcal{B}_Y})^\mathsf{T}\mathbf{Q}\mathbf{K}_Y\mathbf{Q}[f_2]_{\mathcal{B}_Y}$.

We now derive the sample estimate of the conditional covariance matrix $\boldsymbol{\Sigma} = \operatorname{var}(\mathbf{Y}\,|\,\mathbf{X})$. Let $\mathbf{D}_Y$ denote the matrix $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\mathsf{T}$.

Theorem 2. Let $\hat{\Sigma}_{YY\,|\,X} : \hat{\mathcal{H}}_Y \to \hat{\mathcal{H}}_Y$ be defined by the coordinate representation

$$\begin{aligned}&{}_{\mathcal{B}_Y}[\hat{\Sigma}_{YY\,|\,X}]_{\mathcal{B}_Y} \\ &\quad = n^{-1}\big[\mathbf{Q}\mathbf{K}_Y\mathbf{Q} - \big(\mathbf{Q}\mathbf{K}_X\mathbf{Q}\big)(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^*_{\epsilon_n}(\mathbf{Q}\mathbf{K}_Y\mathbf{Q})\big],\end{aligned} \tag{10}$$

where $*$ can be either $\dagger$ or $\ddagger$, $\mathbf{K}_Y$ is the linear kernel matrix, and $\{\epsilon_n\}$ is a sequence of nonnegative numbers. Then

$$\{\langle y^i, \hat{\Sigma}_{YY\,|\,X} y^j\rangle_{\hat{\mathcal{H}}_Y}\} = n^{-1}\big[\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y - \mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\big(\mathbf{Q}\mathbf{K}_X\mathbf{Q}\big)\big(\mathbf{Q}\mathbf{K}_X\mathbf{Q}\big)^*_{\epsilon_n}\mathbf{Q}\mathbf{D}_Y\big]. \quad (11)$$

Despite its appearance, the matrix $(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^*_\epsilon$ is actually a symmetric matrix. One can also show that Equation (11) is a positive-semidefinite matrix. We denote the matrix (11) as $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ if $* = \dagger$, and as $\hat{\mathbf{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$ if $* = \ddagger$. In the following, $\mathbf{e}_i$ represents a p-dimensional vector whose ith entry is 1 and other entries are 0. For an expression that represents a vector, such as $[f]_{\mathcal{B}_Y}$, let $([f]_{\mathcal{B}_Y})_i$ denote its ith entry.

Proof of Theorem 1. Note that, for an $f \in \mathcal{H}_Y$, $[f]_{\mathcal{B}_Y}$ is any vector $\mathbf{a} \in \mathbb{R}^p$ such that $f(\mathbf{y}) = \mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\mathbf{y}$. Because $y^i = \mathbf{e}_i^{\mathsf{T}}\mathbf{y} = \mathbf{e}_i^{\mathsf{T}}(\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y)^{-1}\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\mathbf{y}$, we have

$$[y^i]_{\mathcal{B}_Y} = \mathbf{D}_Y\big(\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\big)^{-1}\mathbf{e}_i.$$

Then, by Lemma 2,

$$\langle y^i, \hat{\Sigma}_{YY\,|\,X} y^j\rangle_{\mathcal{H}_Y} = \mathbf{e}_i^{\mathsf{T}}\big(\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\big)^{-1}\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{K}_Y\mathbf{Q}([\hat{\Sigma}_{YY\,|\,X}]_{\mathcal{B}_Y})\mathbf{D}_Y\big(\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\big)^{-1}\mathbf{e}_j.$$

When κ_Y is the linear kernel, $\mathbf{K}_Y = \mathbf{D}_Y\mathbf{D}_Y^{\mathsf{T}}$. Hence, the right-hand side reduces to

$$\mathbf{e}_i^{\mathsf{T}}\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}([\hat{\Sigma}_{YY\,|\,X}]_{\mathcal{B}_Y})\mathbf{D}_Y\big(\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\big)^{-1}\mathbf{e}_j.$$

Now substitute Equation (10) into the above expression to complete the proof. ■

4. INTRODUCING SPARSE PENALTY

Let $\hat{\mathbf{\Sigma}}$ denote either of the RKHS estimators given by Equation (11). We are interested in the sparse estimation of $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. Let

$$L_n(\mathbf{\Theta}) = -\log\det(\mathbf{\Theta}) + \mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{\Sigma}}). \quad (12)$$

This objective function has the same form as that used in unconditional Gaussian graphical models, such as those considered by Lauritzen (1996), Yuan and Lin (2007), and Friedman, Hastie, and Tibshirani (2008), except that the sample covariance matrix therein is replaced by the RKHS estimate of conditional covariance matrix.

To achieve sparsity in $\mathbf{\Theta}$, we introduce two types of penalized versions of the objective function (12):

$$\text{lasso:} \quad \Upsilon_n(\mathbf{\Theta}) = L_n(\mathbf{\Theta}) + \lambda_n\sum_{i\neq j}|\theta_{ij}|, \quad (13)$$

$$\text{adaptive lasso:} \quad \Lambda_n(\mathbf{\Theta}) = L_n(\mathbf{\Theta}) + \lambda_n\sum_{i\neq j}|\tilde{\theta}_{ij}|^{-\gamma}|\theta_{ij}|, \quad (14)$$

where, in Equation (14), γ is a positive number and $\{\tilde{\theta}_{ij}\}$ is a $\sqrt{n}$-consistent, nonsparse estimate of Θ. For more details about the development of these two types of penalty functions; see Tibshirani (1996), Zou (2006), Zhao and Yu (2006), and Zou and Li (2008). Yuan and Lin (2007) used lasso and nonnegative garrote (Breiman 1995) penalty functions for the unconditional graphical model, the latter of which is similar to adaptive lasso with $\gamma = 1$. In the following, we will write

$$P_n(\mathbf{\Theta}) = \lambda_n\sum_{i\neq j}|\theta_{ij}|, \quad \Pi_n(\mathbf{\Theta}) = \lambda_n\sum_{i\neq j}|\tilde{\theta}_{ij}|^{-\gamma}|\theta_{ij}|.$$

5. VON MISES EXPANSIONS OF THE RKHS ESTIMATORS

The sparse and oracle properties of the estimators introduced in Section 4 depend heavily on the asymptotic properties of the RKHS estimators $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ and $\hat{\mathbf{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$. In this section, we derive their von Mises expansions (von Mises 1947). For simplicity, we base our asymptotic development on the polynomial kernel. That is, we let

$$\kappa_X(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^{\mathsf{T}}\mathbf{b} + 1)^r, \quad r = 1, 2, \ldots. \quad (15)$$

For a matrix $\mathbf{A}$ and an integer $k \geq 2$, let $\mathbf{A}^{\otimes k}$ be the k-fold Kronecker product $A \otimes \cdots \otimes A$. For $k = 1$, we adopt the convention $A^{\otimes 1} = A$. For a k-dimensional array $\mathbf{B} = \{b_{i_1\cdots i_k} : i_1, \ldots, i_k = 1, \ldots, q\}$, we define the "vec half" operator as

$$\mathrm{vech}(\mathbf{B}) = \{b_{i_1\cdots i_k} : 1 \leq i_k \leq \cdots \leq i_1 \leq q\}.$$

This is a generalization of the vech operator for matrices introduced by Henderson and Searle (1979). It can be shown that there exists a matrix $\mathbf{G}_{k,q}$, of full column rank, such that $\mathrm{vec}(\mathbf{B}) = \mathbf{G}_{k,q}\,\mathrm{vech}(\mathbf{B})$. The specific form of $\mathbf{G}_{k,q}$ is not important to us. For $k = 1$, we adopt the convention $\mathrm{vech}(\mathbf{B}) = \mathbf{B}$, $\mathbf{G}_{1,q} = \mathbf{I}_q$. Let

$$\mathbf{U} = \begin{pmatrix} \mathrm{vech}(\mathbf{X}^{\otimes 1}) \\ \vdots \\ \mathrm{vech}(\mathbf{X}^{\otimes r}) \end{pmatrix},$$

$$\mathbf{U}_i = \begin{pmatrix} \mathrm{vech}(\mathbf{X}_i^{\otimes 1}) \\ \vdots \\ \mathrm{vech}(\mathbf{X}_i^{\otimes r}) \end{pmatrix}, \quad \text{and} \quad \mathbf{D}_U = (\mathbf{U}_1, \ldots, \mathbf{U}_n)^{\mathsf{T}}.$$

The estimators $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ and $\hat{\mathbf{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$ involve $n \times n$ matrices $\mathbf{Q}\mathbf{K}_X\mathbf{Q}$ and $(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^*_{\epsilon_n}$, which are difficult to handle asymptotically because their dimensions grow with n. We now give an asymptotically equivalent expression which only involves matrices of fixed dimensions. Let

$$\tilde{\mathbf{\Sigma}} = n^{-1}\big[\mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y - \mathbf{D}_Y^{\mathsf{T}}\mathbf{Q}\mathbf{D}_U\big(\mathbf{D}_U^{\mathsf{T}}\mathbf{Q}\mathbf{D}_U\big)^{-1}\mathbf{D}_U^{\mathsf{T}}\mathbf{Q}\mathbf{D}_Y\big]. \quad (16)$$

Lemma 3. Suppose κ_X is the polynomial kernel (15) and $\mathrm{var}(\mathbf{U})$ is positive definite.

1. If $\epsilon_n = o(n)$, then, with probability tending to 1, $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n) = \tilde{\mathbf{\Sigma}}$;
2. If $\epsilon_n = o(n^{\frac{1}{2}})$, then $\hat{\mathbf{\Sigma}}_{\mathrm{RR}}(\epsilon_n) = \tilde{\mathbf{\Sigma}} + o_P(n^{-\frac{1}{2}})$.

It is interesting to note that different choices of inversion requires different convergence rates for ϵ_n.

Proof. By simple computation, we find

$$\big(\mathbf{X}_i^{\mathsf{T}}\mathbf{X}_j + 1\big)^r = 1 + \sum_{k=1}^{r}\binom{r}{k}\big[\mathrm{vech}\big(\mathbf{X}_i^{\otimes k}\big)\big]^{\mathsf{T}}\mathbf{G}_{k,q}^{\mathsf{T}}\mathbf{G}_{k,q}\,\mathrm{vech}\big(\mathbf{X}_i^{\otimes k}\big).$$

Hence, $\mathbf{K}_X = \mathbf{J}_n + \mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}$ where $\mathbf{C} = \operatorname{diag}(\binom{r}{1}\mathbf{G}_{1,q}^\mathsf{T}\mathbf{G}_{1,q}, \ldots, \binom{r}{r}\mathbf{G}_{r,q}^\mathsf{T}\mathbf{G}_{r,q})$. So,

$$\mathbf{Q}\mathbf{K}_X\mathbf{Q} = \mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q}. \tag{17}$$

Let s be the dimension of $\mathbf{U}$. Then, $\operatorname{rank}(\mathbf{Q}\mathbf{K}_X\mathbf{Q}) = s$. Let $\lambda_1 \geq \cdots \geq \lambda_s > 0$ be the nonzero eigenvalues and $\mathbf{v}_1, \ldots, \mathbf{v}_s$ be the corresponding vectors of this matrix. Since $\lambda_1, \ldots, \lambda_s$ are also eigenvalues of $n(n^{-1}\mathbf{C}^{\frac{1}{2}}\mathbf{D}_U^\mathsf{T}\mathbf{Q}\mathbf{D}_U\mathbf{C}^{\frac{1}{2}})$, which converges in probability to the positive-definite matrix $\mathbf{C}^{\frac{1}{2}}\operatorname{var}(\mathbf{U})\mathbf{C}^{\frac{1}{2}}$, we have $\lambda_s = nc + o_P(n)$ for some $c > 0$. This, together with $\epsilon_n = o(n)$, implies $(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger_{\epsilon_n} = (\mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q})^\dagger$ with probability tending to 1. Consequently, with probability tending to 1,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n) = n^{-1}\big[&\mathbf{D}_Y^\mathsf{T}\mathbf{Q}\mathbf{D}_Y \\ &- \mathbf{D}_Y^\mathsf{T}\mathbf{Q}\big(\mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q}\big)\big(\mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q}\big)^\dagger\mathbf{Q}\mathbf{D}_Y\big].\end{aligned} \tag{18}$$

Now it is easy to verify that if $\mathbf{B} \in \mathbb{R}^{s\times t}$ is a matrix of full column rank and $\mathbf{A} \in \mathbb{R}^{t\times t}$ is a positive-definite matrix, then

$$(\mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T})(\mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T})^\dagger = \mathbf{B}(\mathbf{B}^\mathsf{T}\mathbf{B})^{-1}\mathbf{B}^\mathsf{T}. \tag{19}$$

Thus, the right-hand side of Equation (18) is $\tilde{\boldsymbol{\Sigma}}$, which proves part 1. To prove part 2, we first note that

$$\begin{aligned}&(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\ddagger_{\epsilon_n} \\ &\quad= \left(\mathbf{Q}\mathbf{K}_X\mathbf{Q} - \sum_{i=1}^{s}\frac{\lambda_i\epsilon_n}{\lambda_i+\epsilon_n}\mathbf{v}_i\mathbf{v}_i^\mathsf{T}\right)(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger.\end{aligned} \tag{20}$$

Since the function $-\epsilon_n/(\epsilon_n + \lambda)$ is increasing for $\lambda > 0$, we have

$$\begin{aligned}&-\frac{\epsilon_n}{\lambda_s+\epsilon_n}(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger \leq (\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\ddagger_{\epsilon_n} \\ &\quad- (\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger \leq 0.\end{aligned}$$

Hence,

$$\begin{aligned}0 &\geq n^{-1}\mathbf{D}_Y^\mathsf{T}\mathbf{Q}\big[(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\ddagger_{\epsilon_n} - (\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger\big]\mathbf{Q}\mathbf{D}_Y \\ &\geq -\frac{\epsilon_n}{\lambda_s+\epsilon_n}n^{-1}\mathbf{D}_Y^\mathsf{T}\mathbf{Q}(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger\mathbf{Q}\mathbf{D}_Y,\end{aligned} \tag{21}$$

