On Sufficient Graphical Models

Bing Li BXL9@PSU.EDU
Department of Statistics, Pennsylvania State University
326 Thomas Building, University Park, PA 16802

Kyongwon Kim KIMK@EWHA.AC.KR
Department of Statistics, Ewha Womans University
52 Ewhayeodae-gil, Seodaemun-gu, Seoul, Republic of Korea, 03760

Editor: Jin Tian

Abstract

We introduce a sufficient graphical model by applying the recently developed nonlinear sufficient dimension reduction techniques to the evaluation of conditional independence. The graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike a fully nonparametric graphical model, which relies on the high-dimensional kernel to characterize conditional independence, our graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and variable selection consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

Keywords: conjoined conditional covariance operator, generalized sliced inverse regression, nonlinear sufficient dimension reduction, reproducing kernel Hilbert space

1. Introduction

In this paper we propose a new nonparametric statistical graphical model, which we call the sufficient graphical model, by incorporating the recently developed nonlinear sufficient dimension reduction techniques to the construction of the distribution-free graphical models.

Let $\mathscr{G} = (\Gamma, \mathcal{E})$ be an undirected graph consisting of a finite set of nodes $\Gamma = \{1, \dots, p\}$ and set of edges $\mathcal{E} \subseteq \{(i,j) \in \Gamma \times \Gamma : i \neq j\}$. Since (i,j) and (j,i) represent the same edge in an undirected graph, we can assume without loss of generality that $i > j$. A statistical graphical model links $\mathscr{G}$ with a random vector $X = (X^1, \dots, X^p)$ by the conditional independence:

$$(i,j) \notin \mathcal{E} \Leftrightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)}, \qquad (1)$$

where $X^{-(i,j)} = \{X^1, \dots, X^p\} \setminus \{X^i, X^j\}$, and $A \perp\!\!\!\perp B | C$ means conditional independence. Thus, nodes i and j are connected if and only if X^i and X^j are dependent given $X^{-(i,j)}$. Our goal is to estimate the set $\mathcal{E}$ based on a sample $X_1, \dots, X_n$ of X. See Lauritzen (1996).

One of the most popular statistical graphical models is the Gaussian graphical model, which assumes that $X \sim N(\mu, \Sigma)$. Under the Gaussian assumption, conditional independence in (1) is encoded in the precision matrix $\Theta = \Sigma^{-1}$ in the following sense

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \Leftrightarrow \theta_{ij} = 0, \tag{2}$$

where θ_{ij} is the (i, j)th entry of the precision matrix Θ. By this equivalence, estimating $\mathcal{E}$ amounts to identifying the positions of the zero entries of the precision matrix, which can be achieved by sparse estimation methods such as the Tibshirani (1996), Fan and Li (2001), and Zou (2006). A variety of methods have been developed for estimating the Gaussian graphical model, which include, for example, Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Bickel and Levina (2008), and Peng et al. (2009). See also Friedman et al. (2008), Guo et al. (2010), and Lam and Fan (2009).

Since the Gaussian distribution assumption is restrictive, many recent advances have focused on relaxing this assumption. A main challenge in doing so is to avoid the curse of dimensionality (Bellman, 1961): a straightforward nonparametric extension would resort to a high-dimensional kernel, which are known to be ineffective. One way to relax the Gaussian assumption without evoking a high dimensional kernel is to use the copula Gaussian distribution, which is the approach taken by Liu et al. (2009), Liu et al. (2012a), and Xue and Zou (2012), and is further extended to the transelliptical model by Liu et al. (2012b).

However, the copula Gaussian assumption could still be restrictive: for example, if A and B are random variables satisfying $B = A^2 + \epsilon$, where A and ϵ are i.i.d. $N(0, 1)$, then (A, B) does not satisfy the copula Gaussian assumption. To further relax the distributional assumption, Li et al. (2014) proposed a new statistical relation called *the additive conditional independence* as an alternative criterion for constructing the graphical model. This relation has the advantage of achieving nonparametric model flexibility without using a high-dimensional kernel, while obeying the same set of semi-graphoid axioms that govern the conditional independence (Dawid, 1979; Pearl and Verma, 1987). See also Lee et al. (2016b) and Li and Solea (2018a). Other approaches to nonparametric graphical models include Fellinghauer et al. (2013) and Voorman et al. (2013).

In this paper, instead of relying on additivity to avoid the curse of dimensionality, we apply the recently developed nonparametric sufficient dimension reduction (Lee et al., 2013; Li, 2018b) to achieve this goal. The estimation proceeds in two steps: first, we use nonlinear sufficient dimension reduction to reduce $X^{-(i,j)}$ to a low-dimensional random vector U^{ij}; second, we use the kernel method to construct a nonparametric graphical model based on (X^i, X^j) and the dimension-reduced random vectors U^{ij}. The main differences between this approach and Li et al. (2014) are, first, we are able to retain conditional independence as the criterion for constructing the network, which is a widely accepted criterion with a more direct interpretation, and second, we are no longer restricted by the additive structure in the graphical model. Another attractive feature of our method is due to the "kernel trick", which means its computational complexity depends on the sample size rather than the size of the networks.

The rest of the paper is organized as follows. In Sections 2 and 3, we introduce the sufficient graphical model and describe its estimation method at the population level. In Section 4 we lay out the detailed algorithms to implement the method. In Section 5 we develop the asymptotic properties such as estimation consistency, variable selection consistency, and convergence rates. In Section 6, we conduct simulation studies to compare of our method with the existing methods. In Section 7, we apply our method to the DREAM 4 Challenge gene network data set. Section 8 concludes the paper with some further discussions. We put all proofs and some additional results in the Appendix.

2. Sufficient graphical model

In classical sufficient dimension reduction, we seek the lowest dimensional subspace $\mathcal{S}$ of $\mathbb{R}^p$, such that, after projecting $X \in \mathbb{R}^p$ on to $\mathcal{S}$, the information about the response Y is preserved; that is, $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $P_{\mathcal{S}}$ is the projection onto $\mathcal{S}$. This subspace is called the central subspace, written as $\mathcal{S}_{Y|X}$. See, for example, Li (1991), Cook (1994), and Li (2018b). Li et al. (2011) and Lee et al. (2013) extended this framework to the nonlinear setting by considering the more general problem: $Y \perp\!\!\!\perp X | \mathcal{G}$, where $\mathcal{G}$ a sub-σ field of the σ-field generated by X. The class of functions in a Hilbert space that are measurable with respect to $\mathcal{G}$ is called the central class, written as $\mathfrak{S}_{Y|X}$. Li et al. (2011) introduced the Principal Support Vector Machine, and Lee et al. (2013) generalized the Sliced Inverse Regression (Li, 1991) and the Sliced Average Variance Estimate (Cook and Weisberg, 1991) to estimate the central class. Precursors of this theory include Bach and Jordan (2002), Wu (2008), and Wang (2008).

To link this up with the statistical graphical model, let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\Omega_X, \mathcal{F}_X)$ a Borel measurable space with $\Omega_X \subseteq \mathbb{R}^p$, and $X : \Omega \to \Omega_X$ a random vector with distribution P_X. The ith component of X is denoted by X^i and its range denoted by Ω_{X^i}. We assume $\Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^p}$. Let $X^{(i,j)} = (X^i, X^j)$ and $X^{-(i,j)}$ be as defined in the Introduction. Let $\sigma(X^{-(i,j)})$ be the σ-field generated by $X^{-(i,j)}$. We assume, for each $(i,j) \in \Gamma \times \Gamma$, there is a proper sub σ-field $\mathcal{G}^{-(i,j)}$ of $\sigma(X^{-(i,j)})$ such that

$$X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}. \tag{3}$$

Without loss of generality, we assume $\mathcal{G}^{-(i,j)}$ is the smallest sub σ-field of $\sigma(X^{-(i,j)})$ that satisfies the above relation; that is, $\mathcal{G}^{-(i,j)}$ is the central σ-field for $X^{(i,j)}$ versus $X^{-(i,j)}$. There are plenty examples of joint distributions of X for which the condition (3) holds for every pair (i,j): see Appendix J. Using the properties of conditional independence developed in Dawid (1979) (with a detailed proof given in Li (2018b)), we can show that (3) implies the following equivalence.

Theorem 1 *If $X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$, then*

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \;\Leftrightarrow\; X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}.$$

This equivalence motivates us to use $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$ as the criterion to construct the graph $\mathcal{G}$ after performing nonlinear sufficient dimension reduction of $X^{(i,j)}$ versus $X^{-(i,j)}$ for each $(i,j) \in \Gamma \times \Gamma$, $i > j$.

Definition 2 *Under condition (3), the graph defined by*

$$(i,j) \notin \mathcal{E} \Leftrightarrow X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$$

is called the sufficient graphical model.

3. Estimation: population-level development

The estimation of the sufficient graphical model involves two steps: the first step is to use nonlinear sufficient dimension reduction to estimate $\mathcal{G}^{-(i,j)}$; the second is to construct a graph $\mathscr{G}$ based on reduced data

$$\{(X^{(i,j)}, \mathcal{G}^{-(i,j)}) : (i,j) \in \Gamma \times \Gamma, i > j\}.$$

In this section we describe the two steps at the population level. To do so, we need some preliminary concepts such as the covariance operator between two reproducing kernel Hilbert spaces, the mean element in an reproducing kernel Hilbert spaces, the inverse of an operator, as well as the centered reproducing kernel Hilbert spaces. These concepts are defined in the Appendix A.2. A fuller development of the related theory can be found in Li (2018b). The symbols $\mathrm{ran}(\cdot)$ and $\overline{\mathrm{ran}}(\cdot)$ will be used to denote the range and the closure of the range of a linear operator.

3.1 Step 1: Nonlinear dimension reduction

We use the generalized sliced inverse regression Lee et al. (2013), (Li, 2018b) to perform the nonlinear dimension reduction. For each pair $(i,j) \in \Gamma \times \Gamma$, $i > j$, let $\Omega_{X^{-(i,j)}}$ be the range of $X^{-(i,j)}$, which is the Cartesian product of $\Omega_{X^1}, \dots, \Omega_{X^p}$ with Ω_{X^i} and Ω_{X^j} removed. Let

$$\kappa_X^{-(i,j)} : \Omega_{X^{-(i,j)}} \times \Omega_{X^{-(i,j)}} \to \mathbb{R}$$

be a positive semidefinite kernel. Let $\mathscr{H}_X^{-(i,j)}$ be the centered reproducing kernel Hilbert space generated by $\kappa_X^{-(i,j)}$. Let $\Omega_{X^{(i,j)}}$, $\kappa_X^{(i,j)}$, and $\mathscr{H}_X^{(i,j)}$ be the similar objects defined for $X^{(i,j)}$.

Assumption 1

$$E[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})] < \infty, \quad E[\kappa_X^{(i,j)}(X^{(i,j)}, X^{(i,j)})] < \infty.$$

This is a very mild assumption that is satisfied by most kernels. Under this assumption, the following covariance operators are well defined:

$$\Sigma_{X^{-(i,j)}X^{(i,j)}} : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{-(i,j)}, \quad \Sigma_{X^{-(i,j)}X^{-(i,j)}} : \mathscr{H}_X^{-(i,j)} \to \mathscr{H}_X^{-(i,j)}.$$

For the formal definition of the covariance operator, see SA.2. Next, we introduce the regression operator from $\mathscr{H}_X^{(i,j)}$ to $\mathscr{H}_X^{-(i,j)}$. For this purpose we need to make the following assumption.

Assumption 2 $\mathrm{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) \subseteq \mathrm{ran}(\Sigma_{X^{-(i,j)}X^{-(i,j)}})$.

As argued in Li (2018b), this assumption can be interpreted as a type of collective smoothness in the relation between $X^{(i,j)}$ and $X^{-(i,j)}$: intuitively, it requires the operator $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ sends all the input functions to the low-frequency domain of the operator $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$. Under Assumption 2, the linear operator

$$R_{X^{-(i,j)}X^{(i,j)}} = \Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}} \Sigma_{X^{-(i,j)}X^{(i,j)}}$$

is defined, and we call it the regression operator from $\mathscr{H}_X^{(i,j)}$ to $\mathscr{H}_X^{-(i,j)}$. The meaning of the inverse $\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}$ is defined in Appenix A.2. The regression operator in this form was formally defined in Lee et al. (2016a), but earlier forms existed in Fukumizu et al. (2004); see also Li (2018a).

Assumption 3 *$R_{X^{-(i,j)}X^{(i,j)}}$ is a finite-rank operator, with rank d_{ij}.*

Intuitively, this assumption means that $R_{X^{-(i,j)}X^{(i,j)}}$ filters out the high frequency functions of $X^{(i,j)}$, so that, for any $f \in \mathscr{H}^{(i,j)}$, $R_{X^{-(i,j)}X^{(i,j)}} f$ is relatively smooth. It will be violated, for example, if one can find an $f \in \mathscr{H}^{(i,j)}$ that makes $R_{X^{-(i,j)}X^{(i,j)}} f$ arbitrarily choppy. The regression operator plays a crucial role in nonlinear sufficient dimension reduction. Let $L_2(P_{X^{-(i,j)}})$ be the L_2-space with respect to the distribution $P_{X^{-(i,j)}}$ of $X^{-(i,j)}$. As shown in Lee et al. (2013), the closure of the range of the regression operator is equal to the central subspace; that is,

$$\overline{\mathrm{ran}}(R_{X^{-(i,j)}X^{(i,j)}}) = \mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}} \tag{4}$$

under the following assumption.

Assumption 4

1. *$\mathscr{H}_X^{-(i,j)}$ is dense in $L_2(P_{X^{-(i,j)}})$ modulo constants; that is, for any $f \in L_2(P_{X^{-(i,j)}})$ and any $\epsilon > 0$, there is a $g \in \mathscr{H}_X^{-(i,j)}$ such that* $\mathrm{var}[f(X^{-(i,j)}) - g(X^{-(i,j)})] < \epsilon$*;*
2. *$\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$ is a sufficient and complete.*

The first condition essentially requires the kernel $\kappa_X^{-(i,j)}$ to be a universal kernel with respect to the $L_2(P_{X^{-(i,j)}})$-norm. It means $\mathscr{H}^{-(i,j)}$ is rich enough to approximate any $L_2(P_{X^{-(i,j)}})$-function arbitrarily closely. For example, it is satisfied by the Gaussian radial basis function kernel, but not by the polynomial kernel. For more information on universal kernels, see Sriperumbudur, Fukumizu, and Lanckriet (2011). The completeness in the second condition means

$$E[g(X^{-(i,j)})|X^{(i,j)}] = 0 \text{ almost surely } \Rightarrow \ g(X^{-(i,j)}) = 0 \text{ almost surely.}$$

This concept is defined in Lee, Li, and Chiaromonte (2013), and is similar to the classical definition of completeness treating $X^{-(i,j)}$ as the parameter. Lee, Li, and Chiaromonte (2013) showed that completeness is a mild condition, and is satisfied by most nonparametric models.