By Equation (19),

$$\begin{aligned}&n^{-1}\mathbf{D}_Y^\mathsf{T}\mathbf{Q}(\mathbf{Q}\mathbf{K}_X\mathbf{Q})(\mathbf{Q}\mathbf{K}_X\mathbf{Q})^\dagger\mathbf{Q}\mathbf{D}_Y \\ &\quad= \big(n^{-1}\mathbf{D}_Y^\mathsf{T}\mathbf{Q}\mathbf{D}_U\big)\big(n^{-1}\mathbf{D}_U^\mathsf{T}\mathbf{Q}\mathbf{D}_U\big)^{-1}(n^{-1}\mathbf{D}_U\mathbf{Q}\mathbf{D}_Y) \\ &\quad\xrightarrow{P} \operatorname{cov}(\mathbf{Y},\mathbf{U})[\operatorname{var}(\mathbf{U})]^{-1}\operatorname{cov}(\mathbf{U},\mathbf{Y}).\end{aligned}$$

Hence, the right-hand side of Equation (21) is of the order $o_P(n^{-\frac{1}{2}})$. In other words,

$$\begin{aligned}&\hat{\boldsymbol{\Sigma}}_{\mathrm{RR}}(\epsilon_n) \\ &\quad= n^{-1}\big[\mathbf{D}_Y^\mathsf{T}\mathbf{Q}\mathbf{D}_Y - \mathbf{D}_Y^\mathsf{T}\mathbf{Q}(\mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q})(\mathbf{Q}\mathbf{D}_U\mathbf{C}\mathbf{D}_U^\mathsf{T}\mathbf{Q})^\dagger\mathbf{Q}\mathbf{D}_Y\big] \\ &\quad+ o_P\big(n^{-\frac{1}{2}}\big).\end{aligned}$$

Here, we evoke Equation (19) again to complete the proof. ■

Note that when $r = 1$ and $p < n$, and $\epsilon_n = 0$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ reduces to

$$\begin{aligned}&n^{-1}\big[\mathbf{D}_Y^\mathsf{T}\mathbf{Q}\mathbf{D}_Y - \mathbf{D}_Y\mathbf{Q}\mathbf{D}_X\big(\mathbf{D}_X^\mathsf{T}\mathbf{Q}\mathbf{D}_X\big)^{-1}\mathbf{D}_X\mathbf{Q}\mathbf{D}_Y\big] \\ &\qquad= \operatorname{var}_n(\mathbf{Y}) - \operatorname{cov}_n(\mathbf{Y},\mathbf{X})[\operatorname{var}_n(\mathbf{X})]^{-1}\operatorname{cov}_n(\mathbf{X},\mathbf{Y}),\end{aligned}$$

where $\operatorname{var}_n(\cdot)$ and $\operatorname{cov}_n(\cdot,\cdot)$ denote the sample variance and covariance matrices. This is exactly the sample estimate of the residual variance for linear regression.

Lemma 3 allows us to derive the asymptotic expansions of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$ from that of $\tilde{\boldsymbol{\Sigma}}$, which is a (matrix-valued) function of sample moments. Let $\mathcal{F}$ be a convex family of probability measures defined on Ω_{XY} that contains all the empirical distributions $\mathbb{P}_n$ and the true distribution $\mathbb{P}_0$ of $(\mathbf{X},\mathbf{Y})$. Let $T : \mathcal{F} \to \mathbb{R}^{p\times p}$ be the following statistical functional:

$$\mathbb{P} \mapsto \operatorname{var}_{\mathbb{P}}(\mathbf{Y}) - \operatorname{cov}_{\mathbb{P}}(\mathbf{Y},\mathbf{U})[\operatorname{var}_{\mathbb{P}}(\mathbf{U})]^{-1}\operatorname{cov}_{\mathbb{P}}(\mathbf{U},\mathbf{Y}). \tag{22}$$

In this notation, $\tilde{\boldsymbol{\Sigma}} = T(\mathbb{P}_n)$. In the following, we use var and cov to denote the variance and covariance under $\mathbb{P}_0$. For the polynomial kernel (15), the evaluation $T(\mathbb{P}_0)$ has a special meaning, as described in the next lemma. Its proof is standard, and is omitted.

Lemma 4. Suppose:

1. Entries of $\mathbb{E}(\mathbf{Y}\mid\mathbf{X})$ are polynomials in $X^1, \ldots, X^q$ of degrees no more than r;
2. $\mathbf{Y}\mid\mathbf{X} \sim N(\mathbb{E}(\mathbf{Y}\mid\mathbf{X}), \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a nonrandom matrix.

Then, $T(\mathbb{P}_0) = \boldsymbol{\Sigma}$.

Let $\mathbb{P}_\alpha = (1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_n$, where $\alpha \in [0,1]$. Let D_α denote the differential operator $\partial/\partial\alpha$, and let $D_{\alpha=0}$ denote the operation of taking derivative with respect to α and then evaluating the derivative at $\alpha = 0$. It is well known that if T is Hadamard differentiable with respect to the norm $\|\cdot\|_\infty$ in $\mathcal{F}$, then

$$T(\mathbb{P}_n) = T(\mathbb{P}_0) + D_{\alpha=0}T(\mathbb{P}_\alpha) + o_P\big(n^{-\frac{1}{2}}\big). \tag{23}$$

Since the functional T in our case is a smooth function of sample moments, it is Hadamard differentiable under very general conditions. See, for example, Reeds (1976), Fernholz (1983), Bickel et al. (1993), and Ren and Sen (1991). The next lemma can be verified by straightforward computation.

Lemma 5. Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be square-integrable random vectors. Then,

$$D_{\alpha=0}[\operatorname{cov}_{\mathbb{P}_\alpha}(\mathbf{V}_1,\mathbf{V}_2)] = \mathbb{E}_n\mathbf{q}(\mathbf{V}_1,\mathbf{V}_2) - \mathbb{E}\,\mathbf{q}(\mathbf{V}_1,\mathbf{V}_2),$$

where $\mathbf{q}(\mathbf{V}_1,\mathbf{V}_2) = (\mathbf{V}_1 - \mathbb{E}\mathbf{V}_1)(\mathbf{V}_2 - \mathbb{E}\mathbf{V}_2)^\mathsf{T}$.

We will adopt the following notational system:

$$\boldsymbol{\mu}_Y = \mathbb{E}(\mathbf{Y}),\ \ \boldsymbol{\mu}_U = \mathbb{E}(\mathbf{U}),\ \ \mathbf{V}_Y = \operatorname{var}(\mathbf{Y}),\ \ \mathbf{V}_U = \operatorname{var}(\mathbf{U}),$$
$$\mathbf{V}_{UY} = \operatorname{cov}(\mathbf{U},\mathbf{Y}).$$

The next theorem gives the first-order von-Mises expansion of $\tilde{\boldsymbol{\Sigma}}$.

Theorem 3. Suppose the functional in Equation (22) is Hadamard differentiable, and the conditions in Lemma 4 hold. Then,

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \mathbb{E}_n\mathbf{M} + o_P\big(n^{-\frac{1}{2}}\big), \tag{24}$$

where $\mathbf{M} = \mathbf{M}_0 - \mathbb{E}\mathbf{M}_0$ and

$$\begin{aligned}\mathbf{M}_0 = &\big[\mathbf{Y} - \boldsymbol{\mu}_Y - \mathbf{V}_{YU}\mathbf{V}_U^{-1}(\mathbf{U} - \boldsymbol{\mu}_U)\big] \\ &\times \big[\mathbf{Y} - \boldsymbol{\mu}_Y - \mathbf{V}_{YU}\mathbf{V}_U^{-1}(\mathbf{U} - \boldsymbol{\mu}_U)\big]^\mathsf{T}.\end{aligned}$$

Proof. We have

$$D_\alpha T(\mathbb{P}_\alpha) = D_\alpha[\mathrm{var}_{\mathbb{P}_\alpha}(\mathbf{Y})] - D_\alpha\{\mathrm{cov}_{\mathbb{P}_\alpha}(\mathbf{Y}, \mathbf{U}) \times [\mathrm{var}_{\mathbb{P}_\alpha}(\mathbf{U})]^{-1}\mathrm{cov}_{\mathbb{P}_\alpha}(\mathbf{U}, \mathbf{Y})\}.$$

We now apply Lemma 5 to obtain

$$\begin{aligned} &D_{\alpha=0}\{\mathrm{cov}_{\mathbb{P}_\alpha}(\mathbf{Y}, \mathbf{U})[\mathrm{var}_{\mathbb{P}_\alpha}(\mathbf{U})]^{-1}\mathrm{cov}_{\mathbb{P}_\alpha}(\mathbf{U}, \mathbf{Y})\} \\ &\quad = [\mathbb{E}_n\mathbf{q}(\mathbf{Y}, \mathbf{U}) - \mathbb{E}\mathbf{q}(\mathbf{Y}, \mathbf{U})]\mathbf{V}_U^{-1}\mathbf{V}_{UY} + \mathbf{V}_{YU}D_{\alpha=0} \\ &\qquad \times\{[\mathrm{var}_{\mathbb{P}_\alpha}(\mathbf{U})]^{-1}\}\mathbf{V}_{UY} + \mathbf{V}_{YU}\mathbf{V}_U^{-1}[\mathbb{E}_n\mathbf{q}(\mathbf{U}, \mathbf{Y}) - \mathbb{E}\mathbf{q}(\mathbf{U}, \mathbf{Y})]. \end{aligned}$$

By the chain rule for differentiation and Lemma 5,

$$D_{\alpha=0}\{[\mathrm{var}_{\mathbb{P}_\alpha}(\mathbf{U})]^{-1}\} = -\mathbf{V}_U^{-1}[\mathbb{E}_n\mathbf{q}(\mathbf{U}, \mathbf{U}) - \mathbb{E}\mathbf{q}(\mathbf{U}, \mathbf{U})]\mathbf{V}_U^{-1}.$$

Hence,

$$\begin{aligned} D_{\alpha=0}T(\mathbb{P}_\alpha) = &[\mathbb{E}_n\mathbf{q}(\mathbf{Y}, \mathbf{Y}) - \mathbb{E}\mathbf{q}(\mathbf{Y}, \mathbf{Y})] \\ &- [\mathbb{E}_n\mathbf{q}(\mathbf{Y}, \mathbf{U})^{\mathsf{T}} - \mathbb{E}\mathbf{q}(\mathbf{Y}, \mathbf{U})]\mathbf{V}_U^{-1}\mathbf{V}_{UY} \\ &+ \mathbf{V}_{YU}\mathbf{V}_U^{-1}[\mathbb{E}_n\mathbf{q}(\mathbf{U}, \mathbf{U}) - \mathbb{E}\mathbf{q}(\mathbf{U}, \mathbf{U})]\mathbf{V}_U^{-1}\mathbf{V}_{UY} \\ &- \mathbf{V}_{YU}\mathbf{V}_U^{-1}[\mathbb{E}_n\mathbf{q}(\mathbf{U}, \mathbf{Y}) - \mathbb{E}\mathbf{q}(\mathbf{U}, \mathbf{Y})]. \end{aligned}$$

This can be rewritten as $\mathbb{E}_n\mathbf{M}_0 - \mathbb{E}\mathbf{M}_0$, where $\mathbf{M}_0$ is the matrix in the theorem. ■

Using this expansion, we can write down the asymptotic distribution of $\tilde{\boldsymbol{\Sigma}}$.

Corollary 1. Suppose the functional in Equation (22) is Hadamard differentiable. Then,

$$\sqrt{n}\mathrm{vec}(\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Xi}), \tag{25}$$

where

$$\boldsymbol{\Xi} = \left(\mathbf{I}_p, -\mathbf{V}_{YU}\mathbf{V}_U^{-1}\right)^{\otimes 2} \mathrm{var}\left[\begin{pmatrix}\mathbf{Y} - \boldsymbol{\mu}_Y \\ \mathbf{U} - \boldsymbol{\mu}_U\end{pmatrix} \otimes \begin{pmatrix}\mathbf{Y} - \boldsymbol{\mu}_Y \\ \mathbf{U} - \boldsymbol{\mu}_U\end{pmatrix}\right]\begin{pmatrix}\mathbf{I}_p \\ -\mathbf{V}_U^{-1}\mathbf{V}_{UY}\end{pmatrix}^{\otimes 2}.$$

By Lemma 3, Theorem 3, and Slutsky's theorem, we arrive at the following von Mises expansions for the two RKHS estimators $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$ of the conditional variance $\boldsymbol{\Sigma}$.

Corollary 2. Suppose the functional in Equation (22) is Hadamard differentiable, and the conditions in Lemma 4 hold.

1. If $\epsilon_n = o(n^{\frac{1}{2}})$, then $\hat{\boldsymbol{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$ has expansion (24).
2. If $\epsilon_n = o(n)$, then $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ has expansion (24).

Although in this article we have only studied the asymptotic distribution for the polynomial kernel, the basic formulation and analysis could potentially be extended to other kernels under some regularity conditions. We leave this to future research.

6. SPARSITY AND ASYMPTOTIC DISTRIBUTION: THE LASSO

In this section, we study the asymptotic properties of the sparse estimator based on the objective function $\Upsilon_n(\Theta)$ in Equation (13). Let $\mathbb{S}^{p\times p}$ denote the class of all $p \times p$ symmetric matrices. For a matrix $\mathbf{A} \in \mathbb{S}^{p\times p}$, let $\sigma_i(\mathbf{A})$ be the ith eigenvalue of $\mathbf{A}$. Note that, for any integer r, we have

$$\mathrm{tr}(\mathbf{A}^r) = \sum_{i=1}^{p}\sigma_i^r(\mathbf{A}). \tag{26}$$

Let E_n be the sample estimate of E; that is, $E_n = \{(i, j) : \hat{\theta}_{ij} \neq 0\}$, where $\hat{\boldsymbol{\Theta}} = \{\hat{\theta}_{ij}\}$ is the minimizer of Equation (13). Following Lauritzen (1996), for a matrix $\mathbf{A} = \{a_{ij}\}$ and a graph $\mathcal{G} = (\Gamma, E)$, let $\mathbf{A}(\mathcal{G})$ denote the matrix that sets a_{ij} to 0 whenever $(i, j) \notin E$. Let

$$\begin{aligned} \mathbb{R}^{p\times p}(\mathcal{G}) &= \{\mathbf{A}(\mathcal{G}) : \mathbf{A} \in \mathbb{R}^{p\times p}\}, \quad \mathbb{S}^{p\times p}(\mathcal{G}) \\ &= \{\mathbf{A}(\mathcal{G}) : \mathbf{A} \in \mathbb{S}^{p\times p}\}. \end{aligned}$$

We define a bivariate sign function as follows. For any numbers a, b, let

$$\mathrm{sign}(a, b) = \mathrm{sign}(a)I(a \neq 0) + \mathrm{sign}(b)I(a = 0).$$

The following lemma will prove useful. Its proof is omitted.

Lemma 6. The bivariate sign function $\mathrm{sign}(a, b)$ has the following properties:

1. For any $c_1 > 0, c_2 > 0$, $\mathrm{sign}(c_1a, c_2b) = \mathrm{sign}(a, b)$;
2. For sufficiently small $|b|$, $|a + b| - |a| = \mathrm{sign}(a, b)b$.

The next theorem generalizes Theorem 1 of Yuan and Lin (2007). It applies to any random matrix $\hat{\boldsymbol{\Sigma}}$ with expansion (24), including $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$.

Theorem 4. Suppose $(\mathbf{X}, \mathbf{Y})$ follows CGGM with respect to a graph (Γ, E). Let $\hat{\Theta}$ be the minimizer of Equation (13), where $\sqrt{n}\lambda_n \to \lambda_0 > 0$ and $\hat{\boldsymbol{\Sigma}}$ having the expansion (24). Let $\mathbf{W} \in \mathbb{R}^{p\times p}$ be a random matrix such that $\mathrm{vec}(\mathbf{W})$ is distributed as $N[0, \mathrm{var}(\mathrm{vec}(\mathbf{M}))]$. Then, the following assertions hold.

1. $0 < \lim_{n\to\infty}\mathbb{P}(E_n = E) < 1$;
2. For any $\epsilon > 0$, there is a $\lambda_0 > 0$ such that $\lim_{n\to\infty}\mathbb{P}(E_n = E) > 1 - \epsilon$;
3. $\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \xrightarrow{\mathcal{D}} \mathrm{argmin}\{\Phi(\boldsymbol{\Delta}, \mathbf{W}) : \boldsymbol{\Delta} \in \mathbb{S}^{p\times p}\}$, where

$$\Phi(\boldsymbol{\Delta}, \mathbf{W}) = \mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta}\boldsymbol{\Sigma}/2 + \boldsymbol{\Delta}\mathbf{W}) + \lambda_0\sum_{i\neq j}\mathrm{sign}(\theta_{0,ij}, \delta_{ij})\delta_{ij}.$$

The inequality $\lim_{n\to\infty}\mathbb{P}(E_n = E) > 0$ implies that, as $n \to \infty$, there is a positive probability to estimate a parameter as 0 when it is 0. A stronger property is that this probability tends to 1. For clarity, we refer to the former property as *sparsity*, and the latter as *sparsistency* (see Fan and Li 2001; Lam and Fan 2009). According to this definition, the two inequalities in part 1 mean that lasso is sparse but not sparsistent. Note that the unconstrained maximum likelihood estimate—and indeed any regular estimate—is not sparse. Part 2 means that, even though lasso is not sparsistent, we can make it as close to sparsistent as we wish by choosing a sufficiently large λ_0. Part 3 gives the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)$. It is not the same as that of the maximum likelihood estimate under the constraint $\theta_{ij} = 0$ for $(i, j) \in E^c$. That is, it does not have the oracle property. The proof of part 3 is similar to that of Theorem 1 in Yuan and Lin (2007) in the context of GGM. However, to our knowledge there were no previous results parallel to parts 1 and 2 for GGM. Knight and Fu (2000) contains some basic ideas for the asymptotics for lasso-type estimators.

Proof of Theorem 2. We prove the three assertions in the order *3, 1, 2*.

3. Let $\mathbf{\Theta}_0$ be the true value of $\mathbf{\Theta}$, and let

$$\Phi_n(\mathbf{\Delta}) = n[\Upsilon_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - \Upsilon_n(\mathbf{\Theta}_0)].$$

By an argument similar to Yuan and Lin (2007, Theorem 1), it can be shown that

$$\begin{aligned} &n[L_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - L_n(\mathbf{\Theta}_0)] \\ &\quad = \operatorname{tr}(\mathbf{\Delta\Sigma\Delta\Sigma})/2 + n^{1/2}\operatorname{tr}(\mathbf{\Delta}\mathbb{E}_n\mathbf{M}), \\ &n[P_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - P_n(\mathbf{\Theta}_0)] \\ &\quad = \lambda_0 \sum_{i\neq j} \operatorname{sign}(\theta_{0,ij}, \delta_{ij})\delta_{ij} + o(1). \end{aligned}$$

Since $n^{1/2}\mathbb{E}_n\mathbf{M} \xrightarrow{\mathcal{D}} \mathbf{W}$, we have

$$\begin{aligned} \Phi_n(\mathbf{\Delta}) &\xrightarrow{\mathcal{D}} \operatorname{tr}(\mathbf{\Delta\Sigma\Delta\Sigma}/2 + \mathbf{\Delta W}) + \lambda_0 \sum_{i\neq j} \operatorname{sign}(\theta_{0,ij}, \delta_{ij})\delta_{ij} \\ &= \Phi(\mathbf{\Delta}, \mathbf{W}). \end{aligned}$$

Both $\Phi_n(\mathbf{\Delta})$ and $\Phi(\mathbf{\Delta}, \mathbf{W})$ are strictly convex with probability 1. Applying Theorem 4.4 of Geyer (1994) we see that

$$\operatorname{argmin}\{\Phi_n(\mathbf{\Delta}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}\} \xrightarrow{\mathcal{D}} \operatorname{argmin}\{\Phi(\mathbf{\Delta}, \mathbf{W}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}\}.$$

However, by construction, if $\hat{\mathbf{\Delta}}$ is the (almost surely unique) minimizer of $\Phi_n(\mathbf{\Delta})$, then $\hat{\mathbf{\Delta}} = n^{1/2}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$. This proves part 3.

1. For a generic function $f(\mathbf{t})$ defined on $\mathbf{t} \in \mathbb{R}^s$, let $\partial^L_{t_i}$ and $\partial^R_{t_i}$ be the left and right partial derivatives with respect to the ith component of $\mathbf{t}$. When f is differentiable with respect to t_i, we write $\partial_{t_i} = \partial^L_{t_i} = \partial^R_{t_i}$. Note that $E_n = E$ if and only if $\Phi(\mathbf{\Delta}, \mathbf{W})$ is minimized within $\mathbb{S}^{p\times p}(\mathcal{G})$. This happens if and only if

$$\begin{aligned} &\partial^L_{\delta_{ij}} \Phi(\mathbf{\Delta}, \mathbf{W}) \le 0 \le \partial^R_{\delta_{ij}} \Phi(\mathbf{\Delta}, \mathbf{W}), \quad (i, j) \in \Gamma \times \Gamma, \ i \ge j, \\ &\text{and} \qquad \mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G}). \end{aligned} \tag{27}$$

Here, we only consider the cases $i \ge j$ because $\mathbf{\Delta}$ is a symmetric matrix. Let

$$\begin{aligned} L(\mathbf{\Delta}, \mathbf{W}) &= \operatorname{tr}(\mathbf{\Delta\Sigma\Delta\Sigma}/2 + \mathbf{\Delta W}), \\ P(\mathbf{\Delta}, \mathbf{W}) &= \lambda_0 \sum_{i\neq j} \operatorname{sign}(\theta_{0,ij}, \delta_{ij})\delta_{ij}. \end{aligned}$$

Then, $\partial L(\mathbf{\Delta}, \mathbf{W})/\partial\mathbf{\Delta} = \mathbf{\Sigma\Delta\Sigma} + \mathbf{W}$. For $(i, j) \in E$, $P(\mathbf{\Delta}, \mathbf{W})$ is differentiable with respect to δ_{ij} and $\partial_{\delta_{ij}} P(\mathbf{\Delta}, \mathbf{W}) = \lambda_0 \operatorname{sign}(\theta_{0,ij})$. For $(i, j) \in E^c$, $P(\mathbf{\Delta}, \mathbf{W})$ is not differentiable with respect to δ_{ij}, but has left and right derivatives, given by $\partial^L_{\delta_{ij}} P(\mathbf{\Delta}, \mathbf{W}) = -\lambda_0$ and $\partial^R_{\delta_{ij}} P(\mathbf{\Delta}, \mathbf{W}) = \lambda_0$. Condition (27) now reduces to

$$\begin{cases} (\mathbf{\Sigma\Delta\Sigma})_{ij} + w_{ij} + s_{ij} = 0, & \text{if } (i, j) \in E, i \ge j, \\ (\mathbf{\Sigma\Delta\Sigma})_{ij} + w_{ij} \in [-\lambda_0, \lambda_0], & \text{if } (i, j) \in E^c, i \ge j, \end{cases} \tag{28}$$

where $\mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G})$, $s_{ij} = \lambda_0 \operatorname{sign}(\theta_{0,ij})$ if $(i, j) \in E$ and $i \ne j$ and $s_{ij} = 0$ if $i = j$.

Now consider the event

$$\begin{aligned} G = \{\{w_{ij} : i \ge j\} : &\text{ Equation (28) is satisfied for some} \\ &\mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G})\}. \end{aligned}$$

We need to show that $\mathbb{P}(G) > 1$. Since $\mathbf{\Delta}$ belongs to $\mathbb{S}^{p\times p}(\mathcal{G})$, it has as many free parameters as there are equations in the first line of Equation (28). Since $\mathbf{\Sigma}$ is a nonsingular matrix, for any $\{w_{ij} : i \ge j, (i, j) \in E\}$, there exists a unique $\mathbf{\Delta}$ that satisfies the first line of Equation (28) and it is a linear function of $\{w_{ij} : (i, j) \in E, i \ge j\}$. Writing this function as $\mathbf{\Delta}(\{w_{ij} : (i, j) \in E, i \ge j\})$, we see that Equation (28) is satisfied if and only if

$$\begin{aligned} &[\mathbf{\Sigma\Delta}(\{w_{\mu\nu} : (\mu, \nu) \in E, \mu \ge \nu\})\mathbf{\Sigma}]_{ij} + w_{ij} \in [-\lambda_0, \lambda_0], \\ &\quad \text{for } i \ge j, \ (i, j) \in E^c. \end{aligned}$$

Since the mapping

$$\begin{aligned} \tau : \{w_{ij} : i \ge j\} \mapsto \{&2[\mathbf{\Sigma\Delta}(\{w_{\mu\nu} : (\mu, \nu) \in E, \mu \ge \nu\})\mathbf{\Sigma}]_{ij} \\ &+ 2w_{ij} : i \ge j, (i, j) \in E^c\} \end{aligned}$$

from $\mathbb{R}^{p(p+1)}$ to $\mathbb{R}^{\operatorname{card}(E^c)/2}$ is continuous, the set $\tau^{-1}[(-\lambda_0, \lambda_0)^{\operatorname{card}(E^c)/2}]$ is open in $\mathbb{R}^{p(p+1)/2}$. Furthermore, this open set is nonempty because if we let

$$\begin{aligned} w_{ij} = &-[\mathbf{\Sigma\Delta}(\{w_{\mu\nu} : (\mu, \nu) \in E, \mu \ge \nu\})\mathbf{\Sigma}]_{ij}, \\ &\text{for } (i, j) \in E^c, i \ge j, \end{aligned}$$

then $\tau(\{w_{ij} : i \ge j\}) = 0$. Because $\{w_{ij} : i \ge j\}$ has a multivariate normal distribution with a nonsingular covariance matrix, any nonempty open set in $\mathbb{R}^{p(p+1)/2}$ has positive probability. This proves $\lim_{n\to\infty} \mathbb{P}(E_n = E) > 0$.

Similarly, the set $\tau^{-1}[(\lambda_0, 3\lambda_0)^{\operatorname{card}(E^c)/2}]$ is open in $\mathbb{R}^{p(p+1)/2}$, and if we let

$$\begin{aligned} w_{ij} = &-[\mathbf{\Sigma\Delta}(\{w_{\mu\nu} : (\mu, \nu) \in E, \mu \ge \nu\})\mathbf{\Sigma}]_{ij} + 2\lambda_0, \\ &\text{for } (i, j) \in E^c, i \ge j, \end{aligned}$$

then $\tau(\{w_{ij} : i \ge j\}) = (2\lambda_0, \ldots, 2\lambda_0)^{\mathsf{T}} \in (\lambda_0, 3\lambda_0)^{\operatorname{card}(E^c)/2}$. Hence, the $\mathbf{W}$-probability of $\tau^{-1}[(\lambda_0, 3\lambda_0)^{\operatorname{card}(E^c)/2}]$ is positive, implying $\lim_{n\to\infty} \mathbb{P}(E_n = E) < 1$.

2. For each $(i, j) \in E^c$, $i \ge j$, Let $U_{ij} = [\mathbf{\Sigma\Delta}(\{w_{\mu\nu} : (\mu, \nu) \in E, \mu \ge \nu\})\mathbf{\Sigma}]_{ij} + w_{ij}$. Then, for any $\eta > 0$, there is a $\lambda_0^{ij} > 0$ such that $\mathbb{P}(U_{ij} \in [-\lambda_0^{ij}, \lambda_0^{ij}]) > 1 - \eta$. Let $\lambda_0 = \max\{\lambda_0^{ij} : (i, j) \in E^c, i \ge j\}$. Then, $\mathbb{P}(U_{ij} \in [-\lambda_0, \lambda_0]) > 1 - \eta$. Hence,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(E_n = E) &= \mathbb{P}(U_{ij} \in [-\lambda_0, \lambda_0], (i, j) \in E^c, i \ge j) \\ &\ge 1 - \sum_{(i,j)\in E^c, i\ge j} \{1 - \mathbb{P}(U_{ij} \in [-\lambda_0, \lambda_0])\} \\ &\ge 1 - \operatorname{card}(E^c)\eta/2. \end{aligned}$$

This proves part 2 because η can be arbitrarily small. ■

7. SPARSISTENCY AND ORACLE PROPERTY: THE ADAPTIVE LASSO

We now turn to the adaptive lasso based on Equation (14). We still use $\hat{\mathbf{\Theta}} = \{\hat{\theta}_{ij}\}$ to denote the minimizer of Equation (14). For a matrix $\mathbf{\Delta} = \{\delta_{ij}\} \in \mathbb{S}^{p\times p}$, and a set $C \in \Gamma \times \Gamma$, let $\boldsymbol{\delta}_C = \{\delta_{ij} : (i, j) \in C, i \ge j\}$, which is to be interpreted as a vector where index i moves first, followed by index j.