A basis of the central class $\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$ can be found by solving the generalized eigenvalue problem: for $k = 1, \ldots, d_{ij}$,

$$\begin{aligned} &\text{maximize} \quad \langle f, \Sigma_{X^{-(i,j)}X^{(i,j)}} A \Sigma_{X^{(i,j)}X^{-(i,j)}} f \rangle_{-(i,j)} \\ &\text{subject to} \quad \begin{cases} \langle f_k, \Sigma_{X^{-(i,j)}X^{-(i,j)}} f_k \rangle_{-(i,j)} = 1 \\ \langle f_k, \Sigma_{X^{-(i,j)}X^{-(i,j)}} f_\ell \rangle_{-(i,j)} = 0, \text{ for } \ell = 1, \ldots, k-1 \end{cases} \end{aligned} \tag{5}$$

where $A : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{(i,j)}$ is any nonsingular and self adjoint operator, and $\langle \cdot, \cdot \rangle_{-(i,j)}$ is the inner product in $\mathscr{H}_X^{-(i,j)}$. That is, if $f_1^{ij}, \ldots f_{d_{ij}}^{ij}$ are the first d_{ij} eigenfunctions of this eigenvalue problem, then they span the central class. This type of estimate of the central class is called generalized sliced inverse regression. Convenient choices of A are the identity mapping I or the operator $\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}$. If we use the latter, then we need the following assumption.

Assumption 5 $\mathrm{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) \subseteq \mathrm{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}})$.

This assumption has the similar interpretation as Assumption 2; see Appendix K. At the population level, choosing A to be $\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}$ achieves better scaling because it down weights those components of the output of $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ with larger variances. However, if the sample size is not sufficiently large, involving an estimate of $\Sigma^{-1}_{X^{-(i,j)}X^{(i,j)}}$ in the procedure could incur extra variations that overwhelm the benefit brought by $\Sigma^{-1}_{X^{-(i,j)}X^{(i,j)}}$. In this case, a nonrandom operator such as $A = I$ is preferable. In this paper we use $A = \Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}$. Let U^{ij} denote the random vector $(f_1^{ij}(X^{-(i,j)}), \ldots f_{d_{ij}}^{ij}(X^{-(i,j)}))$. The set of random vectors $\{U^{ij} : (i,j) \in \Gamma \times \Gamma, i > j\}$ is the output for the nonlinear sufficient dimension reduction step.

3.2 Step 2:Estimation of sufficient graphical model

To estimate the edge set of the sufficient graphical model we need to find a way to determine whether $X^i \perp\!\!\!\perp X^j | U^{ij}$ is true. We use a linear operator introduced by Fukumizu et al. (2008) to perform this task, which is briefly described as follows. Let U, V, W be random vectors taking values in measurable spaces $(\Omega_U, \mathcal{F}_U)$, $(\Omega_V, \mathcal{F}_V)$, and $(\Omega_W, \mathcal{F}_W)$. Let $\Omega_{UW} = \Omega_U \times \Omega_W$, $\Omega_{VW} = \Omega_V \times \Omega_W$, $\mathcal{F}_{UW} = \mathcal{F}_U \times \mathcal{F}_V$, and $\mathcal{F}_{VW} = \mathcal{F}_V \times \mathcal{F}_W$. Let

$$\kappa_{UW} : \Omega_{UW} \times \Omega_{UW} \to \mathbb{R}, \quad \kappa_{VW} : \Omega_{VW} \times \Omega_{VW} \to \mathbb{R}, \quad \kappa_W : \Omega_W \times \Omega_W \to \mathbb{R}$$

be positive kernels. For example, for $(u_1, w_1), (u_2, w_2) \in \Omega_{UW} \times \Omega_{UW}$, κ_{UW} returns a real number denoted by $\kappa_{UW}[(u_1, w_1), (u_2, w_2)]$. Let $\mathscr{H}_{UW}$, $\mathscr{H}_{VW}$, and $\mathscr{H}_W$ be the centered reproducing kernel Hilbert space's generated by the kernels κ_{UW}, κ_{VW}, and κ_W. Define the covariance operators

$$\begin{aligned} &\Sigma_{(UW)(VW)} : \mathscr{H}_{VW} \to \mathscr{H}_{UW}, \quad \Sigma_{(UW)W} : \mathscr{H}_W \to \mathscr{H}_{UW}, \\ &\Sigma_{(VW)W} : \mathscr{H}_W \to \mathscr{H}_{VW}, \quad \Sigma_{WW} : \mathscr{H}_W \to \mathscr{H}_W \end{aligned} \tag{6}$$

as before. The following definition is due to Fukumizu et al. (2008). Since it plays a special role in this paper, we give it a name – "conjoined conditional covariance operator" that figuratively depicts its form.

Definition 3 *Suppose*

1. *If S is W, or (U, W), or (V, W), then $E[\kappa_S(S, S)] < \infty$;*

2. $\mathrm{ran}(\Sigma_{W(VW)}) \subseteq \mathrm{ran}(\Sigma_{WW})$, $\mathrm{ran}(\Sigma_{W(UW)}) \subseteq \mathrm{ran}(\Sigma_{WW})$.

Then the operator $\Sigma_{\ddot{U}\ddot{V}|W} = \Sigma_{(UW)(VW)} - \Sigma_{(UW)W}\Sigma_{WW}^{-1}\Sigma_{W(VW)}$ is called the conjoined conditional covariance operator between U and V given W.

The word "conjoined" describes the peculiar way in which W appears in $\Sigma_{(UW)W}$ and $\Sigma_{W(VW)}$, which differs from an ordinary conditional covariance operator, where these operators are replaced by Σ_{UW} and Σ_{WV}. The following proposition is due to Fukumizu et al. (2008), a proof of a special case of which is given in Fukumizu et al. (2004).

Proposition 4 *Suppose*

1. *$\mathscr{H}_{UW} \otimes \mathscr{H}_{VW}$ is probability determining;*

2. *for each $f \in \mathscr{H}_{UW}$, the function $E[f(U, W)|W = \cdot]$ belongs to $\mathscr{H}_W$;*

3. *for each $g \in \mathscr{H}_{VW}$, the function $E[g(V, W)|W = \cdot]$ belongs to $\mathscr{H}_W$;*

Then $\Sigma_{\ddot{U}\ddot{V}|W} = 0$ if and only if $U \perp\!\!\!\perp V | W$.

The notion of probability determining in the context of reproducing kernel Hilbert space was defined in Fukumizu et al. (2004). For a generic random vector X, an reproducing kernel Hilbert space $\mathscr{H}_X$ based on a kernel κ_X is probability determining if and only if the mapping $P \mapsto E_P[\kappa_X(\cdot, X)]$ is injective. Intuitively, this requires the family of expectations $\{E_P f(X) : f \in \mathcal{H}_X\}$ to be rich enough to identify P. For example, the Gaussian radial basis function is probability determining, but a polynomial kernel is not. We apply the above proposition to X^i, X^j, U^{ij} for each $(i,j) \in \Gamma \times \Gamma$, $i > j$. Let

$$\kappa_{XU}^{i,ij} : (\Omega_{X^i} \times \Omega_{U^{ij}}) \times (\Omega_{X^i} \times \Omega_{U^{ij}}) \to \mathbb{R}$$

be a positive definite kernel, and $\mathscr{H}_{XU}^{i,ij}$ the centered reproducing kernel Hilbert space generated by $\kappa_{XU}^{i,ij}$. Similarly, let $\kappa_U^{ij} : \Omega_{U^{ij}} \times \Omega_{U^{ij}} \to \mathbb{R}$ be a positive kernel, and $\mathscr{H}_U^{ij}$ the centered reproducing kernel Hilbert space generated by κ_U^{ij}.

Assumption 6 *Conditions (1) and (2) of Definition 3 and conditions (1), (2), and (3) of Proposition 4 are satisfied with U, V, and W therein replaced by X^i, X^j, and U^{ij}, respectively, for each $(i,j) \in \Gamma \times \Gamma$ and $i > j$.*

Under this assumption, the conjoined conditional covariance operator $\Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}$ is well defined and has the following property.

Corollary 5 *Under Assumption 6, we have $(i,j) \notin \mathcal{E} \Leftrightarrow \Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}} = 0$.*

This corollary motivates us to estimate the graph by thresholding the norm of the estimated conjoined conditional covariance operator.

4. Estimation: sample-level implementation

4.1 Implementation of step 1

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). At the sample level, the centered reproducing kernel Hilbert space $\mathscr{H}_X^{-(i,j)}$ is spanned by the functions

$$\{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] : a = 1, \ldots, n\}, \tag{7}$$

where $\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})$ stands for the function $u \mapsto \kappa_X^{-(i,j)}(u, X^{-(i,j)})$, and $E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]$ the function $u \mapsto E_n[\kappa_X^{-(i,j)}(u, X^{-(i,j)})]$.

We estimate the covariance operators $\Sigma_{X^{-(i,j)} X^{(i,j)}}$ and $\Sigma_{X^{-(i,j)} X^{-(i,j)}}$ by

$$\begin{aligned}
\hat{\Sigma}_{X^{-(i,j)} X^{(i,j)}} =& E_n\{[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n \kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] \\
& \otimes [\kappa_X^{(i,j)}(\cdot, X^{(i,j)}) - E_n \kappa_X^{(i,j)}(\cdot, X^{(i,j)})]\} \\
\hat{\Sigma}_{X^{-(i,j)} X^{-(i,j)}} =& E_n\{[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n \kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] \\
& \otimes [\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n \kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\},
\end{aligned}$$

respectively. We estimate $\Sigma^{-1}_{X^{(i,j)} X^{(i,j)}}$ by the Tychonoff-regularized inverse $(\hat{\Sigma}_{X^{(i,j)} X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}$, where $I : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{(i,j)}$ is the identity operator. The regularized inverse is used to avoid over fitting. It plays the same role as ridge regression (Hoerl and Kennard, 1970) that alleviates over fitting by adding a multiple of the identity matrix to the sample covariance matrix before inverting it.

At the sample level, the generalized eigenvalue problem (5) takes the following form: at the kth iteration,

$$\begin{aligned} &\text{maximize} && \langle f, \hat{\Sigma}_{X^{-(i,j)} X^{(i,j)}} (\hat{\Sigma}_{X^{(i,j)} X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1} \hat{\Sigma}_{X^{(i,j)} X^{-(i,j)}} f \rangle_{-(i,j)} \\ &\text{subject to} && \begin{cases} \langle f, \hat{\Sigma}_{X^{-(i,j)} X^{-(i,j)}} f \rangle_{-(i,j)} = 1, \\ \langle f, \hat{\Sigma}_{X^{-(i,j)} X^{-(i,j)}} f_\ell \rangle_{-(i,j)} = 0, \quad \ell = 1, \ldots, k-1, \end{cases} \end{aligned} \tag{8}$$

where $f_1, \ldots, f_{k-1}$ are the maximizers in the previous steps. The first d_{ij} eigenfunctions are an estimate of a basis in the central class $\mathfrak{S}_{X^{(i,j)} | X^{-(i,j)}}$.

Let $K_{X^{-(i,j)}}$ be the $n \times n$ matrix whose (a, b)th entry is $\kappa_X^{-(i,j)}(X_a^{-(i,j)}, X_b^{-(i,j)})$, $Q = I_n - 1_n 1_n^{\mathsf{T}}/n$, and $G_{X^{-(i,j)}} = Q K_{X^{-(i,j)}} Q$. Let $a^1, \ldots, a^{d_{ij}}$ be the first d_{ij} eigenvectors of the matrix

$$(G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} G_{X^{-(i,j)}} G_{X^{(i,j)}} (G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1}. \tag{9}$$

In spite of its appearance, the above matrix is actually symmetric, because the matrices $G_{X^{(i,j)}}$ and $(G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1}$ commute. Let $b^r = (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} a^r$ for $r = 1, \ldots, d_{ij}$. As shown in Section SL.2, the eigenfunctions $f_1^{ij}, \ldots, f_{d_{ij}}^{ij}$ are calculated by

$$f_r^{ij} = \sum_{a=1}^{n} b_a^r \{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\}.$$

The statistics $\hat{U}_a^{ij} = (f_1^{ij}(X_a^{-(i,j)}), \ldots, f_{d_{ij}}^{ij}(X_a^{-(i,j)}))$, $a = 1, \ldots, n$, will be used as the input for the second step.

4.2 Implementation of step 2

This step consists of estimating the conjoined conditional covariance operator's for each (i, j) and thresholding their norms. At the sample level, the centered reproducing kernel Hilbert space's generated by the kernels $\kappa_{XU}^{i,ij}$, $\kappa_{XU}^{j,ij}$, and κ_U^{ij} are

$$\begin{aligned} \mathscr{H}_{XU}^{i,ij} &= \operatorname{span}\{\kappa_{XU}^{i,ij}(\cdot, (X_a^i, U_a^{ij})) - E_n[\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))] : a = 1, \ldots, n\}, \\ \mathscr{H}_{XU}^{j,ij} &= \operatorname{span}\{\kappa_{XU}^{j,ij}(\cdot, (X_a^j, U_a^{ij})) - E_n[\kappa_{XU}^{j,ij}(\cdot, (X^j, U^{ij}))] : a = 1, \ldots, n\}, \\ \mathscr{H}_U^{ij} &= \operatorname{span}\{\kappa_U^{ij}(\cdot, U_a^{ij}) - E_n[\kappa_U^{ij}(\cdot, U^{ij})] : a = 1, \ldots, n\}, \end{aligned}$$

where, for example, $\kappa_{XU}^{i,ij}(\cdot, (X_a^i, U_a^{ij}))$ denotes the function

$$\Omega_{X^i} \times \Omega_{U^{ij}} \to \mathbb{R}, \quad (x^i, u^{ij}) \mapsto \kappa_{XU}^{i,ij}((x^i, u^{ij}), (X_a^i, U_a^{ij}))$$

and $E_n[\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))]$ denotes the function

$$\Omega_{X^i} \times \Omega_{U^{ij}} \to \mathbb{R}, \quad (x^i, u^{ij}) \mapsto E_n[\kappa_{XU}^{i,ij}((x^i, u^{ij}), (X^i, U^{ij}))].$$

We estimate the covariance operators $\Sigma_{(X^i U^{ij})(X^i U^{ij})}$, $\Sigma_{(X^i U^{ij}) U^{ij}}$, $\Sigma_{X^j (X^j U^{ij})}$, and $\Sigma_{U^{ij} U^{ij}}$ by

$$\begin{aligned}
\hat{\Sigma}_{(X^i U^{ij})(X^j U^{ij})} &= E_n\{[\kappa_{XU}^{i,ij}(\cdot,(X^i,U^{ij})) - E_n \kappa_{XU}^{i,ij}(\cdot,(X^i,U^{ij}))] \\
&\qquad \otimes [\kappa_{XU}^{j,ij}(\cdot,(X^j,U^{ij})) - E_n \kappa_{XU}^{j,ij}(\cdot,(X^j,U^{ij}))]\} \\
\hat{\Sigma}_{(X^i U^{ij}) U^{ij}} &= E_n\{[\kappa_{XU}^{i,ij}(\cdot,(X^i,U^{ij})) - E_n \kappa_{XU}^{i,ij}(\cdot,(X^i,U^{ij}))] \\
&\qquad \otimes [\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n \kappa_{U}^{ij}(\cdot,U^{ij})]\} \\
\hat{\Sigma}_{U^{ij} (X^j U^{ij})} &= E_n\{[\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n \kappa_{U}^{ij}(\cdot,U^{ij})] \\
&\qquad \otimes [\kappa_{XU}^{j,ij}(\cdot,(X^j,U^{ij})) - E_n \kappa_{XU}^{j,ij}(\cdot,(X^j,U^{ij}))]\} \\
\hat{\Sigma}_{U^{ij} U^{ij}} &= E_n\{[\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n \kappa_{U}^{ij}(\cdot,U^{ij})] \\
&\qquad \otimes [\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n \kappa_{U}^{ij}(\cdot,U^{ij})]\},
\end{aligned} \tag{10}$$

respectively. We then estimate the conjoined conditional covariance operator by

$$\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}} = \hat{\Sigma}_{(X^i U^{ij})(X^j U^{ij})} - \hat{\Sigma}_{(X^i U^{ij}) U^{ij}} (\hat{\Sigma}_{U^{ij} U^{ij}} + \epsilon_U^{(i,j)} I)^{-1} \hat{\Sigma}_{U^{ij} (X^j U^{ij})},$$

where, again, we have used Tychonoff regularization to estimate the inverted covariance operator $\Sigma_{U^{ij} U^{ij}}$. Let $K_{U^{ij}}$, $K_{X^i U^{ij}}$, and $K_{X^j U^{ij}}$ be the Gram matrices

$$\begin{aligned}
K_{U^{ij}} &= \{\kappa_U^{ij}(U_a^{ij}, U_b^{ij})\}_{a,b=1}^n, \\
K_{X^i U^{ij}} &= \{\kappa_{XU}^{i,ij}((X_a^i, U_a^{ij}), (X_b^i, U_b^{ij}))\}_{a,b=1}^n, \\
K_{X^j U^{ij}} &= \{\kappa_{XU}^{j,ij}((X_a^j, U_a^{ij}), (X_b^j, U_b^{ij}))\}_{a,b=1}^n,
\end{aligned}$$

and $G_{X^i U^{ij}}$, $G_{X^j U^{ij}}$, and $G_{U^{ij}}$ their centered versions

$$G_{X^i U^{ij}} = Q K_{X^i U^{ij}} Q, \quad G_{X^j U^{ij}} = Q K_{X^j U^{ij}} Q, \quad G_{U^{ij}} = Q K_{U^{ij}} Q.$$

As shown in Appendix L,

$$\|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} = \left\| G_{X^i U^{ij}}^{1/2} G_{X^j U^{ij}}^{1/2} - G_{X^i U^{ij}}^{1/2} G_{U^{ij}} (G_{U^{ij}} + \epsilon_U^{(i,j)} Q)^\dagger G_{X^j U^{ij}}^{1/2} \right\|_{\mathrm{F}},$$

where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm. Estimation of the edge set is then based on thresholding this norm; that is,

$$\hat{\mathcal{E}} = \{(i,j) \in \Gamma \times \Gamma : \, i > j, \, \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} > \rho_n\}$$

for some chosen $\rho_n > 0$.

4.3 Tuning

We have three types of tuning constants: those for the kernels, those for Tychonoff regularization, and the threshold ρ_n. For the Tychonoff regularization, we have $\epsilon_X^{(i,j)}$ and $\epsilon_X^{-(i,j)}$ for step 1, and $\epsilon_U^{(i,j)}$ for step 2. In this paper we use the Gaussian radial basis function as the kernel:

$$\kappa(u,v) = \exp(-\gamma \|u - v\|^2). \tag{11}$$

For each (i,j), we have five γ's to determine: $\gamma_X^{(i,j)}$ for the kernel $\kappa_X^{(i,j)}$, $\gamma_X^{-(i,j)}$ for $\kappa_X^{-(i,j)}$, $\gamma_{XU}^{i,ij}$ for $\kappa_{XU}^{i,ij}$, $\gamma_{XU}^{j,ij}$ for $\kappa_{XU}^{j,ij}$, and γ_U^{ij} for κ_U^{ij}, which are chosen by the following formula (see, for example, Li (2018b))

$$1/\sqrt{\gamma} = \binom{n}{2}^{-1} \sum_{a<b} \|s_a - s_b\|, \tag{12}$$

where $s_1, \dots, s_n$ are the sample of random vectors corresponding to the mentioned five kernels. For example, for the kernel $\kappa_{XU}^{j,ij}$, $s_a = (X_a^j, U_a^{ij})$. For the tuning parameters in Tychonoff regularization, we use the following generalized cross validation scheme (GCV; see Golub et al. (1979)):

$$\mathrm{GCV}(\epsilon) = \mathrm{argmin}_\epsilon \sum_{i<j} \frac{\|G_1 - G_2^\intercal [G_2 + \epsilon\, \lambda_{\max}(G_2)]^{-1} G_1\|_{\mathrm{F}}}{\frac{1}{n}\mathrm{tr}\{I_n - G_2^\intercal [G_2 + \epsilon\, \lambda_{\max}(G_2)]^{-1}\}}, \tag{13}$$

where $G_1, G_2 \in \mathbb{R}^{n\times n}$ are positive semidefinite matrices, and $\lambda_{\max}(G_2)$ is the largest eigenvalue of G_2. The matrices G_1 and G_2 are the following matrices for the three tuning parameters:

1. $G_1 = G_{X^{-(i,j)}}$, $G_2 = G_{X^{(i,j)}}$ for $\epsilon_X^{(i,j)}$,

2. $G_1 = G_{X^{(i,j)}}$, $G_2 = G_{X^{-(i,j)}}$ for $\epsilon_X^{-(i,j)}$,

3. $G_1 = G_{X^{(i,j)}}$, $G_2 = G_{U^{ij}}$ for $\epsilon_U^{(i,j)}$,

We minimize (13) over a grid to choose ϵ, as detailed in Section 6.

We also use generalized cross validation to determine the thresholding parameter ρ_n. Let $\hat{\mathcal{E}}(\rho)$ be the estimated edge set using a threshold ρ, and, for each $i \in \Gamma$, let $C^i(\rho) = \{X^j : j \in \Gamma,\ (i,j) \in \hat{\mathcal{E}}(\rho)\}$ be the subset of components of X at the neighborhood of the node i in the graph $(\Gamma, \hat{E}(\rho))$. The basic idea is to apply the generalized cross validation to the regression of the feature of X^i on the feature of $C^i(\rho)$. The generalized cross validation for this regression takes the form

$$\mathrm{GCV}(\rho) = \sum_{i=1}^{p} \frac{\|G_{X^i} - G^\intercal_{C^i(\rho)}[G_{C^i(\rho)} + \epsilon\, \lambda_{\max}(G_{C^i(\rho)}) I_n]^{-1} G_{X^i}\|_{\mathrm{F}}}{\frac{1}{n}\mathrm{tr}\{I_n - G^\intercal_{C^i(\rho)}[G_{C^i(\rho)} + \epsilon\, \lambda_{\max}(G_{C^i(\rho)}) I_n]^{-1}\}}, \tag{14}$$

where $G_{C^i(\rho)} = Q K_{C^i(\rho)} Q$, and $K_{C^i(\rho)}$ is the $n \times n$ kernel matrix for the sample of $C^i(\rho)$. We minimize GCV(ρ) over the grid $\rho \in \{\ell \times 10^{-2} : \ell = 2, \dots, 7\}$ to determine the optimal threshold ρ_n.

Regarding the selection of the dimension of U^{ij}, to our knowledge there has been no systematic procedure available to determine the dimension of the central class for nonlinear sufficient dimension reduction. While some recently developed methods for order determination for linear sufficient dimension reduction, such as the ladle estimate and predictor augmentation estimator (Luo and Li, 2016, 2020), may be generalizable to the nonlinear sufficient dimension reduction setting, we will leave this topic to future research. Our experiences and intuitions indicate that a small dimension, such as 1 or 2, for the central class would be sufficient in most cases. For example, in the classical nonparametric regression problems $Y = f(X) + \epsilon$ with $X \perp\!\!\!\perp \epsilon$, the dimension of the central class is by definition equal to 1.

4.4 Algorithm

We next provide a detailed algorithm of our two-step procedure.

Algorithm Sufficient graphical model

```
for $i > j$ do
    For each $(i,j)$, standardize $X^{(i,j)}$ and $X^{-(i,j)}$ marginally.
    Choose $\gamma_X^{(i,j)}, \gamma_X^{-(i,j)}, \epsilon_X^{(i,j)}$ and $\epsilon_X^{-(i,j)}$ according to (12) and (13) in Section 4.3.
    Extract the first $d_{ij}$ eigenvectors from the matrix

        $$(G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} G_{X^{-(i,j)}} G_{X^{(i,j)}} (G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1}.$$

     Then, derive the sufficient predictor according to Section 4.1 and set this as $U^{ij}$.
    Choose $\gamma_{XU}^{i,ij}, \gamma_{XU}^{j,ij}, \gamma_U^{ij}, \epsilon_U^{(i,j)}$ according to Section 4.3.
    Calculate $\|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} = \left\| G^{1/2}_{X^i U^{ij}} G^{1/2}_{X^j U^{ij}} - G^{1/2}_{X^i U^{ij}} G_{U^{ij}} (G_{U^{ij}} + \epsilon_U^{(i,j)} Q)^\dagger G^{1/2}_{X^j U^{ij}} \right\|_{\mathrm{F}}$.
    Determine the threshold, $\rho_n$, by minimizing (14) over the grid $\rho \in \{\ell \times 10^{-2} : \ell = 2, \ldots, 7\}$.
    Estimate edges by

        $$\hat{\mathcal{E}} = \{(i,j) \in \Gamma \times \Gamma : \, i > j, \, \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} > \rho_n\}.$$

end for
```

5. Asymptotic theory

In this section we develop the consistency and convergence rates of our estimate and related operators. The challenge of this analysis is that our procedure involves two steps: we first extract the sufficient predictor using one set of kernels, and then substitute it into another set of kernels to get the final result. Thus we need to understand how the error propagates from the first step to the second. We also develop the asymptotic theory allowing p to go to infinity with n, which is presented in the Appendix.

5.1 Overview

Our goal is to derive the convergence rate of

$$\left| \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}}\|_{\mathrm{HS}} - \|\Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} \right|,$$

as $\|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}}\|_{\mathrm{HS}}$ is the quantity we threshold to determine the edge set. By the triangular inequality,

$$\begin{aligned} \left| \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}}\|_{\mathrm{HS}} - \|\Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} \right| &\le \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}} - \Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} \\ &\le \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} + \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}} - \Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}}. \end{aligned}$$

So we need to derive the convergence rates of the following quantities:

$$
\begin{aligned}
&\text{(i)} \quad \|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}, \\
&\text{(ii)} \quad \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}}, \\
&\text{(iii)} \quad \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}} - \Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}},
\end{aligned}
\tag{15}
$$

where, to avoid overly crowded subscripts, we have used $\mathscr{H}^{-(i,j)}(X)$ to denote $\mathscr{H}_X^{-(i,j)}$ when it occurs as a subscript.

The first and third convergence rates can be derived using the asymptotic tools for linear operators developed in Fukumizu et al. (2007), Li and Song (2017), Lee et al. (2016a), and Solea and Li (2022). Theorems 10 and 11 below are concerned with these rates. At the sample level, the linear operators involved in these rates are essentially sample averages of tensor products of kernels, which can be dealt with by Chebychev's inequality. A more delicate issue is to deal with the inverses: since, at the population level, these are compact operators, their inverses are unbounded, and cannot be estimated directly. To get around this, we employ Tychonoff regularization with a tuning parameter converging to 0 at a certain rate. It is this aspect of the asymptotic analysis that reflects the infinite-dimensional nature of the problem, which makes the convergence rate slower than $n^{-1/2}$. In fact, without these inverses, all moment estimators are convergent at the parametric $n^{-1/2}$ rate, and the problem is no different from a finite-dimensional problem. Once the convergence rates of the linear operators are determined, the convergence of eigenvectors are then calculated by perturbation theory. For further references for asymptotic analysis of linear operators and perturbation theory, see Koltchinskii and Giné (2000), Blanchard et al. (2007), Fukumizu et al. (2009), Li and Solea (2018b), and Li (2018b).

The second convergence rate in (15) is, however, a new problem specific to the current two-step procedure. It reveals how the error produced in extracting the sufficient predictor $\hat{U}_{ij}$ in the first step propagates into the estimation of the linear operator $\Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}$ in the second step. Theorems 7 through 9 and Theorem 12 below are concerned with this rate. For developing this rate we appeal to the reproducing property of the kernel to establish a uniform substitution error of $\hat{U}^{ij}$ for U^{ij}. This approach is novel and we expect it to be useful in other settings where one uses the sufficient predictors produced by nonlinear sufficient dimension reduction to replace the original random vectors conditioned upon. In some sense, this problem is akin to the post dimension reduction problem considered in Kim et al. (2020).

Once the convergence rates are established, we can then optimize them by varying the convergence rates of the tuning parameters in Tychonoff regularization. This is done in Theorem 14. The technique used in the optimization is similar to that used in Li and Song (2017, Corollary 4) and Li and Solea) and (2018b, Theorem 5).

In the following, if $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, then we write $a_n \prec b_n$ if $a_n/b_n \to 0$. We write $a_n \asymp b_n$ if $0 < \liminf_n (b_n/a_n) \leq \limsup_n (b_n/a_n) < \infty$. We write $b_n \preceq a_n$ if either $b_n \prec a_n$ or $b_n \asymp a_n$. Because (i,j) is fixed in the asymptotic development, and also to emphasize the dependence on n, in the rest of this section we denote $\epsilon_X^{(i,j)}$, $\epsilon_X^{-(i,j)}$, and $\epsilon_U^{(i,j)}$ by ϵ_n, η_n, and δ_n, respectively.

5.2 Transparent kernel

We first develop what we call the "transparent kernel" that passes information from step 1 to step 2 efficiently. Let Ω be a nonempty set, and $\kappa : \Omega \times \Omega \to \mathbb{R}$ a positive kernel.

Definition 6 *We say that κ is a transparent kernel if, for each $t \in \Omega$, the function $s \mapsto \kappa(s,t)$ is twice differentiable and*

1. $\partial\kappa(s,t)/\partial s|_{s=t} = 0$*;*
2. *the matrix $H(s,t) = \partial^2\kappa(s,t)/\partial s \partial s^{\mathsf{T}}$ has a bounded operator norm; that is, there exist $-\infty < C_1 \le C_2 < \infty$ such that*

$$C_1 \le \lambda_{\min}(H(s,t)) \le \lambda_{\max}(H(s,t)) < C_2$$

for all $(s,t) \in \Omega \times \Omega$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ indicate the largest and smallest eigenvalues.

For example, the Gaussian radial basis function kernel is transparent, but the exponential kernel $\kappa(u,v) = \tau^2 \exp(-\gamma\|u - v\|)$ is not. This condition implies a type of Lipschitz continuity in a setting that involves two reproducing kernels κ_0 and κ_1, where the argument of κ_1 is the evaluation of a member of the reproducing kernel Hilbert space generated by κ_0.