Theorem 5. Suppose that $(\mathbf{X}, \mathbf{Y})$ follows CGGM with respect to a graph (Γ, E). Let $\hat{\mathbf{\Theta}}$ be the minimizer of Equation (14), where $\hat{\mathbf{\Sigma}}$ has expansion (24), and

$$\lim_{n\to\infty} n^{1/2}\lambda_n = 0, \quad \lim_{n\to\infty} n^{(1+\gamma)/2}\lambda_n = \infty.$$

Suppose $\tilde{\mathbf{\Theta}}$ in Equation (14) is a $\sqrt{n}$-consistent estimate of $\mathbf{\Theta}_0$ with $\mathbb{P}(\tilde{\theta}_{ij} \ne 0) = 1$ for $(i, j) \in E^c$. Then,

1. $\lim_{n\to\infty} \mathbb{P}(E_n = E) = 1$;
2. $\sqrt{n}\,(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0) \xrightarrow{\mathcal{D}} \operatorname{argmin}\{L(\mathbf{\Delta}, \mathbf{G}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G})\}$.

Part 1 asserts that the adaptive lasso is sparsistent; part 2 asserts that it is asymptotically equivalent to the minimizer of $L(\mathbf{\Delta}, \mathbf{G})$ when E is known. In this sense $\hat{\mathbf{\Theta}}$ is oracle. The condition $P(\tilde{\theta}_{ij} \neq 0) = 1$ means that $\tilde{\theta}_{ij}$ is not sparse, which guarantees that $|\tilde{\theta}_{ij}|^{-\gamma}$ is well defined. For example, we can use the inverse of one of the RKHS estimates as $\tilde{\mathbf{\Theta}}$.

Proof. Let $\mathbf{\Delta} \in \mathbb{S}^{p\times p}$ and $\Psi_n(\mathbf{\Delta}) = n[\Lambda_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - \Lambda_n(\mathbf{\Theta}_0)]$. Consider the difference

$$n[\Pi_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - \Pi_n(\mathbf{\Theta}_0)] = \lambda_n n \sum_{i\neq j} |\tilde{\theta}_{ij}|^{-\gamma}(|\theta_{0,ij} + n^{-1/2}\delta_{ij}| - |\theta_{0,ij}|).$$

By Lemma 6, for large enough n, the right-hand side can be rewritten as

$$\sum_{i\neq j} n^{1/2}\lambda_n |\tilde{\theta}_{ij}|^{-\gamma} \operatorname{sign}(\theta_{0,ij}, n^{-1/2}\delta_{ij})\delta_{ij}. \tag{29}$$

If $(i, j) \in E$, then $|\tilde{\theta}_{ij}|^{-\gamma} = O_P(1)$. Because $n^{1/2}\lambda_n \to 0$, the summand for such (i, j) converges to 0 in probability. Hence, Equation (29) reduces to

$$\sum_{(i,j)\in E^c} n^{1/2}\lambda_n |\tilde{\theta}_{ij}|^{-\gamma} \operatorname{sign}(\delta_{ij})\delta_{ij} + o_P(1) = \sum_{(i,j)\in E^c} n^{1/2}\lambda_n |\tilde{\theta}_{ij}|^{-\gamma}|\delta_{ij}| + o_P(1). \tag{30}$$

If $(i, j) \notin E$, then $\tilde{\theta}_{ij} = O_P(n^{-1/2})$, and hence

$$n^{1/2}\lambda_n |\tilde{\theta}_{ij}|^{-\gamma} = \lambda_n n^{(1+\gamma)/2} |n^{1/2}\tilde{\theta}_{ij}|^{-\gamma} = \frac{\lambda_n n^{(1+\gamma)/2}}{O_P(1)} \xrightarrow{P} \infty.$$

From this, we see that Equation (30) converges in probability to ∞ unless $\boldsymbol{\delta}_{E^c} = 0$, in which case it converges in probability to 0. In other words,

$$n[\Pi_n(\mathbf{\Theta}_0 + n^{-1/2}\mathbf{\Delta}) - \Pi_n(\mathbf{\Theta}_0)] \xrightarrow{P} \begin{cases} 0, & \boldsymbol{\delta}_{E^c} = 0, \\ \infty, & \boldsymbol{\delta}_{E^c} \neq 0, \end{cases}$$

This implies that

$$\Psi_n(\mathbf{\Delta}) \xrightarrow{\mathcal{D}} \Psi(\mathbf{\Delta}, \mathbf{G}), \quad \text{where } \Psi(\mathbf{\Delta}, \mathbf{G}) = \begin{cases} L(\mathbf{\Delta}, \mathbf{G}), & \boldsymbol{\delta}_{E^c} = 0, \\ \infty, & \boldsymbol{\delta}_{E^c} \neq 0. \end{cases}$$

Since both $\Psi_n(\mathbf{\Delta})$ and $\Psi(\mathbf{\Delta}, \mathbf{G})$ are convex and $\Psi(\mathbf{\Delta}, \mathbf{G})$ has a unique minimum, by the epi-convergence results of Geyer (1994), we have

$$\operatorname{argmin}\{\Psi_n(\mathbf{\Delta}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}\} \xrightarrow{\mathcal{D}} \operatorname{argmin}\{\Psi(\mathbf{\Delta}, \mathbf{G}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}\}.$$

This proves part 2 because the right-hand side is, in fact, $\operatorname{argmin}\{L(\mathbf{\Delta}, \mathbf{G}) : \mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G})\}$, and the left-hand side is $\sqrt{n}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$.

The function $\Psi(\mathbf{\Delta}, \mathbf{W})$ is always minimized in a region of $\mathbf{\Delta}$ in which it is not ∞. As a consequence, if $\mathbf{\Delta}$ minimizes $\Psi(\mathbf{\Delta}, \mathbf{W})$ over $\mathbb{S}^{p\times p}$, then $\boldsymbol{\delta}_{E^c} = 0$. Hence, $\mathbb{P}(E_n \supseteq E) \to 1$. In the meantime, since $\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0 = O_P(n^{-\frac{1}{2}})$, we have $\hat{\theta}_{ij} = \theta_{0,ij} + O_P(n^{-\frac{1}{2}})$. If $(i, j) \in E$, then $\theta_{0,ij} \neq 0$. Thus, we see that $\mathbb{P}(E_n \subseteq E) \to 1$. This proves part 1. ■

We now derive the explicit expression of the asymptotic distribution of $\hat{\mathbf{\Theta}}$. Let $\mathbf{G}$ be the unique matrix in $\mathbb{R}^{p^2\times[\operatorname{card}(E)+p]/2}$ such that for any $\mathbf{\Delta} \in \mathbb{S}^{p\times p}(\mathcal{G})$, $\operatorname{vec}(\mathbf{\Delta}) = \mathbf{G}\boldsymbol{\delta}_E$. For example, if $p = 3$ and $E = \{(1, 1), (2, 2), (3, 3), (1, 2), (2, 1), (1, 3), (3, 1)\}$, then $\boldsymbol{\delta}_E = (\delta_{11}, \delta_{21}, \delta_{31}, \delta_{22}, \delta_{33})^{\mathsf{T}}$, and $\mathbf{G}$ is defined by

$$\mathbf{G}^{\mathsf{T}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Corollary 3. Under the assumptions of Theorem 5,

$$\sqrt{n}\operatorname{vec}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0) \xrightarrow{\mathcal{D}} N(0, \mathbf{V}),$$

where

$$\mathbf{V} = \mathbf{G}[\mathbf{G}^{\mathsf{T}}(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\mathbf{G}]^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{\Xi}\mathbf{G}[\mathbf{G}^{\mathsf{T}}(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\mathbf{G}]^{-1}\mathbf{G}^{\mathsf{T}}. \tag{31}$$

Proof. Note that

$$\operatorname{tr}(\mathbf{\Delta}\mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) = \operatorname{vec}^{\mathsf{T}}(\mathbf{\Delta}\mathbf{\Sigma})\operatorname{vec}(\mathbf{\Delta}\mathbf{\Sigma}) = \operatorname{vec}^{\mathsf{T}}(\mathbf{\Delta})(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\operatorname{vec}(\mathbf{\Delta})$$
$$\operatorname{tr}(\mathbf{\Delta}\mathbf{W}) = \operatorname{vec}^{\mathsf{T}}(\mathbf{W})\operatorname{vec}(\mathbf{\Delta}).$$

Hence,

$$L(\mathbf{\Delta}, \mathbf{W}) = \boldsymbol{\delta}_E^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\mathbf{G}\boldsymbol{\delta}_E/2 + \operatorname{vec}^{\mathsf{T}}(\mathbf{W})\mathbf{G}\boldsymbol{\delta}_E.$$

This is a quadratic function minimized by $\boldsymbol{\delta}_E = -[\mathbf{G}^{\mathsf{T}}(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\mathbf{G}]^{-1}\mathbf{G}^{\mathsf{T}}\operatorname{vec}(\mathbf{W})$. In terms of $\mathbf{\Delta}$, the minimizer is $\operatorname{vec}(\mathbf{\Delta}) = -\mathbf{G}[\mathbf{G}^{\mathsf{T}}(\mathbf{\Sigma}^2 \otimes \mathbf{I}_p)\mathbf{G}]^{-1}\mathbf{G}^{\mathsf{T}}\operatorname{vec}(\mathbf{W})$. The corollary now follows from Theorem 5. ■

The asymptotic variance of $\sqrt{n}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$ can be estimated by replacing the moments in Equation (25) and (31) by their sample estimates.

8. CONVERGENCE RATE FOR HIGH-DIMENSIONAL GRAPH

In the last two sections, we have studied the asymptotic properties of our sparse CGGM estimators with the number of nodes p in the graph $\mathcal{G}$ held fixed. We now investigate the case where $p = p_n$ tends to infinity. Due to the limited space, we shall focus on the lasso-penalized RKHS estimator with PC regularization. That is, the minimizer of $\Upsilon_n(\mathbf{\Theta})$ with $\hat{\mathbf{\Sigma}}$ in Equation (12) is taken to be the RKHS estimator $\hat{\mathbf{\Sigma}}_{\text{PC}}(\epsilon_n)$ based on $\kappa_X(\mathbf{a}, \mathbf{b}) = (1 + \mathbf{a}^{\mathsf{T}}\mathbf{b})^r$. Throughout this section, $\hat{\mathbf{\Theta}}$ denotes this estimator. The large-p_n convergence rate for the lasso estimator of the (unconditional) GGM has been studied by Rothman et al. (2008).

In this case, $\mathbf{\Sigma}$, $\mathbf{\Theta}$, and $\mathbf{Y}$ should in principle be written as $\mathbf{\Sigma}^{(n)}$, $\mathbf{\Theta}^{(n)}$, and $\mathbf{Y}^{(n)}$ because they now depend on n. However, to avoid complicated notation, we still use $\mathbf{\Sigma}$, $\mathbf{\Theta}$, and $\mathbf{Y}$, keeping in mind their dependence on n. Following Rothman et al. (2008), we develop the convergence rate in Frobenius norm. Let $\|\cdot\|_1$, $\|\cdot\|_F$, and $\|\cdot\|_\infty$ be the L_1-norm, Frobenius norm, and the

L_∞-norm of a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$:

$$\|\mathbf{A}\|_1 = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}|a_{ij}|, \quad \|\mathbf{A}\|_F = \left(\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}a_{ij}^2\right)^{1/2},$$
$$\|\mathbf{A}\|_\infty = \max|a_{ij}|,$$

where a_{ij} denotes the (i, j)th entry of $\mathbf{A}$. Let $\rho(\mathbf{A})$ denote the number of nonzero entries of $\mathbf{A}$. For easy reference, we list some properties of these matrix functions in the following proposition.

Proposition 1. Let $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{B} \in \mathbb{R}^{d_2 \times d_3}$. Then,

$$\|\mathbf{AB}\|_\infty \le d_2\|\mathbf{A}\|_\infty\|\mathbf{B}\|_\infty, \quad |\mathrm{tr}(\mathbf{AB})| \le \|\mathbf{A}\|_1\|\mathbf{B}\|_\infty,$$
$$\|\mathbf{A}\|_1 \le \sqrt{\rho(\mathbf{A})}\|\mathbf{A}\|_F.$$

The first inequality follows from the definition of $\|\cdot\|_\infty$; the second from Hölder's inequality; the third from the Cauchy-Schwarz inequality. The last two inequalities were used in Rothman et al. (2008). Let $\mathbb{N} = \{1, 2, \ldots\}$. Consider the array of random matrices: $\{\mathbf{A}_i^{(n)} : i = 1, \ldots, n,\ n \in \mathbb{N}\}$, where $\mathbf{A}^{(n)} \in \mathbb{R}^{p_n \times q}$, p_n may depend on n but q is fixed. Let $(\mathbf{A}^{(n)})_{rs}$ denote the (r, s)th entry of $\mathbf{A}^{(n)}$.

Lemma 7. Let $\{\mathbf{A}_i^{(n)} : i = 1, \ldots, n,\ n \in \mathbb{N}\}$ be an array of random matrices in $\mathbb{R}^{p_n \times q}$, each of whose rows is an iid sample of a random matrix $\mathbf{A}^{(n)}$. Suppose that the moment generating functions of $(\mathbf{A}^{(n)})_{st}$, say ϕ_{nst}, are finite on an interval $(-\delta, \delta)$, and their second derivatives are uniformly bounded over this interval for all $s = 1, \ldots, p_n, t = 1, \ldots, q, n \in \mathbb{N}$. If $p_n \to \infty$, then $\|\mathbb{E}_n(\mathbf{A}^{(n)}) - \mathbb{E}(\mathbf{A}^{(n)})\|_\infty = O_P(\log p_n/\sqrt{n})$.

Proof. Let $\mu_{nst} = \mathbb{E}(\mathbf{A}^{(n)})_{st}$. Since $|\mu_{nst}| \le 1 + \phi''_{nst}(0)$, there is $C > 0$ such that $|\mu_{nst}| \le C$ for all s, t, n. Let $(\mathbf{B}^{(n)})_{st} = (\mathbf{A}^{(n)})_{st} - \mu_{nst}$, and ψ_{nst} be the moment generating function of $(\mathbf{B}^{(n)})_{st}$. Then, $\psi_{nst}(\tau) = e^{-\mu_{nst}\tau}\phi_{nst}(\tau)$. For any $a > 0$,

$$\mathbb{P}\left(\sqrt{n}\left(\mathbb{E}_n\left(\mathbf{B}^{(n)}\right)_{st}\right) > a\right) = \mathbb{P}\left(e^{\sqrt{n}\mathbb{E}_n(\mathbf{B}^{(n)})_{st}} > e^a\right)$$
$$\le e^{-a}\left[\psi_{nst}\left(n^{-1/2}\right)\right]^n.$$

By Taylor's theorem and noticing that $\psi'_{nst}(0) = 0$, we have

$$\psi_{nst}(n^{-1/2}) = 1 + \psi''_{nst}(\xi_{nst})/(2n) \le 1 + e^C\phi''_{nst}(\xi_{nst})/(2n)$$

for some $0 \le \xi_{nst} \le n^{-1/2}$. By assumption, there is $C_1 > 0$ such that $\limsup_{n\to\infty}\phi''(\xi_{nst}) \le C_1$ for all $s = 1, \ldots, p_n$ and $t = 1, \ldots, q$. Hence,

$$[\psi_{nst}(n^{-1/2})]^n \le [1 + e^C C_1/(2n)]^n \to e^{e^C C_1/2} \equiv C_2.$$

Thus, we have $\limsup_{n\to\infty} P(\sqrt{n}\mathbb{E}_n(\mathbf{B}^{(n)})_{st} > a) \le e^{-a}C_2$. By the same argument, we can show that $\limsup_{n\to\infty} P(\sqrt{n}\mathbb{E}_n(\mathbf{B}^{(n)})_{st} < -a) \le e^{-a}C_2$. Therefore,

$$\limsup_{n\to\infty} P\left(\sqrt{n}|\mathbb{E}_n\left(\mathbf{B}^{(n)}\right)_{st}| > a\right) \le 2e^{-a}C_2.$$

It follows that

$$\limsup_{n\to\infty}\mathbb{P}\left(\|\mathbb{E}_n\left(\mathbf{B}^{(n)}\right)\|_\infty > c_n\right)$$
$$\le \limsup_{n\to\infty}\sum_{s=1}^{p_n}\sum_{t=1}^{q}\mathbb{P}\left(\sqrt{n}|\mathbb{E}_n\left[\left(\mathbf{A}^{(n)}\right)_{st}\right| > \sqrt{n}c_n\right)$$
$$\le 2qC_2e^{-\sqrt{n}c_n + \log p_n}.$$

In particular,

$$\limsup_{n\to\infty}\mathbb{P}\left(\|\mathbb{E}_n\left(\mathbf{B}^{(n)}\right)\|_\infty > 2\log p_n/\sqrt{n}\right) \le qC_2/p_n \to 0,$$

which implies the desired result. ■

In the following, we call any array of random matrices satisfying the conditions in Lemma 7 a *standard* array. We now establish the convergence rate of $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F$. Let s_n denote the number of nonzero off-diagonal entries of $\mathbf{\Theta}_0$. Let $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}_Y - E(\mathbf{Y} \mid \mathbf{X})$. Let X^t and Y^s denote the components of $\mathbf{X}$ and $\mathbf{Y}$. Their powers are denoted by $(X^t)^r$ and $(Y^s)^r$. For a symmetric matrix $\mathbf{A}$, let $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues of $\mathbf{A}$.

Theorem 6. Let $\hat{\mathbf{\Theta}}$ be the sparse estimator defined in the first paragraph of this section with $\lambda_n \sim (\log p_n/n)^{1/2}$, $\epsilon_n = o(n)$, and $\kappa(\mathbf{a}, \mathbf{b}) = (1 + \mathbf{a}^\mathsf{T}\mathbf{b})^r$. Suppose that $(\mathbf{X}, \mathbf{Y})$ follows a CGGM, and satisfies the following additional assumptions:

1. $\mathbf{Y}$, $\mathbf{YU}^\mathsf{T}$, and $\mathbf{ZU}^\mathsf{T}$ are standard arrays of random matrices;
2. for all $n \in \mathbb{N}$, $\sigma_{\max}(\mathbf{\Sigma}) < \infty$ and $\sigma_{\min}(\mathbf{\Sigma}) > 0$;
3. $p_n \to \infty$ and $p_n(p_n + s_n)^{1/2}(\log p_n)^{5/2} = o(n^{3/2})$;
4. the fixed-dimensional matrix $\mathbf{V}_U = \mathrm{var}(\mathbf{U})$ is nonsingular;
5. each component of $\mathbb{E}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ is a polynomial in $x^1, \ldots, x^q$ of at most rth order.

Then, $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F = O_P([(p_n + s_n)\log p_n/n]^{1/2})$.

Note that we can allow $p_n(p_n + s_n)^{1/2}$ to get arbitrarily close to $n^{3/2}$. This condition is slightly stronger than the corresponding condition in Rothman et al. (2008) for the unconditional case, which requires $(p_n + s_n)\log p_n = o(n^{1/2})$. Also note that $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F$ is the *sum*, instead of *average*, of p_n^2 elements (roughly $p_n + s_n$ nonzero elements). With this in mind, the convergence rate $[(p_n + s_n)\log p_n/n]^{1/2}$ is quite fast. This is the same rate as that given in Rothman et al. (2008) for the unconditional GGM.

Proof of Theorem 3. Let $r_n = [(p_n + s_n)\log p_n/n]^{1/2}$, and $\mathcal{D}_n = \{\mathbf{\Delta} \in \mathbb{S}^{p_n \times p_n} : \|\mathbf{\Delta}\|_2 = Mr_n\}$ for some $M > 0$. Let,

$$\mathbf{G}_n(\mathbf{\Delta}) = \Upsilon_n(\mathbf{\Theta}_0 + \mathbf{\Delta}) - \Upsilon_n(\mathbf{\Theta}_0),$$

where

$$\Upsilon_n(\mathbf{\Theta}) = -\log\det(\mathbf{\Theta}) + \mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)) + \lambda_n\sum_{i\ne j}|\theta_{ij}|. \quad (32)$$

Then, $\hat{\mathbf{\Theta}}$ minimizes $L(\mathbf{\Theta})$ if and only if $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Theta}} - \mathbf{\Theta}_0$ minimizes $\mathbf{G}_n(\mathbf{\Delta})$. As argued by Rothman et al. (2008), since $\mathbf{G}_n(\mathbf{\Delta})$ is convex in $\mathbf{\Delta}$ and $\mathbf{G}_n(\hat{\mathbf{\Delta}}) \le 0$, the minimizer $\hat{\mathbf{\Delta}}$ resides within the sphere $\mathcal{D}_n$ if $G_n(\mathbf{\Delta})$ is positive and bounded away from 0 on this sphere. That is, it suffices to show

$$\mathbb{P}(\inf\{G_n(\mathbf{\Delta}) : \mathbf{\Delta} \in \mathcal{D}_n\} > 0) \to 1.$$

The proof of Lemma 3 shows, in the context of fixed p, that $(\mathbf{QK}_X\mathbf{Q})^\dagger_{\epsilon_n} = (\mathbf{QD}_U\mathbf{CD}_U^\mathsf{T}\mathbf{Q})^\dagger$ with probability tending to 1 if $\epsilon_n = o(n)$. This result still holds here because the dimension q of $\mathbf{X}$ remains fixed. Consequently $\mathbb{P}(\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n) = \tilde{\mathbf{\Sigma}}) \to 1$, where $\tilde{\mathbf{\Sigma}}$ is as defined in Equation (16) but now its dimension increases with n. Thus, we can replace the $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ in Equation (32) by $\tilde{\mathbf{\Sigma}}$. Let

$$\hat{\boldsymbol{\mu}}_U = \mathbb{E}_n(\mathbf{U}), \quad \hat{\mathbf{V}}_{YU} = n^{-1}\mathbf{D}_Y^\mathsf{T}\mathbf{QD}_U, \quad \hat{\mathbf{V}}_U = n^{-1}\mathbf{D}_U^\mathsf{T}\mathbf{QD}_U.$$

Let $\hat{\boldsymbol{\mu}}_{Y|U} = \hat{\mathbf{V}}_{YU}\hat{\mathbf{V}}_U^{-1}(\mathbf{U} - \hat{\boldsymbol{\mu}}_U)$, and $\boldsymbol{\mu}_{Y|U} = E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{V}_{YU}\mathbf{V}_U^{-1}(\mathbf{U} - \boldsymbol{\mu}_U)$. Then,

$$\tilde{\boldsymbol{\Sigma}} = \mathbb{E}_n(\mathbf{Z} + \boldsymbol{\mu}_{Y|U} - \hat{\boldsymbol{\mu}}_{Y|U} + \boldsymbol{\mu}_Y - \hat{\boldsymbol{\mu}}_Y)(\mathbf{Z} + \boldsymbol{\mu}_{Y|U} - \hat{\boldsymbol{\mu}}_{Y|U} + \boldsymbol{\mu}_Y - \hat{\boldsymbol{\mu}}_Y)^{\mathsf{T}}.$$

The term $\boldsymbol{\mu}_{Y|U} - \hat{\boldsymbol{\mu}}_{Y|U}$ can be further decomposed as $\mathbf{Z}_I + \mathbf{Z}_{II} + \mathbf{Z}_{III}$, where

$$\begin{aligned}
\mathbf{Z}_I &= -(\hat{\mathbf{V}}_{YU} - \mathbf{V}_{YU})\hat{\mathbf{V}}_U^{-1}(\mathbf{U} - \hat{\boldsymbol{\mu}}_U),\\
\mathbf{Z}_{II} &= \mathbf{V}_{YU}[\mathbf{V}_U^{-1}(\mathbf{U} - \boldsymbol{\mu}_U) - \hat{\mathbf{V}}_U^{-1}(\mathbf{U} - \hat{\boldsymbol{\mu}}_U)],\\
\mathbf{Z}_{III} &= \boldsymbol{\mu}_Y - \hat{\boldsymbol{\mu}}_Y.
\end{aligned}$$

The function $G_n(\boldsymbol{\Delta})$ can now be rewritten as

$$\begin{aligned}
G_n(\boldsymbol{\Delta}) = L_n^*(\boldsymbol{\Theta}_0 + \boldsymbol{\Delta}) - L_n^*(\boldsymbol{\Theta}_0) &+ \sum_{v\in\{I,II,III\}} \operatorname{tr}\big(\boldsymbol{\Delta}\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_v^{\mathsf{T}}\big)\big)\\
&+ \sum_{v\in\{I,II,III\}} \operatorname{tr}(\boldsymbol{\Delta}\mathbb{E}_n(\mathbf{Z}_v\mathbf{Z}^{\mathsf{T}}))\\
&+ \sum_{v\in\{I,II,III\}}\sum_{\tau\in\{I,II,III\}} \operatorname{tr}\big(\boldsymbol{\Delta}\mathbb{E}_n\big(\mathbf{Z}_v\mathbf{Z}_\tau^{\mathsf{T}}\big)\big),
\end{aligned}$$

where $L_n^*(\boldsymbol{\Theta}) = \operatorname{tr}(\boldsymbol{\Theta}\mathbb{E}_n(\mathbf{Z}\mathbf{Z}^{\mathsf{T}})) - \log\det(\boldsymbol{\Theta}) + \lambda_n\sum_{i\neq j}|\theta_{ij}|$. Since $\mathbf{Z} \sim N(0, \boldsymbol{\Sigma})$, we can use the same argument in the proof of Theorem 1 in Rothman et al. (2008) to show that

$$\mathbb{P}\left(\inf\{L_n^*(\boldsymbol{\Theta}_0 + \boldsymbol{\Delta}) - L_n^*(\boldsymbol{\Theta}_0) : \boldsymbol{\Delta} \in \mathcal{D}_n\} > 0\right) \to 1.$$

Thus, our theorem will be proved if we can show that

$$\sup\{|\operatorname{tr}(\boldsymbol{\Delta}\mathbb{E}_n(\mathbf{A}))| : \boldsymbol{\Delta} \in \mathcal{D}_n\} = o_P(1)$$

for $\mathbf{A}$ being any one of the following eight random matrices

$$\mathbf{Z}\mathbf{Z}_v^{\mathsf{T}},\ \mathbf{Z}_v\mathbf{Z}^{\mathsf{T}},\ \mathbf{Z}_v\mathbf{Z}_\tau^{\mathsf{T}},\quad v, \tau = I, II, III. \tag{33}$$

By Proposition 1, inequalities (2) and (3), we have, for $\boldsymbol{\Delta} \in \mathcal{D}_n$,

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Delta}\mathbb{E}_n(\mathbf{A}))| &\le \|\mathbb{E}_n(\mathbf{A})\|_\infty \|\boldsymbol{\Delta}\|_1 \le \|\mathbb{E}_n(\mathbf{A})\|_\infty \sqrt{\rho(\boldsymbol{\Delta})}\|\boldsymbol{\Delta}\|_2\\
&\le M\|\mathbb{E}_n(\mathbf{A})\|_\infty p_n r_n.
\end{aligned}$$

Thus, it suffices to show that,

$$\|\mathbb{E}_n(\mathbf{A})\|_\infty p_n(p_n + s_n)^{1/2}(\log p_n)^{1/2}n^{-1/2} = o_P(1). \tag{34}$$

Since $\|\mathbb{E}_n(\mathbf{A})\|_\infty = \|\mathbb{E}_n(\mathbf{A}^{\mathsf{T}})\|_\infty$, we only need to consider the following $\mathbf{A}$:

$$\mathbf{Z}\mathbf{Z}_I^{\mathsf{T}}, \mathbf{Z}\mathbf{Z}_{II}^{\mathsf{T}}, \mathbf{Z}\mathbf{Z}_{III}^{\mathsf{T}}, \mathbf{Z}_I\mathbf{Z}_I^{\mathsf{T}}, \mathbf{Z}_I\mathbf{Z}_{II}^{\mathsf{T}}, \mathbf{Z}_I\mathbf{Z}_{III}^{\mathsf{T}}, \mathbf{Z}_{II}\mathbf{Z}_{II}^{\mathsf{T}}, \mathbf{Z}_{II}\mathbf{Z}_{III}^{\mathsf{T}}, \mathbf{Z}_{III}\mathbf{Z}_{III}^{\mathsf{T}}.$$