Theorem 7 *Suppose $\mathscr{H}_0$ is the reproducing kernel Hilbert space generated by κ_0, $\mathscr{H}_0^d$ is the d-fold Cartesian product of $\mathscr{H}_0$ with inner product defined by*

$$\langle U, V\rangle_{\mathscr{H}_0^d} = \langle u_1, v_1\rangle_{\mathscr{H}_0} + \cdots + \langle u_d, v_d\rangle_{\mathscr{H}_0}$$

where $U = (u_1, \ldots, u_d)$ and $V = (v_1, \ldots, v_d)$ are members of $\mathscr{H}_0^d$, $\mathscr{H}_1$ is the reproducing kernel Hilbert space generated by κ_1. Then:

(i) for any $U, V \in \mathscr{H}_0^d$, $a \in \Omega$, we have

$$\|U(a) - V(a)\|_{\mathbb{R}^d} \le [\kappa_0(a,a)]^{1/2} \, \|U - V\|_{\mathscr{H}_0^d};$$

(ii) if $\kappa_1(s,t)$ is a transparent kernel, then there exists a $C > 0$ such that, for each $U, V \in \mathscr{H}_0^d$ and $a \in \Omega$,

$$\|\kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\|_{\mathscr{H}_1} \le C \, [\kappa_0(a,a)]^{1/2} \, \|U - V\|_{\mathscr{H}_0^d}.$$

A direct consequence of this theorem is that, if $\hat{U}$ is an estimate of some U, a member of $\mathscr{H}_0^d$, with $\|\hat{U} - U\|_{\mathscr{H}_0^d} = O_P(b_n)$ for some $0 < b_n \to 0$, $\hat{\Sigma}(\hat{U})$ is a linear operator estimated from the sample $\hat{U}_1, \ldots, \hat{U}_n$ (and perhaps some other random vectors), and $\hat{\Sigma}(U)$ is a linear operator estimated from the sample $U_1, \ldots, U_n$, then,

$$\|\hat{\Sigma}(\hat{U}) - \hat{\Sigma}(U)\|_{\mathrm{HS}} = O_P(b_n). \tag{16}$$

This result is somewhat surprising, because sample estimates such as $\hat{\Sigma}(\hat{U})$ can be viewed as $E_n\mathbb{G}(X, \hat{U})$, where $\hat{U}$ is an estimate of a function U in a functional space with norm $\|\cdot\|$ and

$\mathbb{G}$ is an operator-valued function. If $\|\hat{U} - U\| = O_P(b_n)$ for some $b_n \to 0$, then it is not necessarily true that

$$\|E_n \mathbb{G}(X, \hat{U}) - E_n \mathbb{G}(X, U)\| = O_P(b_n),$$

particularly when U is an infinite dimensional object. Yet relation (16) states exactly this. The reason behind this is that the reproducing kernel property separates the function $\hat{U}$ and its argument X_a (i.e. $\hat{U}(x) = \langle \hat{U}, \kappa(\cdot, x) \rangle$), which implies a type of uniformity among $\hat{U}(X_1), \ldots, \hat{U}(X_n)$. This point will be made clear in the proof in the Appendix. Statement (16) is made precise by the next theorem.

Theorem 8 *Suppose conditions (1) and (2) of Definition 3 are satisfied with U, V, W therein replaced by X^i, X^j, and U^{ij}. Suppose, furthermore:*

(a) κ_U^{ij}, $\kappa_{XU}^{i,ij}$, *and* $\kappa_{XU}^{j,ij}$ *are transparent kernels;*

(b) $\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = O_P(b_n)$ *for some* $0 < b_n \to 0$.

Then

(i) $\|\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} - \hat{\Sigma}_{U^{ij}U^{ij}}\|_{\mathrm{HS}} = O_P(b_n)$*;*

(ii) $\|\hat{\Sigma}_{(X^i\hat{U}^{ij})\hat{U}^{ij}} - \hat{\Sigma}_{(X^iU^{ij})U^{ij}}\|_{\mathrm{HS}} = O_P(b_n)$*;*

(iii) $\|\hat{\Sigma}_{(X^i\hat{U}^{ij})(X^j\hat{U}^{ij})} - \hat{\Sigma}_{(X^iU^{ij})(X^jU^{ij})}\|_{\mathrm{HS}} = O_P(b_n)$*.*

Using Theorem 8 we can derive the convergence rate of $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}}$.

Theorem 9 *Suppose conditions in Theorem 8 are satisfied and, furthermore,*

(a) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(b) $b_n \preceq \delta_n \prec 1$.

Then $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} = O_P(b_n)$.

Note that, unlike in Theorem 8, where our assumptions imply

$$\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}$$

is a finite-rank operator, here, we do not assume $\Sigma^{-1}_{U^{ij}(U^{ij})}\Sigma_{U^{ij}(X^jU^{ij})}$ to be a finite-rank (or even Hilbert-Schmidt) operator; instead, we assume it to be a bounded operator. This is because (X^j, U^{ij}) contains U^{ij}, which makes it unreasonable to assume $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ to be finite-rank or Hilbert Schmidt. For example, when X^j is a constant, $\Sigma_{U^{ij}(X^jU^{ij})}$ is the same as $\Sigma_{U^{ij}U^{ij}}$ and $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}U^{ij}}$ is not a Hilbert Schmidt operator, though it is bounded. Theorem 9 shows that convergence rate of (ii) in (15) is the same as the convergence rate of (i) in (15); it now remains to derive the convergence rate of (i) and (iii).

5.3 Convergence rates of (i) and (iii) in (15)

We first present the convergence rate of $\hat{U}^{ij}$ to U^{ij}. The proof is similar to that of Theorem 5 of Li and Song (2017) but with two differences. First, Li and Song (2017) took A in (5) to be I, whereas we take it to be Σ_{YY}. In particular, the generalized sliced inverse regression in Li and Song (2017) only has one tuning parameter η_n, but we have two tuning parameters η_n and ϵ_n. Second, Li and Song (2017) defined (in the current notation) f_r^{ij} to be the eigenfunctions of

$$\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}\Sigma_{X^{(i,j)}X^{-(i,j)}}\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}},$$

which is different from the generalized eigenvalue problem (5). For these reasons we need to rederive the convergence rate of $\hat{U}^{ij}$.

Theorem 10 *Suppose*

(a) Assumption 1 is satisfied;

(b) $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *is a finite-rank operator with*

$$\begin{aligned}\operatorname{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) &\subseteq \operatorname{ran}(\Sigma^2_{X^{-(i,j)}X^{-(i,j)}}),\\ \operatorname{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) &\subseteq \operatorname{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});\end{aligned}$$

(c) $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$;

(d) for each $r = 1, \dots, d_{ij}$, $\lambda_1^{ij} > \cdots > \lambda^{ij}_{d_{ij}}$.

Then, $\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n)$.

An immediate consequence is that, under the transparent kernel assumption, the b_n in Theorem 9 is the same as this rate. We next derive the convergence rate in (iii) of (15). This rate depends on the tuning parameter δ_n in the estimate of conjoined conditional covariance operator, and it reaches b_n for the optimal choice of δ_n.

Theorem 11 *Suppose conditions (1) and (2) of Definition 3 are satisfied with U, V, W therein replaced by* X^i, X^j, *and* U^{ij}. *Suppose, furthermore,*

(a) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(b) $b_n \preceq \delta_n \prec 1$.

Then $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} = O_P(\delta_n)$. *Consequently, if* $\delta_n \asymp b_n$, *then*

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} = O_P(b_n).$$

Finally, we combine Theorem 9 through Theorem 11 to come up with the convergence rate of $\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}}$. Since there are numerous cross references among the conditions in these theorems, to make a clear presentation we list all the original conditions in the next theorem, even if they already appeared. These conditions are of two categories: those for the step 1 that involves sufficient dimension reduction of $X^{(i,j)}$ versus $X^{-(i,j)}$, and those for the step 2 that involves the estimation of the conjoined conditional covariance operator. We refer to them as the first-level and second-level conditions, respectively.

Theorem 12 *Suppose the following conditions hold:*

(a) *(First-level kernel)* $E[\kappa(S,S)] < \infty$ *for* $\kappa = \kappa_X^{(i,j)}$ *and* $\kappa = \kappa_X^{-(i,j)}$*;*

(b) *(First-level operator)* $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *is a finite-rank operator with rank* d_{ij} *and*

$$\begin{aligned}\operatorname{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) &\subseteq \operatorname{ran}(\Sigma^2_{X^{-(i,j)}X^{-(i,j)}}),\\ \operatorname{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) &\subseteq \operatorname{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});\end{aligned}$$

all the nonzero eigenvalues of $\Sigma_{X^{(i,j)}X^{-(i,j)}}\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *are distinct;*

(c) *(First-level tuning parameters)* $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$, $\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n \prec 1$*;*

(d) *(Second-level kernel)* $E[\kappa(S,S)] < \infty$ *is satisfied for* $\kappa = \kappa_U^{ij}$, $\kappa_{XU}^{i,ij}$, *and* $\kappa_{XU}^{j,ij}$*; furthermore, they are transparent kernels;*

(e) *(Second-level operators)* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(f) *(Second-level tuning parameter)* $\delta_n \asymp \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n$.

Then

$$\|\hat\Sigma_{\ddot X^i\ddot X^j|\hat U^{ij}} - \Sigma_{\ddot X^i\ddot X^j|U^{ij}}\|_{\mathrm{HS}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n). \tag{17}$$

Using this result we immediately arrive at the variable selection consistency of the Sufficient Graphical Model.

Corollary 13 *Under the conditions in Theorem 12, if*

$$\begin{aligned}&\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n \prec \rho_n \prec 1, \ \ \textit{and}\\ &\hat{\mathcal{E}} = \{(i,j) \in \Gamma\times\Gamma :\ i > j,\ \|\hat\Sigma_{\ddot X^i\ddot X^j|\hat U^{ij}}\|_{\mathrm{HS}} < \rho_n\}\end{aligned}$$

then $\lim_{n\to\infty} P(\hat{\mathcal{E}} = \mathcal{E}) \to 1$.

5.4 Optimal rates of tuning parameters

The convergence rate in Theorem 12 depends on ϵ_n and η_n explicitly, and δ_n implicitly (in the sense that $\delta_n \asymp \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n$ is optimal for fixed ϵ_n and η_n). Intuitively, when ϵ_n, η_n, and δ_n increase, the biases increase and variances decrease; when they decrease, the biases decrease and the variances increase. Thus there should be critical rates for them that balance the bias and variance, which are the optimal rates.

Theorem 14 *Under the conditions in Theorem 12, if* ϵ_n, η_n, *and* δ_n *are of the form* n^a, n^b, *and* n^c *for some* $a > 0$, $b > 0$, *and* $c > 0$, *then*

(i) *the optimal rates the tuning parameters are*

$$n^{-3/8} \preceq \epsilon_n \preceq n^{-1/4}, \quad \eta_n \asymp n^{-1/4}, \quad \delta_n \asymp n^{-1/4};$$

(ii) the optimal convergence rate of the estimated conjoined conditional covariance operator is

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} = O_P(n^{-1/4}).$$

Note that there is a range of ϵ_n are optimal, this is because the convergence rate does not have a unique minimizer. This also means the result is not very sensitive to this tuning parameter.

In the above asymptotic analysis, we have treated p as fixed when $n \to \infty$. We have also developed the consistency and convergence rate in the scenario where the dimension of p_n of X goes to infinity with n, which is placed in the Appendix I.

6. Simulation

In this section we compare the performance of our sufficient graphical model with previous methods such as Yuan and Lin (2007), Liu et al. (2009), Voorman et al. (2013), Fellinghauer et al. (2013), Lee et al. (2016b), and a Naïve method which is based on the conjoined conditional covariance operator without the dimension reduction step. The code is publicly available on Github: https://github.com/kyongwonkim/Sufficient-Graphical-Model.git.

By design, the sufficient graphical model has advantages over these existing methods under the following circumstances. First, since the sufficient graphical model does not make any distributional assumption, it should outperform Yuan and Lin (2007) and Liu et al. (2009) when the Gaussian or copula Gaussian assumptions are violated; second, due to the sufficient dimension reduction in sufficient graphical model, it avoids the curse of dimensionality and should outperform Voorman et al. (2013), Fellinghauer et al. (2013), and a Naïve method in the high-dimensional setting; third, since sufficient graphical model does not require additive structure, it should outperform Lee et al. (2016b) when there is severe nonadditivity in the model. Our simulation comparisons will reflect these aspects.

For the sufficient graphical model, Lee et al. (2016b), and the Naïve method, we use the Gaussian radial basis function as the kernel. The regularization constants $\epsilon_X^{(i,j)}$, $\epsilon_X^{-(i,j)}$, and $\epsilon_U^{(i,j)}$ are chosen by the generalized cross validation criterion described in Section 4.3 with the grid $\{10^{-\ell} : \ell = -1, 0, 1, 2, 3, 4\}$. The kernel parameters $\gamma_X^{(i,j)}$, $\gamma_X^{-(i,j)}$, $\gamma_{XU}^{i,ij}$, $\gamma_{XU}^{j,ij}$, and γ_U^{ij} are chosen according to (12). Because the outcomes of tuning parameters are stable, for each model, we compute the generalized cross validation for the first five samples and use their average value for the rest of the simulation.

The performance of each estimate is accessed using the averaged receiver operating characteristic (ROC) curve, which is a convenient visual representation of the accuracy of a classifier. Suppose that we have a set of subjects whose binary labels (say 0 and 1) are known, and a classifier that depends on a turning parameter ρ. For each value of ρ, the classifier gives two percentages: the percentage of true positive (i.e. classifying 1 as 1) and the percentage of false positive (i.e. classifying 0 as 1). Denoting the two percentages as $a(\rho)$ and $b(\rho)$, then the ROC curve is the set of points $\{(a(\rho), b(\rho)) : \rho \in I\}$, where I is an interval. Obviously, for any ρ, we prefer $a(\rho)$ to be large and $b(\rho)$ small, which means the area under the curve (AUC) measures the accuracy of a classifier across all tuning parameter values. In our setting, the set of subjects is the edge set, the labels 0 and 1 correspond to absence and presence of an edge, the classifier is the decision rule for an edge, and the turning parameter is the threshold ρ_n.

To isolate the factors that affect accuracy, we first consider two models with relatively small dimensions and large sample sizes, which are

$$
\begin{aligned}
\text{Model I}: \quad & X^1 = \epsilon_1,\ X^2 = \epsilon_2,\ X^3 = \sin(2X^1) + \epsilon_3 \\
& X^4 = (X^1)^2 + (X^2)^2 + \epsilon_4,\ X^5 = \epsilon_5, \\
\text{Model II}: \quad & X^1 = \epsilon_1,\ X^2 = X^1 + \epsilon_2,\ X^3 = \epsilon_3,\ X^4 = (X^1 + X^3)^2 + \epsilon_4, \\
& X^5 = \cos(2X^2X^3) + \epsilon_5,\ X^6 = X^4 + \epsilon_6,
\end{aligned}
$$

where ϵ_i, $i = 1, \ldots, p$ are from independent and identically distributed standard normal distribution. The edge sets of the two models are

$$
\begin{aligned}
\text{Model I}: \quad & \mathcal{E} = \{(1,3),(1,4),(2,4),(1,2)\} \\
\text{Model II}: \quad & \mathcal{E} = \{(1,2),(1,4),(3,4),(1,3),(2,5),(3,5),(2,3),(4,6)\}.
\end{aligned}
$$

We use $n = 100, 1000$ for each model, and for each n, we generate 50 samples to compute the averaged ROC curves. The dimension d_{ij} for sufficient graphical model is taken to be 2 for all cases (we have also used $d_{ij} = 1$ and the results are very similar to those presented here). The plots in the Figures 1 to 6 show the averaged ROC curves for the seven methods, with the following plotting symbol assignment:

Sufficient graphical model:	red solid line	Voorman et al. (2013):	red dotted line
Lee et al. (2016b):	black solid line	Fellinghauer et al. (2013):	black dotted line
Yuan and Lin (2007):	red dashed line	Naïve:	blue dotted line
Liu et al. (2009):	black dashed line		

From these figures we see that the two top performers are clearly sufficient graphical model and Lee et al. (2016b), and their performances are very similar. Note that none of the two models satisfies the Gaussian or copula Gaussian assumption, which explains why sufficient graphical model and Lee et al. (2016b) outperform Yuan and Lin (2007) and Liu et al. (2009). Sufficient graphical model and Lee et al. (2016b) also outperform Voorman et al. (2013), Fellinghauer et al. (2013), and Naïve method, indicating that curse of dimensionality already takes effect on the fully nonparametric methods. The three nonparametric estimators have similar performances. Also note that Model I has an additive structure, which explains the slight advantage of Lee et al. (2016b) over sufficient graphical model in Figure 1; Model II is not additive, and the advantage of Lee et al. (2016b) disappears in Figure 2.