From the definitions of $\mathbf{Z}$ and $\mathbf{Z}_I$, we have

$$\begin{aligned}
\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_I^{\mathsf{T}}\big) = &-\mathbb{E}_n[\mathbf{Z}(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}}]\hat{\mathbf{V}}_U^{-1}(\hat{\mathbf{V}}_{UY} - \mathbf{V}_{UY})\\
&- \hat{\boldsymbol{\mu}}_Z(\boldsymbol{\mu}_U - \hat{\boldsymbol{\mu}}_U)^{\mathsf{T}}\hat{\mathbf{V}}_U^{-1}(\hat{\mathbf{V}}_{UY} - \mathbf{V}_{UY}),
\end{aligned} \tag{35}$$

where $\hat{\boldsymbol{\mu}}_Z = \mathbb{E}_n(\mathbf{Z})$. By the first inequality of Proposition 1,

$$\begin{aligned}
\|\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_I^{\mathsf{T}}\big)\|_\infty \le\ & q^2\|\mathbb{E}_n[\mathbf{Z}(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}}]\|_\infty\\
&\times \|\hat{\mathbf{V}}_U^{-1}\|_\infty\|\hat{\mathbf{V}}_{UY} - \mathbf{V}_{UY}\|_\infty + q^2\|\hat{\boldsymbol{\mu}}_Z\|_\infty\|(\boldsymbol{\mu}_U - \hat{\boldsymbol{\mu}}_U)^{\mathsf{T}}\|_\infty\\
&\times \|\hat{\mathbf{V}}_U^{-1}\|_\infty\|\hat{\mathbf{V}}_{UY} - \mathbf{V}_{UY}\|_\infty.
\end{aligned} \tag{36}$$

Since $\mathbf{Z} \perp\!\!\!\perp \mathbf{X}$, we have $E[\mathbf{Z}(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}}] = 0$. Hence, by Lemma 7,

$$\|\mathbb{E}_n[\mathbf{Z}(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}}]\|_\infty = O_P(n^{-\frac{1}{2}}\log p_n). \tag{37}$$

Similarly,

$$\begin{aligned}
\|\mathbb{E}_n[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}} - \mathbf{V}_{YU}]\|_\infty &= O_P(n^{-\frac{1}{2}}\log p_n),\\
\|\hat{\boldsymbol{\mu}}_Y - \boldsymbol{\mu}_Y\|_\infty &= O_P(n^{-\frac{1}{2}}\log p_n).
\end{aligned} \tag{38}$$

Since $\hat{\boldsymbol{\mu}}_U - \boldsymbol{\mu}_U$ has a fixed-dimension finite-variance matrix, by the central limit theorem,

$$\|\hat{\boldsymbol{\mu}}_U - \boldsymbol{\mu}_U\|_\infty = O_P(n^{-1/2}). \tag{39}$$

By definition,

$$\begin{aligned}
\hat{\mathbf{V}}_{YU} - \mathbf{V}_{YU} = &\mathbb{E}_n[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{U} - \boldsymbol{\mu}_U)^{\mathsf{T}} - \mathbf{V}_{YU}]\\
&- (\hat{\boldsymbol{\mu}}_Y - \boldsymbol{\mu}_Y)(\hat{\boldsymbol{\mu}}_U - \boldsymbol{\mu}_U)^{\mathsf{T}}.
\end{aligned}$$

By Proposition 1, first inequality,

$$\begin{aligned}
\|\hat{\mathbf{V}}_{YU} - \mathbf{V}_{YU}\|_\infty \le &\|\mathbb{E}_n[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{U} - \boldsymbol{\mu}_U) - \mathbf{V}_{YU}]\|_\infty\\
&+ \|\hat{\boldsymbol{\mu}}_Y - \boldsymbol{\mu}_Y\|_\infty\|\hat{\boldsymbol{\mu}}_U - \boldsymbol{\mu}_U\|_\infty.
\end{aligned}$$

Substituting Equations (38) and (39) into the above inequality, we find

$$\|\hat{\mathbf{V}}_{YU} - \mathbf{V}_{YU}\|_\infty = O_P(n^{-\frac{1}{2}}\log p_n). \tag{40}$$

Since $\mathbf{Z}$ is multivariate normal whose components have means 0 and bounded variance, it is a standard array. Hence,

$$\|\hat{\boldsymbol{\mu}}_Z\|_\infty = O_P(n^{-\frac{1}{2}}\log p_n). \tag{41}$$

Since the dimension of $\hat{\mathbf{V}}_U$ is fixed, its entries have finite variances, and $\mathbf{V}_U$ is nonsingular, we have, by the central limit theorem,

$$\|\hat{\mathbf{V}}_U^{-1}\|_\infty = O_P(1). \tag{42}$$

Substituting Equations (37), (39), (40), (41), and (42) into Equation (36), we find that

$$\|\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_I^{\mathsf{T}}\big)\|_\infty = O_P(n^{-1}(\log p_n)^2),$$

which, by condition (3), satisfies Equation (34).

The order of magnitudes of the rest of the three terms can be derived similarly. We present the results below, omitting the details:

$$\begin{aligned}
\|\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_{II}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}\log p_n),\\
\|\mathbb{E}_n\big(\mathbf{Z}\mathbf{Z}_{III}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}(\log p_n)^2),\\
\|\mathbb{E}_n\big(\mathbf{Z}_I\mathbf{Z}_I^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}(\log p_n)^2),\\
\|\mathbb{E}_n\big(\mathbf{Z}_I\mathbf{Z}_{II}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}\log p_n),\\
\|\mathbb{E}_n\big(\mathbf{Z}_I\mathbf{Z}_{III}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-3/2}(\log p_n)^2),\\
\|\mathbb{E}_n\big(\mathbf{Z}_{II}\mathbf{Z}_{II}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}),\\
\|\mathbb{E}_n\big(\mathbf{Z}_{II}\mathbf{Z}_{III}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-3/2}\log p_n),\\
\|\mathbb{E}_n\big(\mathbf{Z}_{III}\mathbf{Z}_{III}^{\mathsf{T}}\big)\|_\infty &= O_P(n^{-1}(\log p_n)^2).
\end{aligned}$$

By condition (3), all of these terms satisfy the relation in Equation (34). ■

9. IMPLEMENTATION

In this section, we address two issues in implementation: the choice of the tuning parameter and the minimization of the objective functions (13) and (14). For the choice of the tuning parameter, we use a BIC-type criterion (Schwarz 1978) similar to that used in Yuan and Lin (2007). Let $\hat{\boldsymbol{\Theta}}(\lambda) = \{\hat{\theta}_{ij}(\lambda) : i,$

$j \in \Gamma\}$ be the lasso or the adaptive lasso estimate of $\mathbf{\Theta}_0$ in the conditional graphical model for a specific choice of λ of the tuning parameter. Let $E_n(\lambda) = \{(i, j) : \hat{\theta}_{ij}(\lambda) \neq 0\}$, and

$$\mathrm{BIC}(\lambda) = -\log\det[\hat{\mathbf{\Theta}}(\lambda)] + \mathrm{tr}[\hat{\mathbf{\Theta}}(\lambda)\mathbf{\Sigma}_n] + \log n \frac{\mathrm{card}[E_n(\lambda)] + p}{2n}.$$

The tuning parameter is then chosen to be

$$\hat{\lambda}_{\mathrm{BIC}} = \mathrm{argmin}\{\mathrm{BIC}(\lambda) : \lambda \in (0, \infty)\}.$$

Practically, we evaluate this criterion on a grid of points in $(0, \infty)$ and choose the minimizer among these points.

For the minimization of Equations (13) and (14), we follow the graphical lasso procedure proposed by Friedman et al. (2008), but with the sample covariance matrix therein replaced by the RKHS estimates, $\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(\epsilon_n)$ or $\hat{\mathbf{\Sigma}}_{\mathrm{RR}}(\epsilon_n)$, of the conditional covariance matrix $\mathbf{\Sigma}$. The graphical lasso (glasso) procedure is available as a package in the R language.

10. SIMULATION STUDIES

In this section, we compare the sparse estimators for the CGGM with the sparse estimators of the GGM, with the maximum likelihood estimators of the CGGM, and with two naive estimators. We also explore several reproducing kernels and investigate their performances for estimating the CGGM.

10.1 Comparison With Estimators for GGM

We use three criteria for this comparison:

1. False positive rate at $\hat{\lambda}_{\mathrm{BIC}}$. This is defined as the percentage of edges identified as belonging to E when they are not; that is,

$$\mathrm{FP} = \mathrm{card}\{(i, j) : i > j, \theta_{0,ij} = 0, \hat{\theta}_{ij} \neq 0\}/\mathrm{card}\{(i, j) : i > j, \theta_{0,ij} = 0\}.$$

2. False negative rate at $\hat{\lambda}_{\mathrm{BIC}}$. This is defined as the percentage of edges identified as belonging to E^c when they are not; that is,

$$\mathrm{FN} = \mathrm{card}\{(i, j) : i > j, \theta_{0,ij} \neq 0, \hat{\theta}_{ij} = 0\}/\mathrm{card}\{(i, j) : i > j, \theta_{0,ij} \neq 0\}.$$

3. Rate of correct paths. The above two criteria are both specific to the tuning method (in our case, BIC). To assess the potential capability of an estimator of E, independently of the tuning methods used, we use the percentage of cases where E belongs to the path $\{E_n(\lambda) : \lambda \in (0, \infty)\}$, where $E_n(\lambda)$ is an estimator of E for a fixed λ. We write this criterion as PATH.

Example 1. This is the example, illustrated in Figure 1, in which $p = 3$, $q = 1$, and $(X, \mathbf{Y})$ satisfies CGGM with $E = \{(1, 1), (2, 2), (3, 3), (2, 3), (3, 2)\}$. The conditional distribution of $\mathbf{Y} \mid \mathbf{X}$ is specified by

$$\mathbf{Y} = \boldsymbol{\beta} X + \boldsymbol{\varepsilon}, \tag{43}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, 0)^{\mathsf{T}}$, $X \sim N(0, 1)$, and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma})$, $X \perp\!\!\!\perp \boldsymbol{\varepsilon}$, and

$$\mathbf{\Sigma}^{-1} = \mathbf{\Theta} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & 2.53 \\ 0 & 2.53 & 2 \end{pmatrix}.$$

For each simulated sample, β_1 and β_2 are generated, independently, from the uniform distribution defined on the set $(-6, -3) \cup (3, 6)$. We use two sample sizes $n = 50, 100$. The linear kernel $\kappa_X(\mathbf{a}, \mathbf{b}) = 1 + \mathbf{a}^{\mathsf{T}}\mathbf{b}$ is used for the initial RKHS estimate. The results are presented in the first three rows of Table 1. Entries in the table are the means calculated across 200 simulated samples.

The table indicates that the unconditional sparse estimators have much higher false positive rate than false negative rate. This is because they tend to pick up connections among the components of $\mathbf{Y}$ that are due to X. In comparison, the conditional sparse estimators (both lasso and adaptive lasso) can successfully remove the edges effected by X, resulting in more accurate identification of the graph.

Example 2. In this example, we consider three scenarios in which the effect of an external source on the network varies in degree, resulting in different amounts of gain achievable by a conditional graphical model.

We still assume the linear regression model (43), but with $p = 5$. The first scenario is shown in the top two panels in Figure 2, where all the components of $\boldsymbol{\beta}$ are nonzero and $\mathbf{\Sigma}$ diagonal. In this case, the conditional graph is totally disconnected (left panel); whereas the unconditional graph is a complete graph (right panel). Specifically, the parameters are

$$\mathbf{\Sigma} = \mathbf{\Theta} = \mathbf{I}_5, \quad \boldsymbol{\beta} = (0.656, 0.551, 0.757, 0.720, 0.801)^{\mathsf{T}}.$$

The panels in the third row of Figure 2 represent the other extreme, where only Y^1 is related to X, and each Y^i for $i \neq 1$ shares an edge with Y^1 but has no other edges. In this case, the conditional and unconditional graphs are identical. The parameters are specified as follows. The first row (and first column) of $\mathbf{\Theta}$ is $(6.020, -0.827, -1.443, -1.186, -0.922)$; the remaining 4×4 block is $\mathbf{I}_4$. The first entry of $\boldsymbol{\beta}$ is 0.656, and rest entries are 0. Between these two extremes is scenario 2 (second row in Figure 2), where the conditional and unconditional graphs differ only by one edge: $1 \leftrightarrow 4$. The parameters are

$$\mathbf{\Theta} = \begin{pmatrix} 4.487 & -1.186 & -1.443 & 0 & 0 \\ -1.186 & 1.464 & -0.668 & 0.752 & -0.681 \\ -1.443 & -0.668 & 1.963 & -1.084 & 0.981 \\ 0 & 0.752 & -1.084 & 2.220 & -1.105 \\ 0 & -0.681 & 0.981 & -1.105 & 1 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} 0.656 \\ 0.551 \\ 0 \\ 0.601 \\ 0 \end{pmatrix}.$$

For each scenario, we generate 200 samples of sizes $n = 50, 100$ and compute the three criteria across the 200 samples. The results are presented row 4 through row 12 of Table 1. We

Table 1. Comparison of graph estimation accuracy among the lasso and the adaptive lasso estimators of GGM and CGGM for Example 1, Example 2 (including three scenarios), and Example 3. "ALASSO" means adaptive lasso

		$n = 50$				$n = 100$			
		LASSO		ALASSO		LASSO		ALASSO	
Example/scenario	Criteria	GGM	CGGM	GGM	CGGM	GGM	CGGM	GGM	CGGM
	FP	1	0.51	1	0.15	1	0.40	1	0.05
EX1	FN	0	0	0	0	0	0	0	0
	PATH	0	1	0	1	1	1	0	1
	FP	0.84	0.02	0.54	0.06	0.93	0.01	0.67	0.02
EX2-SC1	FN	0	0	0	0	0	0	0	0
	PATH	0	1	0.01	0.63	0	1	0.98	1
	FP	0.98	0.55	0.96	0.31	1	0.51	1	0.22
EX2-SC2	FN	0.06	0.01	0.15	0.02	0.02	0	0.08	0
	PATH	0	0.57	0	0.71	0	0.79	0	0.98
	FP	0.71	0.68	0.18	0.19	0.75	0.77	0.10	0.11
EX2-SC3	FN	0	0	0.01	0.02	0	0	0	0
	PATH	0	0.43	0.80	0.52	0	0.56	1	0.87
	FP	0.79	0.23	0.41	0.09	0.83	0.16	0.59	0.03
EX3	FN	0	0	0	0	0	0	0	0
	PATH	0	0.94	0.95	0.99	0	1	1	1

see significant improvements of the rates of correct identification of the graphical structure by the sparse estimator of CGGM whenever the conditional graph differs from the unconditional graph. Also note that, for scenario 3, where the two graphs are the same, the adaptive lasso estimator for GGM performs better than the adaptive lasso estimator for CGGM. This is because the latter needs to estimate more parameters.