We next consider two models with relatively high dimensions and small sample sizes. A convenient systematic way to generate larger networks is via the hub structure. We choose $p = 200$, and randomly generate ten hubs $h_1, \ldots, h_{10}$ from the 200 vertices. For each h_k, we randomly select a set H_k of 19 vertices to form the neighborhood of h_k. With the network structures thus specified, our two probabilistic models are

$$
\begin{aligned}
\text{Model III}: \quad & X^i = 1 + |X^{h_k}|^2 + \epsilon_i, \quad \text{where} \quad i \in H_k \setminus h_k, \\
\text{Model IV}: \quad & X^i = \sin((X^{h_k})^3)\epsilon_i, \quad \text{where} \quad i \in H_k \setminus h_k,
\end{aligned}
$$

and ϵ_i's are the same as in Models I and II. Note that, in Model III, the dependence of X_i on X_{h_k} is through the conditional mean $E(X_i|X_{h_k})$, whereas in Model IV, the dependence is through

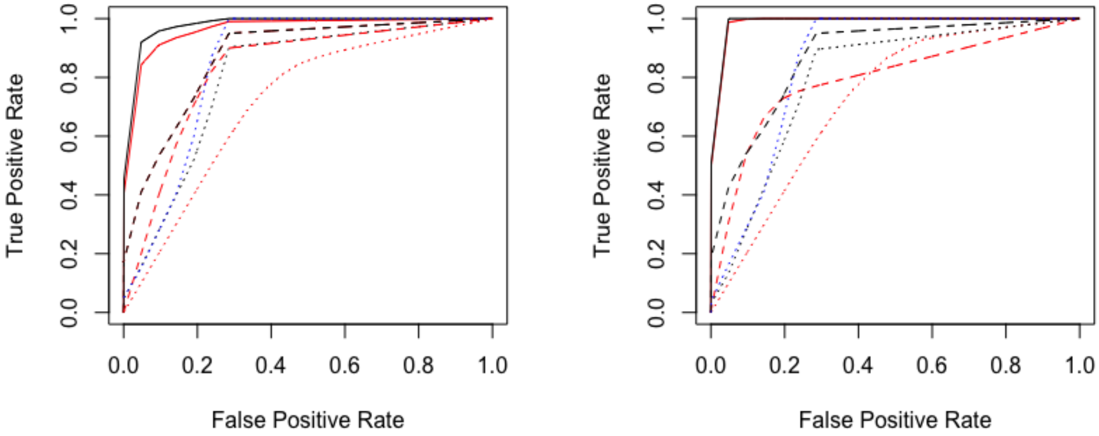

Figure 1: Averaged ROC curves for Model I. Left panel: $n = 100$; right panel: $n = 1000$.

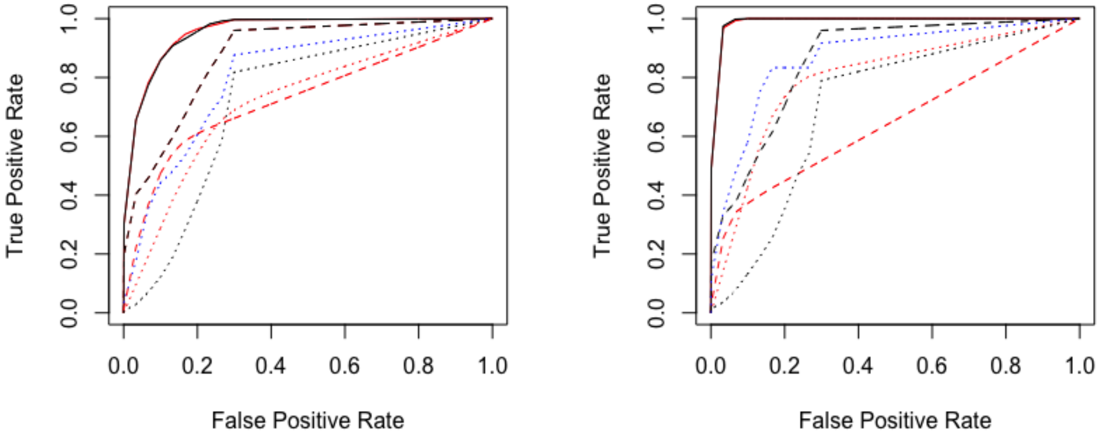

Figure 2: Averaged ROC curves for Model II. Left panel: $n = 100$; right panel: $n = 1000$.

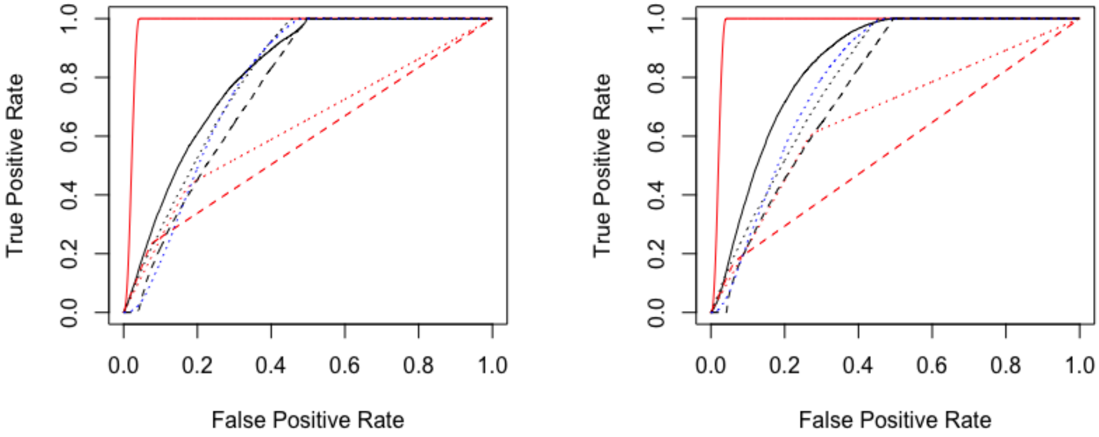

Figure 3: Averaged ROC curves for Model III with $p = 200$ case. Left panel: $n = 50$; right panel: $n = 100$.

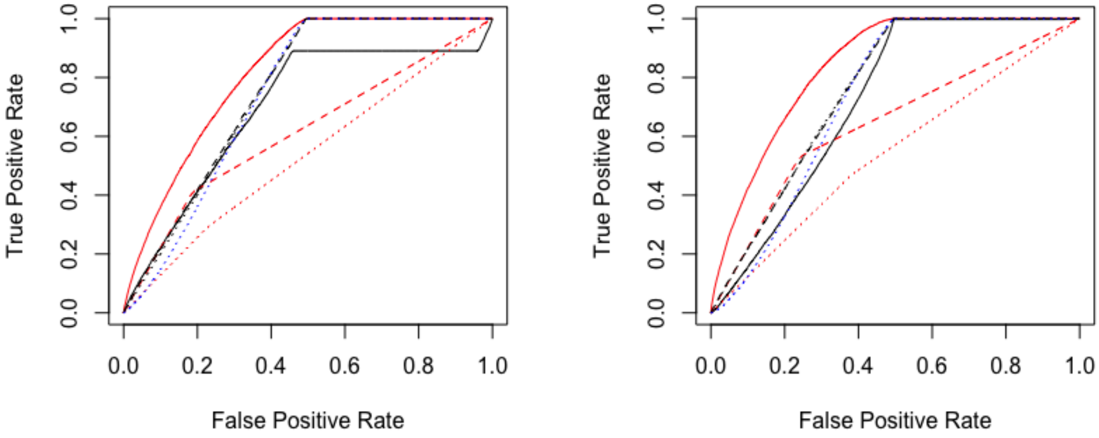

Figure 4: Averaged ROC curves for Model IV. Left panel: $n = 50$; right panel: $n = 100$.

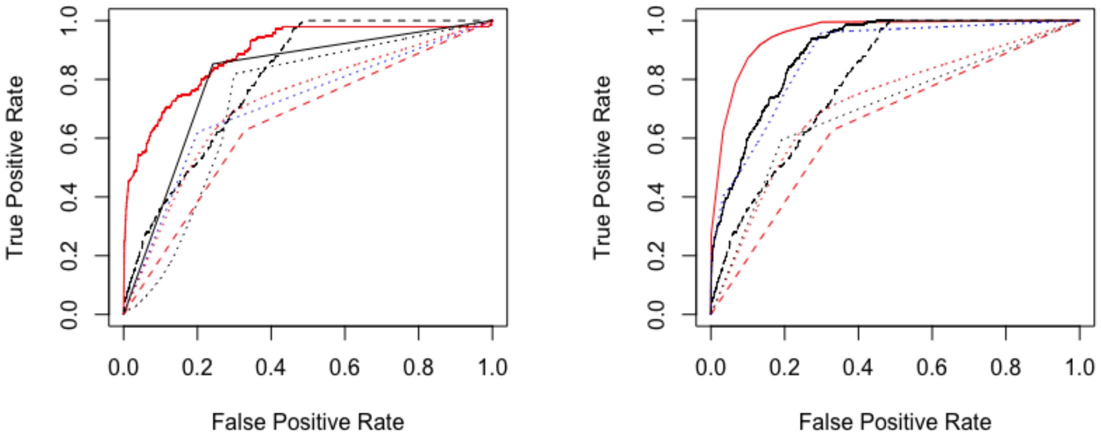

Figure 5: Averaged ROC curves for Model III with $p = 300$ case. Left panel: $n = 300$; right panel: $n = 500$.

the conditional variance $\text{var}(X_i|X_{h_k})$. For each model, we choose two sample sizes $n = 50$ and $n = 100$. The corresponding averaged ROC curves (averaged over 50 samples) are displayed in Figures 3 and 4. In particular, in the context of high-dimensional scenarios where $p > n$, the graphical model with sufficient dimension reduction consistently outperforms alternative methods. This observation underscores the advantages of dimension reduction in the construction of graphical models.

In Figure 5, we further increased the sample size and dimension in Model III to include $p = 300$ and $n = 300, 500$, respectively, while maintaining a hub count of 10. As we can see from this figure, as both the dimension and sample size increase, our method remains competitive, outperforming the alternative approaches.

We now consider a Gaussian graphical model to investigate any efficiency loss incurred by sufficient graphical model. Following the similar structure used in Li et al. (2014), we choose $p = 20$, $n = 100, 200$, and the model

$$\text{Model V} : X \sim N(0, \Theta^{-1}),$$

where Θ is 20×20 precision matrix with diagonal entries 1, 1, 1, 1.333, 3.010, 3.203, 1.543, 1.270, 1.544, 3, 1, 1, 1.2, 1, 1, 1, 1, 3, 2, 1, and nonzero off-diagonal entries $\theta_{3,5} = 1.418$, $\theta_{4,10} = -0.744$, $\theta_{5,9} = 0.519$, $\theta_{5,10} = -0.577$, $\theta_{13,17} = 0.287$, $\theta_{17,20} = 0.542$, $\theta_{14,15} = 0.998$. As expected, Figure 6 shows that Yuan and Lin (2007), Liu et al. (2009), and Lee et al. (2016b) perform better than sufficient graphical model in this case. However, sufficient graphical model still performs reasonably well and significantly outperforms the fully nonparametric methods.

Finally, we conducted some simulation on the generalized cross validation criterion (14) for determining the threshold ρ_n. We generated samples from Models I through V as described above, produced the ROC curves using sufficient graphical model, and determined the threshold ρ_n by

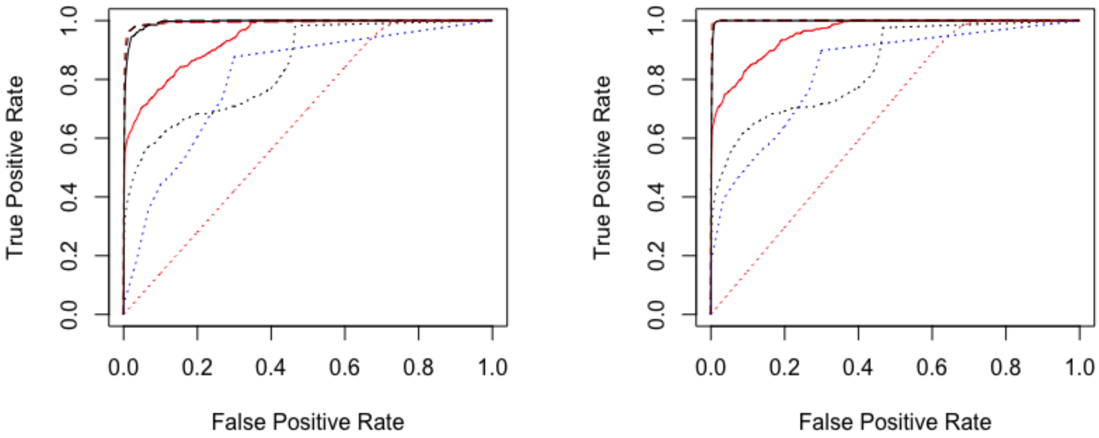

Figure 6: Averaged ROC curves for Model V. Left panel: $n = 100$; right panel: $n = 200$.

(14). The results are presented in Figure 7 in the Appendix. In each penal, the generalized cross validation-determined threshold ρ_n are represented by the black dots on the red ROC curves.

7. Application

We now apply sufficient graphical model to a data set from the DREAM 4 Challenge project and compare it with other methods. The goal of this Challenge is to recover gene regulation networks from simulated steady-state data. A description of this data set can be found in Marbach et al. (2010). Since Lee et al. (2016b) already compared their method with Yuan and Lin (2007), Liu et al. (2009), Voorman et al. (2013), Fellinghauer et al. (2013), and Naïve method for this dataset and demonstrated the superiority of Lee et al. (2016b) among these estimators, here we will focus on the comparison of the sufficient graphical model with Lee et al. (2016b) and the champion method for the DREAM 4 Challenge.

The data set contains data from five networks each of dimension of 100 and sample size 201. We use the Gaussian radial basis function kernel for sufficient graphical model and Lee et al. (2016b) and the tuning methods described in Section 4.3. For sufficient graphical model, the dimensions d_{ij} are taken to be 1. We have also experimented with $d_{ij} = 2$ but the results (not presented here) show no significant difference. Because networks are available, we can compare the ROC curves and their areas under the curve's, which are shown in Table 1.

Table 1: Comparison of sufficient graphical model, Lee et al. (2016b), Naïve and the champion methods in DREAM 4 Challenge

	Network 1	Network 2	Network 3	Network 4	Network 5
Sufficient graphical model	0.85	0.81	0.83	0.83	0.79
Lee et al. (2016b)	0.86	0.81	0.83	0.83	0.77
Champion	0.91	0.81	0.83	0.83	0.75
Naïve	0.78	0.76	0.78	0.76	0.71

As we can see from Table 1, sufficient graphical model has the same areas under the ROC curve values as Lee et al. (2016b) for Networks 2, 3, and 4, performs better than Lee et al. (2016b) for Network 5, but trails slightly behind Lee et al. (2016b) for Network 1; sufficient graphical model has the same areas under the curve as the champion method, performs better for Network 5 and worse for Network 1. Overall, sufficient graphical model and Lee et al. (2016b) perform similarly in this dataset, and they are on a par with the champion method. We should point out that sufficient graphical model and Lee et al. (2016b) are purely empirical; they employ no knowledge about the underlying physical mechanism generating the gene expression data. However, according to Pinna et al. (2010), the champion method did use a differential equation that reflects the underlying physical mechanism. The results for threshold determination are presented in Figure 8 in the Appendix.

8. Discussion

This paper is a first attempt to take advantage of the recently developed nonlinear sufficient dimension reduction method to nonparametrically estimate the statistical graphical model while avoiding the curse of dimensionality. Nonlinear sufficient dimension reduction is used as a module and applied repeatedly to evaluate conditional independence, which leads to a substantial gain in accuracy in the high-dimensional setting. Compared with the Gaussian and copula Gaussian methods, our method is not affected by the violation of the Gaussian and copula Gaussian assumptions. Compared with the additive method (Lee et al., 2016b), our method does not require an additive structure and retains the conditional independence as the criterion to determine the edges, which is a commonly accepted criterion. Compared with fully nonparametric methods, sufficient graphical model avoids the curse of dimensionality and significantly enhances the performance.

The present framework opens up several directions for further research. First, the current model assumes that the central class $\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$ is complete, so that generalized sliced inverse regression is the exhaustive nonlinear sufficient dimension reduction estimate. When this condition is violated, generalized sliced inverse regression is no longer exhaustive and we can employ other nonlinear sufficient dimension reduction methods such as the generalized sliced averaged variance estimation (Lee et al., 2013; Li, 2018b) to recover the part of the central class that generalized sliced inverse regression misses. Second, though we have assumed that there is a proper sufficient sub-σ-field $\mathcal{G}^{-(i,j)}$ for each (i, j), the proposed estimation procedure is still justifiable when no such sub-σ-field exists. In this case, U^{ij} is still the most important set of functions that characterize the statistical dependence of $X^{(i,j)}$ on $X^{-(i,j)}$ – even though it is not sufficient. Without sufficiency, our method may be more appropriately called the Principal Graphical Model than the sufficient graphical model. Third, the current method can be extended to functional graphical model, which are common in medical applications such as EEG and fMRI. Several functional graphical models have been proposed recently, by Zhu et al. (2016), Qiao et al. (2019), Li and Solea (2018b), and Solea and Li (2022). The idea of a sufficient graph can be applied to this setting to improve efficiency. Finally, given the sample size n and the number of nodes p, the proposed 2-step procedure requires inversions of several $n \times n$ matrices for each pair of nodes. This results in computation complexity of $O(p^2n^3)$ for constructing the entire graph, which could be burdensome for large-scale data. Identifying strategies to mitigate the computational cost would be a promising topic for future research.

This paper also contains some theoretical advances that are novel to nonlinear sufficient dimension reduction. For example, it introduces a general framework to characterize how the error of nonlinear sufficient dimension reduction propagates to the downstream analysis in terms of convergence rates. Furthermore, the results for convergence rates of various linear operators allowing the dimension of the predictor to go to infinity are the first of its kind in nonlinear sufficient dimension reduction. These advances will benefit the future development of sufficient dimension reduction in general, beyond the current context of estimating graphical models.

Acknowledgments

We are grateful to the Editor, Action Editor, and three Reviewers for their insightful comments and valuable suggestions, which have helped us greatly in improving our manuscript. Bing Li's research on this work was supported in part by the NSF Grants DMS-1713078, DMS-2210775, and the NIH grant 1 R01 GM15281201. Kyongwon Kim, the corresponding author, was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2021R1F1A1046976, RS-2023-00219212), basic Science Research Program through NRF funded by the Ministry of Education (2021R1A6A1A10039823), and under the High-Potential Individuals Global Training Program (RS-2022-00154879) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

Francis. R. Bach and Michael. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Robert Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*, 3(2), 1961.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294, 2007.

R Dennis Cook. Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the section on Physical and Engineering Sciences*, pages 18–25, 1994.

R Dennis Cook and Sanford Weisberg. Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

A Phillip. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael Von Rhein, and Jan D Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

Kenji Fukumizu, Francis R Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383, 2007.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, pages 489–496, 2008.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, pages 1871–1905, 2009.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 1979.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Alexandros. Karatzoglou, Alex. Smola, Kurt. Hornik, and Achim. Zeileis. Kernlab – an s4 package for kernel methods in r. *Journal of Statististical Software*, 11(9):1–20, 2004.

Kyongwon Kim, Bing Li, Zhou Yu, and Lexin Li. On post dimension reduction statistical inference. to appear in *The Annals of Statistics*, 2020.

Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, pages 113–167, 2000.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

Kuang.-Yao. Lee, Bing. Li, and Hongyu. Zhao. Variable selection via additive conditional independence. *Journal of the Royal Statistical Society: Series B*, 78:1037–1055, 2016a.

Kuang-Yao Lee, Bing Li, and Hongyu Zhao. On an additive partial correlation operator and nonparametric estimation of graphical models. *Biometrika*, 103(3):513–530, 2016b.

Bing Li. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46(1):79–103, 2018a.

Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press, 2018b.

Bing Li and Eftychia Solea. A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113(just-accepted):1637–1655, 2018a.

Bing Li and Eftychia Solea. A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113:1637–1655, 2018b.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095, 2017.

Bing. Li, Adreas. Artemiou, and Lexin. Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39:3182–3210, 2011.

Bing Li, Hyonho Chun, and Hongyu Zhao. On an additive semigraphoid model for statistical networks with application to pathway analysis. *Journal of the American Statistical Association*, 109(507):1188–1204, 2014.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012a.

Han Liu, Fang Han, and Cun-Hui Zhang. Transelliptical graphical models. *Advances in neural information processing systems*, pages 800–808, 2012b.

Wei Luo and Bing Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103:875–887, 2016.

Wei Luo and Bing Li. On order determination by predictor augmentation. *Biometrika (*To appear*)*, 103:875–887, 2020.

Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291, 2010.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

J. Pearl and T. Verma. *The logic of representing dependencies by directed graphs*. University of California (Los Angeles). Computer Science Department, 1987.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Andrea Pinna, Nicola Soranzo, and Alberto de la Fuente. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PLoS One*, 5(10), 2010.

Xinghao Qiao, Shaojun Guo, and Gareth M. James. Functional graphical models. *Journal of the American Statistical Association*, 114:211–222, 2019.

Eftychia Solea and Bing Li. Copula gaussian graphical models for functional data. *Journal of the American Statistical Association*, 117(538):781–793, 2022.

Bharath. K. Sriperumbudur, Kenji. Fukumizu, and Gert RG. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12, 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2013.

Yu. Wang. Nonlinear dimension reduction in feature space. *PhD Thesis, The Pennsylvania State University*, 2008.

Han-Ming. Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Hongxiao Zhu, Nate Strawn, and David B. Dunson. Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 14:1–27, 2016.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix

This Appendix consists of three parts:

1. Proofs of all theorems, lemmas, corollaries, and propositions in the paper. These are done in Appendixes A through H;
2. Some additional theoretical results that are quoted by the paper, including the asymptotic development for the high-dimensional setting (Section I), examples of joint distributions satisfying condition (3) in the paper, and a necessary and sufficient condition for $\mathrm{ran}(B) \subseteq \mathrm{ran}(A)$ (Appendix K);
3. Some additional simulation plots for threshold determination quoted in the paper (Appendix M).

Appendix A. Preliminaries

A.1 Hilbert-Schmidt norm and operator norm

Lemma 15 *If $\mathscr{H}_1$ and $\mathscr{H}_2$ are Hilbert spaces and f and g are members of $\mathscr{H}_1$ and $\mathscr{H}_2$, respectively, then $\|f \otimes g\|_{\mathrm{HS}} = \|f\|_{\mathscr{H}_1} \|g\|_{\mathscr{H}_2}$.*

Proof Because

$$(f \otimes g)(f \otimes g)^* f = (f \otimes g)(g \otimes f) f = f \langle g, g \rangle_{\mathscr{H}_2} \langle f, f \rangle_{\mathscr{H}_1},$$

$\|g\|^2_{\mathscr{H}_2} \|f\|^2_{\mathscr{H}_1}$ is the eigenvalue of the rank-1 operator $(f \otimes g)(f \otimes g)^*$, which is by definition the Hilbert-Schmidt norm $\|f \otimes g\|^2_{\mathrm{HS}}$. □

Lemma 16 *If A and B are linear operators, then*

$$\|AB\|_{\mathrm{HS}} \le \|A\|_{\mathrm{OP}} \|B\|_{\mathrm{HS}}$$

Proof Recall that $\|AB\|^2_{\mathrm{HS}} = \mathrm{tr}(B^* A^* AB)$. Because

$$A^* A \le \lambda_{\max}(A^* A) I,$$

we have

$$\mathrm{tr}(B^* A^* AB) \le \lambda_{\max}(A^* A) \mathrm{tr}(B^* B) = \|A\|^2_{\mathrm{OP}} \|B\|^2_{\mathrm{HS}}. \qquad \square$$

Corollary 17 *If $A_1, \ldots, A_m$ are bounded linear operators with at least one of them, say A_i being a Hilbert-Schmidt operator, then*

$$\|A_1 \cdots A_i \cdots A_m\|_{\mathrm{HS}} \le \|A_1\|_{\mathrm{OP}} \cdots \|A_i\|_{\mathrm{HS}} \cdots \|A_m\|_{\mathrm{OP}}.$$

Lemma 18 *If A and B are self adjoint Hilbert Schmidt operators, then*

$$\|AB\|_{\mathrm{HS}} \le \|A\|_{\mathrm{HS}} \|B\|_{\mathrm{HS}}.$$

Proof This follows from Lemma 16 and the fact that, for any self adjoint operator A, $\|A\|_{\mathrm{OP}} \leq \|A\|_{\mathrm{HS}}$. □

Corollary 19 *If $A_1, \ldots, A_m$ are self adjoint Hilbert Schmidt operators, then*

$$\|A_1 \cdots A_m\|_{\mathrm{HS}} \leq \|A_1\|_{\mathrm{HS}} \cdots \|A_m\|_{\mathrm{HS}}.$$

A.2 Covariance operator and mean element in reproducing kernel Hilbert space

In this subsection we give formal definitions of various concepts used in Section 3, such as the mean element, the covariance operator, the centered reproducing kernel Hilbert space, and the inverse of an operator. For a linear operator A in a Hilbert space $\mathscr{H}$, let $\ker(A) = \{h \in \mathscr{H} : Ah = 0\}$ be the kernel of A, $\mathrm{ran}(A) = \{Ah : h \in \mathscr{H}\}$ the range of A, and $\overline{\mathrm{ran}}(A)$ the closure of $\mathrm{ran}(A)$.

We first introduce the generic notion of the centered reproducing kernel Hilbert space. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and U is a random element defined on $(\Omega, \mathcal{F}, P)$ taking values in $(\Omega_U, \mathcal{F}_U)$. Let $\kappa : \Omega_U \times \Omega_U \to \mathbb{R}$ be a positive kernel. The reproducing kernel Hilbert space generated by the kernel κ is the completion of the linear span of the set of functions $\{\kappa(\cdot, u) : u \in \Omega_U\}$ with the inner product between members of the linear span determined by $\langle u_1, u_2 \rangle = \kappa(u_1, u_2)$. Let us denote this reproducing kernel Hilbert space as $\mathscr{K}_U$. Under the assumption that $E\kappa(U, U) < \infty$, the mean element μ_U of U is a well defined element of $\mathscr{K}_U$ and the covariance operator Σ_{UU} of U is a well defined linear operator that maps from $\mathscr{K}_U$ to $\mathscr{K}_U$, and they are uniquely defined by the relations

1. $\langle f, \mu_U \rangle_{\mathscr{K}_U} = Ef(U)$ for each $f \in \mathscr{K}_U$;

2. $\langle f, \Sigma_{UU} g \rangle_{\mathscr{K}_U} = \mathrm{cov}[f(U), g(U)]$ for all $f, g \in \mathscr{K}_U$.

See, for example, Li (2018b). To study statistical relations, we can, without loss of generality, reset $\mathscr{K}_U$ to be $\overline{\mathrm{ran}}(\Sigma_{UU})$. This is because $\overline{\mathrm{ran}}(\Sigma_{UU}) = \ker(\Sigma_{UU})^{\perp}$, any $f \in \ker(\Sigma_{UU})$ satisfies $\mathrm{var}[f(U)] = 0$. Hence $\ker(\Sigma_{UU})$ consists of functions of U that are almost surely constants. Such functions can be removed from $\mathscr{K}_U$ without affecting any statistical relation. Thus it suffices to consider the subspace $\overline{\mathrm{ran}}(\Sigma_{UU}) \equiv \mathscr{H}_U$ of $\mathscr{K}_U$. We call $\mathscr{H}_U$ the centered reproducing kernel Hilbert space generated by κ. Li and Song (2017) (in Lemma 1) showed that $\mathscr{H}_U$ is the closed subspace of $\mathscr{K}_U$ spanned by the set of functions $\{\kappa_U(\cdot, u) - \mu_U : u \in \Omega_U\}$. Since $\ker(\Sigma_{UU}) = 0$ when Σ_{UU} is restricted on $\mathscr{H}_U$, it is an injective linear operator. We use Σ_{UU}^{-1} to denote the operator from $\mathrm{ran}(\Sigma_{UU})$ to $\overline{\mathrm{ran}}(\Sigma_{UU}) = \mathscr{H}_U$ that sends $\Sigma_{UU} h$ to h. This inverse, however, is not a bounded operator because the operator Σ_{UU}, if it is defined, is a Hilbert-Schmidt operator.

Next, let V be another random element defined on $(\Omega, \mathcal{F}, P)$ taking values in $(\Omega_V, \mathcal{F}_V)$, $\kappa_V : \Omega_V \times \Omega_V \to \mathbb{R}$ a positive definite kernel that satisfies $E\kappa_V(V, V) < \infty$, and $\mathscr{H}_V$ the centered reproducing kernel Hilbert space generated by κ_V. The covariance operator Σ_{UV} is a mapping from $\mathscr{H}_V$ to $\mathscr{H}_U$ that satisfies

$$\langle f, \Sigma_{UV} g \rangle_{\mathscr{H}_U} = \mathrm{cov}[f(U), g(V)]$$

for each $f \in \mathscr{H}_U$ and $g \in \mathscr{H}_V$.

The function μ_U and the linear operators such as Σ_{UU} and Σ_{UV} can be represented explicitly in terms of kernels, as follows:

$$\begin{aligned}\mu_U &= E\kappa_U(\cdot, U),\\ \Sigma_{UU} &= E\{[\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)] \otimes [\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)]\},\\ \Sigma_{UV} &= E\{[\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)] \otimes [\kappa_V(\cdot, V) - E\kappa_V(\cdot, V)]\},\end{aligned} \tag{18}$$

where $\otimes$ is the tensor product, $\kappa_U(\cdot, U)$ is the function $\Omega_U \to \mathbb{R}$, $u \mapsto \kappa_U(u, U)$, and $E\kappa_U(\cdot, U)$ is the function $\Omega_U \to \mathbb{R}$, $\mapsto E[\kappa_U(u, U)]$.

A.3 Sample mean of operators

The following lemma is taken from Fukumizu et al. (2007).

Lemma 20 *Suppose*

(a) U_1 *and* U_2 *are random vectors taking values in* $\Omega_{U_1} \subseteq \mathbb{R}^{p1}$ *and* $\Omega_{U_2} \subseteq \mathbb{R}^{p2}$*, respectively;*

(b) $\kappa_1 : \Omega_{U_1} \times \Omega_{U_1} \to \mathbb{R}$, $\kappa_2 : \Omega_{U_2} \times \Omega_{U_2} \to \mathbb{R}$ *are positive kernel functions such that* $E[\kappa_1(U_1, U_1)] < \infty$ *and* $E[\kappa_2(U_2, U_2)] < \infty$*;*

(c) $(U_{11}, U_{21}), \ldots, (U_{1n}, U_{2n})$ *are an i.i.d. sample of* (U_1, U_2)*.*

Then $\|\hat{\Sigma}_{U_1U_2} - \Sigma_{U_1U_2}\|_{\mathrm{HS}} = O_P(n^{-1/2})$*.*

A.4 Tychonoff regularized inverse

Henceforth, we say that a linear operator A is a CSP operator if it is compact, self-adjoint, and positive semidefinite. Note that A being positive semidefinite implies that A is injective; that is, $\ker(A) = \{0\}$. If $A : \mathscr{H} \to \mathscr{H}$ is an injective linear operator, we define A^{-1} to be the linear operator from $\mathrm{ran}(A)$ to $\mathscr{H}$ such that, for any $g \in \mathrm{ran}(A)$, $A^{-1}g$ is the unique element $f \in \mathscr{H}$ such that $Af = g$. For any $\alpha > 0$, we denote the operator $(A^{\alpha})^{-1}$ by $A^{-\alpha}$. The conditions for the following lemma are not the weakest possible, but they make the proof simple and they are all we will need.