Example 3. In this example, we investigate a situation where the edges in the unconditional graphical model have two external sources. We use model (43) with $p = 5, q = 2, \mathbf{X} \sim N(0, \mathbf{I}_2)$,

$$\mathbf{\Theta} = \begin{pmatrix} 1.683 & -0.827 & 0 & 0 & 0 \\ -0.827 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1.722 & 0.849 & 0 \\ 0 & 0 & 0.849 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\beta = \begin{pmatrix} 0.656 & 0 \\ 0 & 0.551 \\ 0.757 & 0 \\ 0 & 0.720 \\ 0 & 0.800 \end{pmatrix}.$$

The results are summarized in the last three rows of Table 1, from which we can see similar improvements by the sparse CGGM estimators.

10.2 Comparison With Maximum Likelihood Estimates of CGGM

In Section 7, we showed that the adaptive lasso estimate for the CGGM possesses oracle property; that is, its asymptotic variance reaches the lower bound among regular estimators when the graph $\mathcal{G}$ is *assumed known*. Hence, it makes sense to compare adaptive lasso estimate with the maximum likelihood estimate of $\mathbf{\Theta}$ under the constraints $\theta_{ij} = 0, (i, j) \notin E$, which is known to be optimal among regular estimates. In this section, we make such a comparison, using all three examples in Section 10.1. As a benchmark, we also compare these two estimates with the maximum likelihood estimate under the full model, which for the linear kernel is $[\hat{\mathbf{\Sigma}}_{\mathrm{PC}}(0)]^{\dagger}$. For this comparison we use the squared Frobenius norm $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F^2$, which characterizes the closeness of two precision matrices rather than that of graphs.

Table 2 shows that the adaptive lasso estimator is rather close to the constrained MLE, with the unconstrained MLE trailing noticeably behind. In three out of five cases, the constrained MLE performs better than the adaptive lasso, which is not surprising because, although the two estimators are equivalent asymptotically, the former employs the true graphical structure unavailable for adaptive lasso, making it more accurate for the finite sample. When the errors of adaptive lasso are lower than the constrained MLE, the differences are within the margins of error.

Table 2. Comparison of parameter estimation accuracy among adaptive lasso, and unconstrained and constrained MLE for CGGM. Entries are of the form $a \pm b$, where a is the mean, and b the standard deviation, of criterion $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F^2$ computed from 200 simulated samples

		MLE	
Example/scenario	ALASSO	Unconstrained	Constrained
EX1	1.458 ± 0.125	2.256 ± 0.171	1.575 ± 0.155
EX2-SC1	0.142 ± 0.008	0.400 ± 0.018	0.122 ± 0.007
EX2-SC2	1.858 ± 0.120	2.294 ± 0.148	1.663 ± 0.116
EX2-SC3	1.147 ± 0.071	2.139 ± 0.186	0.969 ± 0.081
EX3	0.303 ± 0.016	0.773 ± 0.555	0.327 ± 0.275

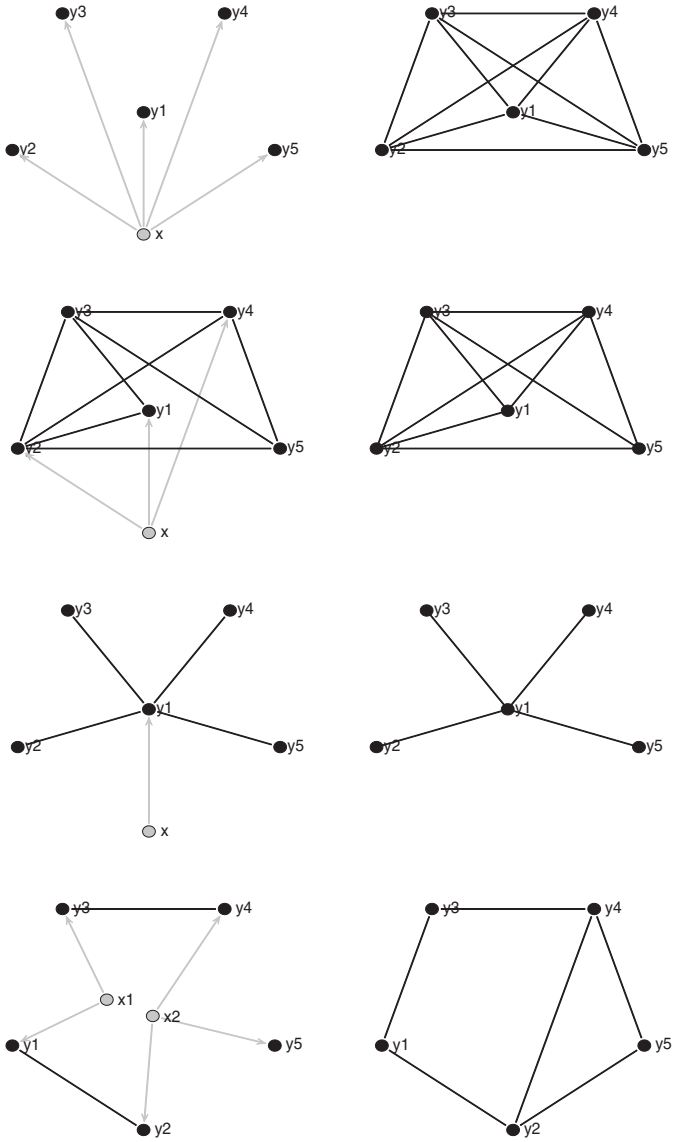

Figure 2. Conditional and unconditional graphical models in Examples 2 and 3. Left panels: the conditional graphical models. Right panels: the corresponding unconditional graphical models. Black nodes indicate response variables; gray nodes are the predictors. Edges in the conditional and unconditional graphs are indicated by black lines; regression of $\mathbf{Y}$ on $\mathbf{X}$ is indicated by directed gray lines with arrows. (The online version of this figure is in color.)

10.3 Exploring Different Reproducing Kernels

In this section, we explore three types of kernels for RKHS

$$\begin{aligned}
\text{polynomial (PN):}\ & \kappa_X(\mathbf{a},\mathbf{b}) = (\mathbf{a}^\mathsf{T}\mathbf{b}+1)^r,\\
\text{Gauss radial basis (RB):}\ & \kappa_X(\mathbf{a},\mathbf{b}) = \exp(-\gamma\|\mathbf{a}-\mathbf{b}\|^2),\\
\text{rational quadratic (RQ):}\ & \kappa_X(\mathbf{a},\mathbf{b}) = 1-\|\mathbf{a}-\mathbf{b}\|^2/\\
& \qquad \times(\|\mathbf{a}-\mathbf{b}\|^2+c), \quad (44)
\end{aligned}$$

and investigate their performances as initial estimates for lasso and adaptive lasso. These kernels are widely used for RKHS (Genton 2001). For the CGGM, we use a nonlinear regression model with four combinations of dimensions: $q = 10, 20$, $p = 50, 100$. The nonlinear regression model is specified by

$$Y^i = \begin{cases} \left(\boldsymbol{\beta}_1^\mathsf{T}\mathbf{X}+1\right)^2+\varepsilon^i, & i = 1,3,6,8,\ldots,p-4,p-2,\\ \left(\boldsymbol{\beta}_2^\mathsf{T}\mathbf{X}\right)^2+\varepsilon^i, & i = 2,4,5,7,9,10,\ldots,\\ & \quad p-3,p-1,p, \end{cases} \quad (45)$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are q-dimensional vectors

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (\underbrace{1,\ldots,1}_{q/2},\underbrace{0,\ldots,0}_{q/2})^\mathsf{T},\\
\boldsymbol{\beta}_2 &= (\underbrace{0,\ldots,0}_{q/2},\underbrace{1,-1,\ldots,(-1)^{q/2+1}}_{q/2})^\mathsf{T}.
\end{aligned}$$

The distribution of $(\varepsilon^1,\ldots,\varepsilon^p)^\mathsf{T}$ is multivariate normal with mean 0 and precision matrix

$$\begin{pmatrix} \boldsymbol{\Gamma} & & 0\\ & \ddots & \\ 0 & & \boldsymbol{\Gamma} \end{pmatrix}$$

where each $\boldsymbol{\Gamma}$ is the precision matrix in Example 3.

The following specifications apply throughout the rest of Section 10: $\gamma = 1/(9q)$ for RB, $c = 200$ for RQ, and $r = 2$ for PN (because the predictors in Equation (45) are quadratic polynomials); the RKHS estimator $\hat{\Sigma}_{\text{PC}}(\epsilon_n)$ is used as the initial estimator for lasso and adaptive lasso, where ϵ_n are chosen so that the first 70 eigenvectors of $\mathbf{Q}\mathbf{K}_x\mathbf{Q}$ are retained. Ideally, the kernel parameters γ, c, and ϵ_n should be chosen by data-driven methods such as cross-validation. However, this is beyond the scope of the present article and will be further developed in a future study. Our choices are based on trial and error in pilot runs. Our experience indicates that the sparse estimators perform well and are reasonably stable when ϵ_n is chosen so that $10\% \sim 30\%$ of the eigenvectors of $\mathbf{Q}\mathbf{K}_x\mathbf{Q}$ are included. The sample size is $n = 100$ and the simulation is based on 200 samples. To save computing time we use the BIC to optimize λ_n for the first sample and use it for the rest 199 samples.

In Table 3, we compare the sparse estimators lasso and adaptive lasso, whose initial estimates are derived from kernels in Equation (44), with the full and constrained MLEs. The full MLE is computed using the knowledge that the predictor is a quadratic polynomial of $\mathbf{x}_1,\ldots,\mathbf{x}_p$. The constrained MLE uses, in addition, the knowledge of the conditional graph $\mathcal{G}$; that is, the positions of the zero entries of the true conditional precision matrix.

Table 3 shows that in all cases the sparse estimators perform substantially better than the full MLEs. For $q = 10$, the adaptive lasso estimates based on all three kernels also perform better than both the full and constrained MLEs; whereas the accuracy of most of the lasso estimates are between the full and the constrained MLEs. For $q = 20$, all sparse estimators perform substantially better than both the full and the constrained MLEs. From these results, we can see the effects of two types of regularization: the sparse regularization of the conditional precision matrix and the kernel-PCA regularization for the predictor. The first regularization counteracts the increase in the number of parameters in the conditional precision matrix as p increases, and the second counteracts the increase in the number

Table 3. Exploration of different reproducing kernels. Entries are $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2$ averaged over 200 simulation samples

		LASSO			ALASSO			MLE	
p	q	PN	RB	RQ	PN	RB	RQ	Full	Constrained
50	10	1.84	6.88	8.14	1.84	2.31	2.71	29.41	3.82
	20	11.03	11.03	11.03	17.15	17.14	17.14	299.28	94.18
100	10	5.96	20.07	24.06	3.35	4.33	5.17	171.32	7.73
	20	20.06	20.05	20.05	29.10	28.30	28.31	1787.32	186.39

of terms in a quadratic polynomial as q increases, both resulting in substantially reduced estimation error.

10.4 Comparisons With Two Naive Estimators

We now compare our sparse RKHS estimators for the CGGM with two simple methods: the linear regression and the simple thresholding.

10.4.1 Naive Linear Regression. A simple estimate of CGGM is to first apply multivariate linear regression of $\mathbf{Y}$ versus $\mathbf{X}$, regardless of the true regression relation, and then apply a sparse penalty to the residual variance matrix. To make a fair comparison with linear regression, we consider the following class of regression models:

$$\mathbf{Y} = (1-a)(\boldsymbol{\beta}^\mathsf{T}\mathbf{X})^2/4 + a(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}) + \boldsymbol{\varepsilon}, \quad 0 \le a \le 1. \quad (46)$$

This is a convex combination of a linear model and a quadratic model: it is linear when $a = 1$, quadratic when $a = 0$, and a mixture of both when $0 < a < 1$. The distribution of $\boldsymbol{\varepsilon}$ is as specified in Section 10.3. The fraction $1/4$ in the quadratic term in Equation (46) is introduced so that the linear and quadratic terms have the similar signal-to-noise ratios. In Table 4, we compare the estimation error of the CGGM based on Equation (46) by sparse linear regression and by sparse RKHS estimator using the adaptive lasso as penalty. We take $q = 10$, $p = 50$, and $n = 500$. The Gauss radial basis is used for the kernel method, with tuning parameters γ and ϵ_n being the same as specified in Section 10.3.

We see that the sparse RKHS estimate performs substantially better in all cases except $a = 1$, where Equation (46) is exactly a linear model. This suggests that linear-regression sparse estimate of CGGM is rather sensitive to nonlinearity: a slight proportion of nonlinearity in the mixture would make the kernel method favorable. In comparison, the sparse RKHS estimate is stable and accurate for different types of regression relations.

10.4.2 Simple Thresholding. A naive approach to sparsity is by dropping small entries of a matrix. For example, we can estimate $\boldsymbol{\Theta}$ by setting to zero the entries of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}^{-1}$ whose absolute values are smaller than some $\tau > 0$. Let $\tilde{\boldsymbol{\Theta}}(\tau)$ denote this estimator. Using the example in Section 10.3 with $p = 50, q = 10, n = 500$, we now compare the sparse RKHS estimators with $\tilde{\boldsymbol{\Theta}}(\tau)$. The curve in Figure 3 is the error $\|\tilde{\boldsymbol{\Theta}}(\tau) - \boldsymbol{\Theta}\|_F^2$ versus $\tau \in [0, 1]$. Each point in the curve is the error over 200 simulated samples. The estimate $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}$ is based on the gauss radial basis, whose tuning parameters γ and ϵ_n are as given in Section 10.3. The two horizontal lines in Figure 3 represent the errors for the lasso and the adaptive lasso estimators based on $\hat{\boldsymbol{\Sigma}}_{\mathrm{PC}}$, which are read off from Table 3.