Lemma 21 *Suppose* $\mathscr{H}$ *and* $\mathscr{K}$ *are Hilbert spaces and*

(a) $A_1 : \mathscr{K} \to \mathscr{K}$ *is a CSP operator;*

(b) $A_2 : \mathscr{H} \to \mathscr{K}$ *is a finite-rank linear operator;*

(c) $\alpha > 0$*, and* $\mathrm{ran}(A_2) \subseteq \mathrm{ran}(A_1^{\alpha})$*.*

Then, for any $\eta > 0$*,* $(A_1 + \eta I)^{-\alpha} A_2$ *is a finite-rank operator with*

$$\|(A_1 + \eta I)^{-\alpha} A_2\|_{\mathrm{HS}} \leq \|A_1^{-\alpha} A_2\|_{\mathrm{HS}}. \tag{19}$$

Condition (c) is equivalent to $\mathrm{ran}(A_2) \subseteq \mathrm{dom}(A_1^{-\alpha})$, so that the operator $A_1^{-\alpha} A_2$ is a well defined finite-rank operator.

Proof First, since $A_1^{-\alpha} A_2$ is a finite-rank operator, it is a Hilbert-Schmidt operator. Because

$$(A_1 + \eta I)^{-\alpha} A_2 = (A_1 + \eta I)^{-\alpha} A_1^{\alpha} A_1^{-\alpha} A_2$$

and $(A_1 + \eta I)^{-\alpha} A_1^{\alpha} \leq (A_1 + \eta I)^{-\alpha} (A_1 + \eta I)^{\alpha} = I$, we have

$$A_2^* [(A_1 + \eta I)^{-\alpha}]^2 A_2 = A_2^* A_1^{-\alpha} [(A_1 + \eta I)^{-\alpha} A_1^{\alpha}]^2 A_1^{-\alpha} A_2 \leq A_2^* (A_1^{-\alpha})^2 A_2,$$

where the first equality holds because $(A_1 + \eta I)^{-\alpha}$ and A_1^{α} commute. Hence the trace norm of the left is no greater than the trace norm of the right, which is equivalent to (19). □

Corollary 22 *Suppose*

1. $A_1 : \mathscr{K} \to \mathscr{K}$ *and* $A_3 : \mathscr{H} \to \mathscr{H}$ *are CSP operators;*
2. $A_2 : \mathscr{H} \to \mathscr{K}$ *is a finite rank linear operator;*
3. $\alpha > 0$, $\beta > 0$, $\mathrm{ran}(A_2) \subseteq \mathrm{ran}(A_1^{\alpha})$, $\mathrm{ran}(A_2^*) \subseteq \mathrm{ran}(A_3^{\beta})$.

Then $(A_1 + \eta I)^{-\alpha} A_2 (A_3 + \epsilon I)^{-\beta}$ *is a finite-rank operator and*

$$\|(A_1 + \eta I)^{-\alpha} A_2 (A_3 + \epsilon I)^{-\beta}\|_{\mathrm{HS}} \leq \|A_1^{-\alpha} A_2 A_3^{-\beta}\|_{\mathrm{HS}}. \tag{20}$$

Proof Again, it is obvious that $A_1^{-\alpha} A_2 A_3^{-\beta}$ is a finite-rank operator, so it has a finite Hilbert-Schmidt norm. Since the conditions in Lemma 21 are satisfied for A_1 and A_2 therein replaced by A_1 and $A_2 A_3^{-\beta}$ in this corollary, the operator $(A_1 + \eta I)^{-\alpha} A_2 A_3^{-\beta}$ is Hilbert Schmidt with

$$\|(A_1 + \eta I)^{-\alpha} A_2 A_3^{-\beta}\|_{\mathrm{HS}} \leq \|A_1^{-\alpha} A_2 A_3^{-\beta}\|_{\mathrm{HS}}. \tag{21}$$

Similarly, since the conditions in Lemma 21 are satisfied for A_1 and A_2 therein replaced by A_3 and $A_2^*(A_1 + \eta I)^{-\alpha}$ in this corollary, the operator $(A_3 + \epsilon I)^{-\beta} A_2^* (A_1 + \eta I)^{-\alpha}$ is Hilbert Schmidt with

$$\begin{aligned}\|(A_3 + \epsilon I)^{-\beta} A_2^* (A_1 + \eta I)^{-\alpha}\| &\leq \|A_3^{-\beta} A_2^* (A_1 + \eta I)^{-\alpha}\|_{\mathrm{HS}} \\ &= \|(A_1 + \eta I)^{-\alpha} A_2^* A^{-\beta}\|_{\mathrm{HS}},\end{aligned}$$

where the right-hand side, by (21), is no greater than $\|A_1^{-\alpha} A_2 A^{-\beta}\|_{\mathrm{HS}}$. □

Lemma 23 *Suppose* $\mathscr{H}$ *and* $\mathscr{K}$ *are Hilbert spaces and*

(a) $A_1 : \mathscr{K} \to \mathscr{K}$ *is a CSP operator;*

(b) $\alpha > 0$, *and* $\mathrm{ran}(A_2) \subseteq \mathrm{ran}(A_1^{\alpha})$; $A_2^{-\alpha} A_1$ *is a bounded linear operator.*

Then, for any $\eta > 0$, $(A_1 + \eta I)^{-\alpha} A_2$ *is a finite-rank operator with*

$$\|(A_1 + \eta I)^{-\alpha} A_2\|_{\mathrm{OP}} \leq \|A_1^{-\alpha} A_2\|_{\mathrm{OP}}. \tag{22}$$

The proof is similar to that of Lemma 21 and is omitted.

A.5 Negative square root

The next lemma can be verified by simple computation (see, for example, Fukumizu, Bach, and Gretton (2007)).

Lemma 24 *If A and B are self adjoint and invertible linear operators, then*

$$\begin{aligned} A^{-1/2} - B^{-1/2} &= A^{-3/2}(B^{3/2} - A^{3/2})B^{-1/2} + A^{-3/2}(A - B) \\ &= A^{-1/2}(B^{3/2} - A^{3/2})B^{-3/2} + (A - B)B^{-3/2}. \end{aligned}$$

A.6 Notations for order of magnitude

If $\{A_n\}$ is a sequence of random operators and $\{a_n\}$ is a sequence of positive numbers such that $\|A_n\|_{\mathrm{OP}} = O_P(a_n)$, then we write $A_n = \dot{O}_P(a_n)$. If $\|A_n\|_{\mathrm{HS}} = O_P(a_n)$, then we write $A_n = \ddot{O}_P(a_n)$. Note that $A_n = \ddot{O}_P(a_n)$ implies $A_n = \dot{O}_P(a_n)$, and $\ddot{O}_P(a_n)\dot{O}_P(b_n) = \ddot{O}_P(a_n b_n)$. Similarly, if $\|A_n\|_{\mathrm{OP}} = o_P(a_n)$, then we write $A_n = \dot{o}_P(a_n)$. If $\|A_n\|_{\mathrm{HS}} = o_P(a_n)$, then we write $A_n = \ddot{o}_P(a_n)$.

Also, as already mentioned in the main manuscript, if $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, then we write $a_n \prec b_n$ if $a_n/b_n \to 0$. We write $a_n \asymp b_n$ if $0 < \liminf_n(b_n/a_n) \leq \limsup_n(b_n/a_n) < \infty$. We write $b_n \preceq a_n$ if either $b_n \prec a_n$ or $b_n \asymp a_n$.

Appendix B. Proof of Theorem 1

Proof of $X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \Rightarrow X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$. Since $\mathcal{G}^{-(i,j)} \subseteq \sigma(X^{-(i,j)})$, we have

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \Leftrightarrow X^i \perp\!\!\!\perp X^j | (X^{-(i,j)}, \mathcal{G}^{-(i,j)}).$$

Hence

$$\begin{aligned} \begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \\ (X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} &\Rightarrow \begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)}, \mathcal{G}^{-(i,j)} \\ (X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} \\ &\Rightarrow \begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)}, \mathcal{G}^{-(i,j)} \\ X^i \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} \\ &\Rightarrow X^i \perp\!\!\!\perp (X^j, X^{-(i,j)}) | \mathcal{G}^{-(i,j)} \\ &\Rightarrow X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}, \end{aligned}$$

where the first implication follows from statement 2 of Theorem 2.1 of Li (2018b); the second from statement 4; the third from statement 2 again.

Proof of $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)} \Rightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)}$. Let $A \in \sigma(X^i)$, $B \in \sigma(X^j)$. It suffices to show that

$$P(X^i \in A, X^j \in B | X^{-(i,j)}) = P(X^i \in A | X^{-(i,j)}) P(X^j \in B | X^{-(i,j)}).$$

Because $(X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$, the left hand side is

$$P(X^i \in A, X^j \in B | \mathcal{G}^{-(i,j)}),$$

which, by condition 2, is equal to $P(X^i \in A|\mathcal{G}^{-(i,j)})P(X^j \in B|\mathcal{G}^{-(i,j)})$. However, by statement 2 of Theorem 2.1 of Li (2018b), we have $X^i \perp\!\!\!\perp X^{-(i,j)}|\mathcal{G}^{-(i,j)}$ and $X^j \perp\!\!\!\perp X^{-(i,j)}|\mathcal{G}^{-(i,j)}$. Hence

$$P(X^i \in A|\mathcal{G}^{-(i,j)})P(X^j \in B|\mathcal{G}^{-(i,j)}) = P(X^i \in A|X^{-(i,j)})P(X^j \in B|X^{-(i,j)}),$$

as desired.

Appendix C. Proof of Theorem 7

(i) Because, by the reproducing property of an reproducing kernel Hilbert space,

$$u_r(a) = \langle u_r,\, \kappa_0(\cdot, a)\rangle_{\mathscr{H}}, \quad v_r(a) = \langle v_r,\, \kappa_0(\cdot, a)\rangle_{\mathscr{H}}, \quad r = 1, \ldots, d,$$

we have

$$\begin{aligned}\|U(a) - V(a)\|^2_{\mathbb{R}^d} &= \sum_{r=1}^{d} \langle u_r - v_r, \kappa_0(\cdot, a)\rangle^2_{\mathscr{H}} \le \sum_{r=1}^{d} \|u_r - v_r\|^2_{\mathscr{H}} \|\kappa_0(\cdot, a)\|^2_{\mathscr{H}} \\ &= \kappa_0(a, a) \sum_{r=1}^{d} \|u_r - v_r\|^2_{\mathscr{H}} = \kappa_0(a, a)\|U - V\|^2_{\mathscr{H}^d}.\end{aligned}$$

Now take square root on both sides to complete the proof of (a).
(ii) By the definition of the inner product in an reproducing kernel Hilbert space,

$$\begin{aligned}&\|\kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\|^2_{\mathscr{H}_1} \\ &= \langle \kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a)), \kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\rangle_{\mathscr{H}_1} \\ &= \kappa_1(U(a), U(a)) - 2\kappa_1(U(a), V(a)) + \kappa_1(V(a), V(a)) \\ &\le |\kappa_1(U(a), U(a)) - \kappa_1(V(a), U(a))| \\ &\quad + |\kappa_1(U(a), V(a)) - \kappa_1(V(a), V(a))|.\end{aligned} \tag{23}$$

By Taylor's mean value theorem

$$\begin{aligned}&\kappa_1(V(a), U(a)) - \kappa_1(U(a), U(a)) \\ &= \left[\frac{\partial \kappa_1(s, U(a))}{\partial s}\right]_{s=U(a)} [V(a) - U(a)] \\ &\quad + \frac{1}{2}[V(a) - U(a)]^{\mathsf{T}} \left[\frac{\partial^2 \kappa_1(s, U(a))}{\partial s \partial s^{\mathsf{T}}}\right]_{s=\xi} [V(a) - U(a)]\end{aligned}$$

for some ξ in the line joining $U(a)$ and $V(a)$. Since, by assumption, the first derivative above is 0, and the second derivative has bounded eigenvalues, there is a constant $C_1 > 0$ such that

$$\begin{aligned}|\kappa_1(V(a), U(a)) - \kappa_1(U(a), U(a))| &\le C_1\|V(a) - U(a)\|^2_{\mathbb{R}^d} \\ &\le C_1\|V - U\|^2_{\mathscr{H}^d}\, \kappa_0(a, a),\end{aligned} \tag{24}$$

where the second inequality follows from part (i). By similar computation, we have, for a constant $C_1 > 0$ (which can be taken as the same constant above),

$$|\kappa_1(U(a), V(a)) - \kappa_1(V(a), V(a))| \le 2^{-1}C_1\|V - U\|^2_{\mathscr{H}^d}\, \kappa_0(a, a). \tag{25}$$

Substitute (24) and (25) into the right-hand side of (23) to prove (ii). □

Appendix D. Proof of Theorem 8

(i) Tentatively abbreviating $U^{ij}(X^{-(i,j)})$ and $\hat{U}^{ij}(X^{-(i,j)})$ by U^{ij} and $\hat{U}^{ij}$, we have

$$\begin{aligned}
&\|\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} - \hat{\Sigma}_{U^{ij}U^{ij}}\|_{\text{HS}} \\
&= \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij})) \\
&\quad - [E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\
&\qquad\qquad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\|_{\text{HS}} \\
&\leq \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij}))\|_{\text{HS}} \\
&\quad + \|[E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\
&\qquad\qquad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\|_{\text{HS}} \\
&\equiv \|\Delta_n^{(1)}\|_{\text{HS}} + \|\Delta_n^{(2)}\|_{\text{HS}},
\end{aligned} \tag{26}$$

where, for example, $E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}))$ is the abbreviation of

$$n^{-1}\sum_{a=1}^{n} \kappa_U^{ij}(\cdot, \hat{U}_a^{ij})$$

We now derive the order of magnitude of $\|\Delta_n^{(1)}\|_{\text{HS}}$. Because

$$\begin{aligned}
&E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij})) \\
&= E_n[(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))] \\
&\quad + E_n[(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})) \otimes \kappa_U^{ij}(\cdot, U^{ij})] \\
&\quad + E_n[\kappa_U^{ij}(\cdot, U^{ij}) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))],
\end{aligned}$$

we have, by the triangular inequality,

$$\begin{aligned}
\|\Delta_n^{(1)}\|_{\text{HS}} \leq{}& E_n\|(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))\|_{\text{HS}} \\
&+ 2E_n\|(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})) \otimes \kappa_U^{ij}(\cdot, U^{ij})\|_{\text{HS}}.
\end{aligned}$$

By Lemma 15, the right-hand side can be rewritten as

$$\begin{aligned}
&E_n\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2E_n(\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)} \|\kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)}).
\end{aligned} \tag{27}$$

Now applying Theorem 7, part (ii), with

$$\begin{aligned}
&\Omega = \Omega^{-(i,j)}, \quad \kappa_0 = \kappa_X^{-(i,j)}, \quad \mathscr{H}_0 = \mathscr{H}^{-(i,j)}(X), \\
&\mathbb{R}^d = \mathbb{R}^{d_{ij}}, \quad \kappa_1 = \kappa_U^{ij}, \quad \mathscr{H}_1 = \mathscr{H}^{ij}(U),
\end{aligned}$$

we see that, for some $C > 0$, (27) is bounded from above by

$$\begin{aligned}
&C^2\|\hat{U}^{ij} - U^{ij}\|^2_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}\; E_n\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) \\
&+ 2C\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \\
&\times E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})\kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\}.
\end{aligned} \tag{28}$$

By the weak law of large numbers,

$$
\begin{aligned}
&E_n \kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) = O_P(1) \\
&E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) \kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\} = O_P(1).
\end{aligned}
$$

Hence (28) is of the order $O_P(b_n^2)O_P(1) + O_P(b_n)O_P(1) = O_P(b_n)$.