Table 4. Comparison with linear regression. Entries are $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2$ averaged over 200 samples

a	0	0.2	0.4	0.6	0.8	1
Linear	10.49	10.48	10.48	10.42	10.10	0.75
Kernel	1.56	2.13	1.77	1.72	1.67	1.56

The figure shows that the simple thresholding estimate, even for the best threshold, does not perform as well as either of our sparse estimates. Note that in practice $\boldsymbol{\Theta}_0$ is unknown, and the optimal τ cannot be obtained by minimizing the curve in Figure 3. With this in mind, we expect the actual gap between the thresholding estimate and the sparse estimates to be even greater than that shown in the figure.

11. NETWORK INFERENCE FROM EQTL DATA

In this section, we apply our CGGM sparse estimators to two datasets to infer gene networks from expression quantitative trait loci (eQTL) data. The dataset is collected from an F2 intercross between inbred lines C3H/HeJ and C57BL/6J (Ghazalpour et al. 2006). It contains 3,421 transcripts and 1,065 markers from the liver tissues of 135 female mice

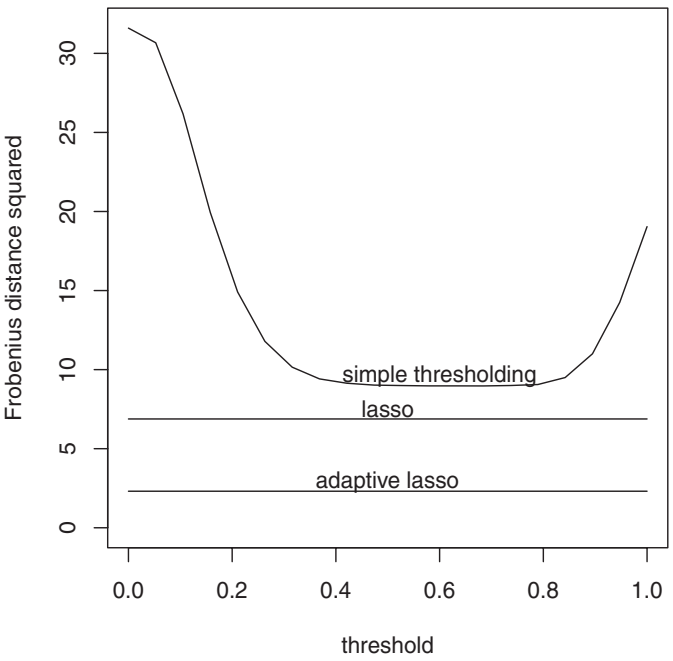

Figure 3. Comparison of lasso, adaptive lasso, and simple thresholding.

($n = 135$). The purpose of our analysis is to identify direct gene interactions by fitting the CGGM to the eQTL dataset. Although a gene network can be inferred from expression data alone, such a network would contain edges due to confounders such as shared genetic causal variants. The available marker data in the eQTL dataset allow us to isolate the confounded edges by conditioning on the genomic information.

We restrict our attention to subsets of genes, partly to accommodate the small sample size. In the eQTL analysis tradition, subsets of genes can be identified by two methods: co mapping and co expression. For co-mapping, each gene expression trait is mapped to markers, and the transcripts that are mapped to the same locus are grouped together. For co-expression, the highly correlated gene expressions are grouped together.

We first consider the subset of genes identified by co-mapping. It has been reported (Neto, Keller, Attie, and Yandell 2010) that 14 transcripts are mapped to a marker on chromosome 2 (at 55.95 cM). As this locus is linked to many transcripts, it is called a hot-spot locus. It is evident that this marker should be included as a covariate for CGGM. In addition, we include a marker on chromosome 15 (at 76.65 cM) as a covariate, because it is significantly linked to gene *Apbb1ip* (permutation p-value < 0.005), conditioning on the effect of the marker on chromosome 2. The transcript mapping is performed using the `qtl` package in R (Broman, Wu, Sen, and Churchill 2003).

The GGM detects 52 edges; the CGGM detects 34 edges, all in the set detected by the GGM. The two graphs are presented in the upper panel of Figure 4, where edges detected by both GGM and CGGM are represented by black solid lines, and edges detected by GGM alone are represented by blue dotted lines. In the left panel of Table 5, we compare the connectivity of genes of each graph. Among the 14 transcripts, *Pscdbp* contains the hot spot locus within the ± 100 kb boundaries of its location. Interestingly, in CGGM, this cis transcript has the lowest connectivity among the 14 genes, but one of the genes with which it is associated (*Apbb1ip*) is a hub gene, connected with seven other genes. In sharp contrast, GGM shows high connectivity of *Pscdbp* itself.

We next study the subset identified by co-expression. We use a hierarchical clustering approach in conjunction with the average agglomeration procedure to partition the transcripts into 10 groups. The relevant dissimilarity measure is $1 - |\rho_{ij}|$, where ρ_{ij} is the Pearson correlation between transcripts i and j. Among the 10 groups, we choose a group that contains 15 transcripts with the mean absolute correlation equal to 0.78. Thirteen of the 15 transcripts have annotations, and they are used in our analysis. Using the qtl package in R, each transcript is mapped to markers, and seven markers on chromosome 2 are significantly linked to transcripts (permutation p-value 0.005). Among those, we drop two markers that are identical to the adjacent markers. We thus use five markers (Chr2@100.18, Chr2@112.75, Chr2@115.95, Chr2@120.72, Chr2@124.12) as covariates.

The GGM identifies 52 edges and the CGGM identifies 30 edges. Among these edges, 28 are shared by both methods. The two graphs are presented in the lower panel of Figure 4, where edges detected by both GGM and CGGM are represented by black solid lines, edges detected by GGM alone are represented by blue dotted lines, and edges detected by CGGM alone are represented by red broken lines. Among the 13 transcripts, *Dtwd1* is the closest to all markers (distances < 300 kbp). The

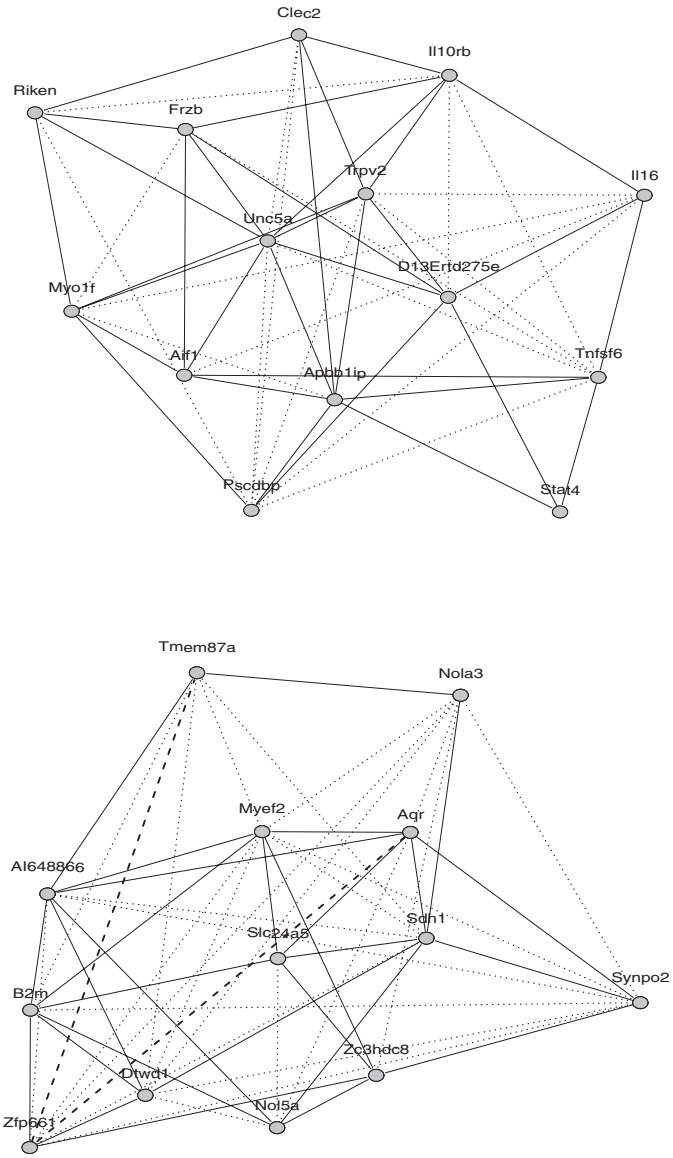

Figure 4. Gene networks based on GGM and CGGM. Upper panel: data based on co mapping selection. Lower panel: data based on coexpression selection. An edge from both GGM and CGGM is represented by a solid line; an edge from the GGM alone is represented by a dotted line; an edge from CGGM alone is represented by a broken line. (The online version of this figure is in color.)

connectivity of each method is shown in the right panel of Table 5. The number of genes that are connected to *Dtwd1* is reduced by 5 by CGGM. Unlike in the co-mapping network, in the co-expression network, CGGM detects two edges (Aqr–Zfp661, Zfp661–Tmem87a) that are not detected by GGM. These additional edges could be caused by the error in regression estimation. For example, if the regression coefficients for some of markers are 0, but are estimated to be nonzero, then spurious correlations arise among residuals. This indicates that the accurate identification of the correct covariates is more important in the co-expression network.

As a summary, in the network inference from the transcripts identified by co-mapping, we see that after conditioning on the markers, the cis-regulated transcripts (*Pscdbp*) are connected to relatively few genes of high connectivity. However, without conditioning on the markers, they appear to have high

Table 5. Connectivity of genes in the conditional and unconditional graphical models. The cis-regulated transcripts are indicated with boldface

Co-mapping				Co-expression			
Gene	GGM	CGGM	Diff.	Gene	GGM	CGGM	Diff.
Il16	7	3	4	Aqr	7	6	1
Frzb	7	5	2	Nola3	8	2	6
Apbb1ip	8	7	1	AI648866	9	6	3
Clec2	6	4	2	Zfp661	8	5	3
Riken	6	4	2	Myef2	11	5	6
Il10rb	8	5	3	Synpo2	9	3	6
Myo1f	8	5	3	Slc24a5	6	5	1
Aif1	6	5	1	Nol5a	7	4	3
Unc5a	11	8	3	Sdh1	10	6	4
Trpv2	9	6	3	Dtwd1	9	4	5
Tnfsf6	9	4	5	B2m	8	6	2
Stat4	3	3	0	Tmem87a	6	3	3
Pscdbp	9	3	6	Zc3hdc8	6	5	1
D13Ertd275e	7	6	1				

connectivity themselves. In other words, without conditioning on markers, these cis-regulated transcripts might be misinterpreted as hub-genes themselves. In the network inference from the transcripts identified by co expression, we see that after conditioning on the markers, a few edges are additionally detected, which may result from inaccurate identification of covariates for an individual transcript. Thus, including the correct set of markers for an individual transcript can be important.

ACKNOWLEDGMENTS

We would like to thank three referees and an Associate Editor for their many excellent suggestions, which lead to substantial improvement on an earlier draft. Especially, the asymptotic development in Section 8 is inspired by two reviewers' comments.

[Received October 2010. Revised July 2011.]

REFERENCES

Aronszajn, N. (1950), "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, 68, 337–404. [153]

Bickel, P. J., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [152]

Bickel, P. J., Ritov, Y., Klaassenn, C. A. J., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: The Johns Hopkins University Press. [156]

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384. [155]

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003), "R/qtl: QTLMapping in Experimental Crosses," *Bioinformatics,* 19, 889–890. [166]

Dempster, A. P. (1972), "Covariance Selection," *Biometrika*, 32, 95–108. [152]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [157]

Fernholz, L. T. (1983), "Von Mises Calculus for Statistical Functionals," *Lecture Notes in Statistics*, 19, New York: Springer. [156]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [155,162]

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009), "Kernel Dimension Reduction in Regression," *The Annals of Statistics*, 4, 1871–1905. [153,154]

Genton, M. G. (2001), "Classes of Kernels for Machine Learning: A Statistics Perspective," *Journal of Machine Learning Research*, 2, 299–312. [164]

Geyer, C. J. (1994), "On the Asymptotics of Constrained M-estimation," *The Annals of Statistics*, 22, 1998–2010. [158,159]

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Shadt, E. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2006), "Integrating Genetic and Network Analysis to Characterize Gene Related to Mouse Weight," *PLoS Genetics*, 2, e130. [165]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2009), "Joint Estimation of Multiple Graphical Models," unpublished manuscript available at *http://www.stat. lsa.umich.edu/gmichail/manuscript-jasa-09.pdf*. [152]

Henderson, H. V., and Searle, S. R. (1979), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *Canadian Journal of Statistics*, 7, 65–81. [155]

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*, New York: Cambridge University Press. [154]

Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378. [157]

Lafferty, J., McCallum, A., and Pereira, F. (2001), "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML 2001), 282–289. [152]

Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *The Annals of Statistics*, 37, 4254–4278. [152,157]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press. [152,155,157]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [152]

Mises, R. v. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *The Annals of Mathematical Statistics*, 18, 309–348. [155]

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010), "Causal Graphical Models in Systems Genetics: A Unified Framework for Joint Inference of Causal Network and Genetic Architecture for Correlated Phenotypes," *The Annals of Applied Statistics*, 4, 320–339. [166]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of American Statistical Association*, 104, 735–746. [152]

Reeds, J. A. (1976), "On the Definition of Von Mises Functionals," Ph.D. dissertation, Harvard University. [156]

Ren, J., and Sen, P. K. (1991), "On Hadamard Differentiability of Extended Statistical Functional," *Journal of Multivariate Analysis*, 39, 30–43. [156]

Rothman, A. J., Bickel, P., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [159,160,161]

Schwarz, G. E. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [161]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [152,155]

Vapnik, N. V. (1998), *Statistical Learning Theory*, New York: Wiley. [153]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [152,155,157,158,161]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [155]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [152,155]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509–1533. [155]