Next, we derive the order of magnitude of $\|\Delta_n^{(2)}\|_{\mathrm{HS}}$, which can be rewritten as

$$
\begin{aligned}
&\|\{[E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij})) + E_n(\kappa_U^{ij}(\cdot, U^{ij}))] \\
&\qquad\qquad \otimes [E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij})) + E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\} \\
&\qquad\qquad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\mathrm{HS}} \\
&\le \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2\|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\mathscr{H}^{ij}(U)} \, \|E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\mathscr{H}^{ij}(U)} \\
&\le E_n\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2E_n\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)} \, E_n\|\kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)}.
\end{aligned}
$$

As shown in the proof of $\|\Delta_n^{(1)}\|_{\mathrm{HS}}$, this too is of the order $O_P(b_n^2)O_P(1) + O_P(b_n)O_P(1) = O_P(b_n)$.

(ii) and (iii): The proofs of (ii) and (iii) are essentially the same as the proof of (i), so we just highlight the proof of (ii) and omit the proof of (iii). Similar to (26), we have

$$
\|\hat{\Sigma}_{(X^i\hat{U}^{ij})\hat{U}^{ij}} - \hat{\Sigma}_{(X^iU^{ij})U^{ij}}\|_{\mathrm{HS}} \le \|\Delta_n^{(1)}\|_{\mathrm{HS}} + \|\Delta_n^{(2)}\|_{\mathrm{HS}},
$$

where

$$
\begin{aligned}
\Delta_n^{(1)} &= E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij})) \otimes \kappa_U^{ij}(\cdot, U^{ij})) \\
\Delta_n^{(2)} &= E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij}))) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\
&\quad - E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij})).
\end{aligned}
$$

Because

$$
\begin{aligned}
\Delta_n^{(1)} &= E_n[(\kappa_U^{ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_U^{ij}(\cdot, (X^i, U^{ij}))) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))] \\
&\quad + E_n[(\kappa_U^{ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_U^{ij}(\cdot, (X^i, U^{ij}))) \otimes \kappa_U^{ij}(\cdot, U^{ij})] \\
&\quad + E_n[\kappa_U^{ij}(\cdot, (X^i, U^{ij})) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))],
\end{aligned}
$$

we have

$$
\begin{aligned}
&\|\Delta_n^{(1)}\|_{\mathrm{HS}} \\
&\le E_n\|(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2E_n(\|(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))\|_{\mathscr{H}^{ij}(U)} \|\kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)}) \\
&\le C^2\|\hat{U}^{ij} - U^{ij}\|^2_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \, E_n\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) \\
&\qquad + 2C\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \\
&\qquad \times E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) \kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\}.
\end{aligned}
$$

The rest of the proof is the same as the corresponding part of the proof of part (i). □

Appendix E. Proof of Theorem 9

Denote

$$\begin{aligned}
&\hat{\Sigma}_{(X^i \hat{U}^{ij})\hat{U}^{ij}},\ (\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} + \delta_n I)^{-1},\ \hat{\Sigma}_{\hat{U}^{ij}(X^j \hat{U}^{ij})},\\
&\hat{\Sigma}_{(X^i U^{ij})U^{ij}},\ (\hat{\Sigma}_{U^{ij}U^{ij}} + \delta_n I)^{-1},\ \hat{\Sigma}_{U^{ij}(X^j U^{ij})},\\
&\Sigma_{(X^i U^{ij})U^{ij}},\ \Sigma^{-1}_{U^{ij}U^{ij}},\ \Sigma_{U^{ij}(X^j U^{ij})}
\end{aligned} \tag{29}$$

by $\hat{A},\ \hat{B},\ \hat{C},\ \tilde{A},\ \tilde{B},\ \tilde{C},\ A,\ B,\ C$, respectively. Then, by the definition of conjoined conditional covariance operator and the triangular inequality,

$$\begin{aligned}
&\|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}}\\
&\le \|\hat{\Sigma}_{(X^i \hat{U}^{ij})(X^j \hat{U}^{ij})} - \hat{\Sigma}_{(X^i U^{ij})(X^j U^{ij})}\|_{\mathrm{HS}} + \|\hat{A}\hat{B}\hat{C} - \tilde{A}\tilde{B}\tilde{C}\|_{\mathrm{HS}}
\end{aligned} \tag{30}$$

By Theorem 8, the first term is $O_P(b_n)$. The second term (without the norm) is

$$\begin{aligned}
\hat{A}\hat{B}\hat{C} - \tilde{A}\tilde{B}\tilde{C} &= (\hat{A} - \tilde{A})\hat{B}\hat{C} + \tilde{A}(\hat{B} - \tilde{B})\hat{C} + \tilde{A}\tilde{B}(\hat{C} - \tilde{C})\\
&= (\hat{A} - \tilde{A})\hat{B}\hat{C} + \tilde{A}\hat{B}(\tilde{B}^{-1} - \hat{B}^{-1})\tilde{B}\hat{C} + \tilde{A}\tilde{B}(\hat{C} - \tilde{C}).
\end{aligned} \tag{31}$$

Since, by Theorem 8,

$$\hat{A} - \tilde{A} = \ddot{O}_P(b_n), \quad \tilde{B}^{-1} - \hat{B}^{-1} = \ddot{O}_P(b_n), \quad \hat{C} - \tilde{C} = \ddot{O}_P(b_n),$$

in order for (31) to hold it suffices to show that

$$\hat{B}\hat{C} = \dot{O}_P(1), \quad \tilde{A}\hat{B} = \dot{O}_P(1), \quad \tilde{B}\hat{C} = \dot{O}_P(1), \quad \tilde{A}\tilde{B} = \dot{O}_P(1). \tag{32}$$

To simplify the notation, let

$$\check{B} = (\Sigma_{U^{ij}U^{ij}} + \delta_n I)^{-1}, \quad \hat{D} = \hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}}, \quad \tilde{D} = \hat{\Sigma}_{U^{ij}U^{ij}}, \quad D = \Sigma_{\hat{U}^{ij}\hat{U}^{ij}}.$$

For the first relation (32), by Theorem 8, $\hat{C} - \tilde{C} = \ddot{O}_P(b_n)$; by Lemma 20, $\tilde{C} - C = \ddot{O}_P(n^{-1/2})$. Hence

$$\begin{aligned}
\hat{B}\hat{C} &= \hat{B}(\hat{C} - \tilde{C}) + \hat{B}(\tilde{C} - C) + (\hat{B} - \tilde{B})C + (\tilde{B} - \check{B})C\\
&\qquad + (\check{B} - B)C + BC\\
&= \ddot{O}_P(\delta_n^{-1} b_n) + \ddot{O}_P(\delta_n^{-1} n^{-1/2}) + (\hat{B} - \tilde{B})C + (\tilde{B} - \check{B})C + \check{B}C.
\end{aligned} \tag{33}$$

The third term on the right is

$$\begin{aligned}
(\hat{B} - \tilde{B})C &= \hat{B}\ddot{O}_P(b_n)\tilde{B}C\\
&= \ddot{O}_P(\delta_n^{-1} b_n)(\tilde{B} - \check{B})C + \ddot{O}_P(\delta_n^{-1} b_n)\check{B}C\\
&= \ddot{O}_P(\delta_n^{-1} b_n)\tilde{B}\ddot{O}_P(n^{-1/2})\check{B}C + \ddot{O}_P(\delta_n^{-1} b_n)\check{B}C\\
&= \ddot{O}_P(\delta_n^{-1} b_n)\ddot{O}_P(\delta_n^{-1} n^{-1/2})\dot{O}_P(1) + \ddot{O}_P(\delta_n^{-1} b_n)\dot{O}_P(1).
\end{aligned}$$

So, by condition (b) and the fact that $n^{-1/2} \prec b_n$, this term is $\ddot{o}_P(1)$. The fourth term on the right-hand side of (33) is

$$(\tilde{B} - \check{B})C = \tilde{B}\ddot{O}_P(n^{-1/2})\check{B}C = \ddot{O}_P(\delta_n^{-1} b_n) = \ddot{o}_P(1).$$

By Lemma 23, $\check{B}C = \dot{O}_P(1)$. Hence the first relation in (32) holds. For later use, note that in this process we also proved

$$\tilde{B}C = \dot{O}_P(1), \text{ (and hence also) } A\tilde{B} = \dot{O}_P(1). \tag{34}$$

For the second relation in (32):

$$\tilde{A}\hat{B} = (\tilde{A} - A)\hat{B} + A(\hat{B} - \tilde{B}) + A\tilde{B}. \tag{35}$$

The first term is of the order $\ddot{O}_P(n^{-1/2}\delta_n^{-1})$. The second term is

$$A\tilde{B}\ddot{O}_P(b_n)\hat{B} = A\tilde{B}\ddot{O}_P(b_n)\dot{O}_P(\delta_n^{-1}) = \ddot{O}_P(b_n\delta_n^{-1}) = \ddot{O}_P(1),$$

where the second equality follows from (34), and the third from condition (b). The third term, again by (34), is of the order $\dot{O}_P(1)$. Thus the second relation in (32) holds.

For the third relation in (32):

$$\tilde{B}\hat{C} = \tilde{B}(\hat{C} - \tilde{C}) + \tilde{B}\tilde{C} = \ddot{O}_P(\delta_n^{-1} b_n) + \tilde{B}\tilde{C}. \tag{36}$$

Using an argument similar to the next step, we can show that

$$\tilde{B}\tilde{C} = \dot{O}_P(1). \tag{37}$$

Hence the third relation in (32) holds.

For the fourth relation in (32):

$$\tilde{A}\tilde{B} = (\tilde{A} - C)\tilde{B} + A\tilde{B} = \dot{O}_P(\delta_n^{-1})\dot{O}_P(\delta_n + n^{-1/2}) + A\tilde{B}.$$

By (34), the last term is of the order $\dot{O}_P(1)$. Thus the fourth relation in (32) holds.

Appendix F. Proof of Theorem 10

Lemma 25 *Suppose*

(a) the conditions in Corollary 22 are satisfied for $\alpha = 2$;

(b) $\|\hat{A}_1 - A_1\|_{\text{HS}} = O_P(n^{-1/2})$, $\|\hat{A}_2 - A_2\|_{\text{HS}} = O_P(n^{-1/2})$;

(c) $n^{-1} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.

Then

$$\begin{aligned}
&\|(\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1),\\
&\|A_1^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1),\\
&\|[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}]A_2\|_{\text{HS}} = O_P(\eta_n^{-1/2} n^{-1/2} + \eta_n),\\
&\|A_1^{-1/2} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\text{HS}} = O_P(n^{-1/2} + \epsilon_n).
\end{aligned} \tag{38}$$

Proof Let

$$\begin{aligned}
B_1 &= (\hat{A}_1 + \eta_n I)^{-1/2} - (A_1 + \eta_n I)^{-1/2}, \\
B_2 &= (A_1 + \eta_n I)^{-1/2} - A_1^{-1/2}, \\
B_3 &= A_1^{-1/2}, \\
C_1 &= (\hat{A}_3 + \epsilon_n I)^{-1} - (A_3 + \epsilon_n I)^{-1}, \\
C_2 &= (A_3 + \epsilon_n I)^{-1} - A_3^{-1}, \\
C_3 &= A_3^{-1}.
\end{aligned} \tag{39}$$

Then we can reexpress

$$\begin{aligned}
(\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} &= (B_1 + B_2 + B_3) A_2 (C_1 + C_2 + C_3) \\
&= \sum_{i=1}^{3} \sum_{j=1}^{3} B_i A_2 C_j.
\end{aligned} \tag{40}$$

We now analyze the nine terms in (40). By Lemma 24,

$$\begin{aligned}
B_1 &= \{(\hat{A}_1 + \eta_n I)^{-1/2} [(A_1 + \eta_n I)^{3/2} - (\hat{A}_1 + \eta_n I)^{3/2}] + (\hat{A}_1 - A_1)\} \\
&\quad \times (A_1 + \eta_n I)^{-3/2} \\
&= \dot{O}_P(\eta_n^{-1/2} n^{-1/2} + n^{-1/2})(A_1 + \eta_n I)^{-3/2} \\
&= \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2}.
\end{aligned}$$

Similarly,

$$B_2 = \{(A_1 + \eta_n I)^{-1/2} [A_1^{3/2} - (A_1 + \eta_n I)^{3/2}] + \eta_n I\} A_1^{-3/2}.$$

Since $(A_1 + \eta_n I)^{-1/2}$ commutes with $A_1^{3/2}$ and $(A_1 + \eta_n I)^{3/2}$, we have

$$B_2 = \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1/2} A_1^{-3/2}. \tag{41}$$

The terms C_1 and C_2 are

$$\begin{aligned}
C_1 &= (A_3 + \epsilon_n I)^{-1}(A_3 - \hat{A}_3)(\hat{A}_3 + \epsilon_n I)^{-1} = (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2} \epsilon_n^{-1}) \\
C_2 &= A_3^{-1}(A_3 + \epsilon_n I)^{-1}(-\epsilon_n I) = A_3^{-1}(A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n)
\end{aligned}$$

Hence

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2} \sum_{j=1}^{2} B_i A_2 C_j \\
&= \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2} A_2 (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2} \epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2} A_2 A_3^{-1} (A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n) \\
&+ \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1/2} A_1^{-3/2} A_2 (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2} \epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1/2} A_1^{-3/2} A_2 A_3^{-1} (A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n).
\end{aligned} \tag{42}$$

Because

$$A_1^{-3/2}A_2A_3^{-1}, \quad A_1^{-3/2}A_2A_3^{-2}, \quad A_1^{-2}A_2A_3^{-1}, \quad A_1^{-2}A_2A_3^{-2}$$

are finite-rank operators, by Lemma 21 and Corollary 22, the four operators in the middle of the four terms in (42) all have finite Hilbert-Schmidt norms which do not depend on n. Thus

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2}\sum_{j=1}^{2}B_iA_2C_j \\
&= \ddot{O}_P(\eta_n^{-1/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2}\epsilon_n + \eta_n n^{-1/2}\epsilon_n^{-1} + \eta_n\epsilon_n) = \ddot{o}_P(1),
\end{aligned} \tag{43}$$

where the last equality follows from condition (c). Let R be the indices of the rest of the terms except the last term: $R = \{(1,3),(2,3),(3,1),(3,2)\}$. Then

$$\textstyle\sum_{(i,j)\in R}B_iA_2C_j = \ddot{O}_P(\eta_n^{-1/2}n^{-1/2} + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = \ddot{o}_P(1), \tag{44}$$

where the last equality follows from condition (c). Thus we have

$$(\hat{A}_1 + \eta_n I)^{-1/2}A_2(\hat{A}_3 + \epsilon_n I)^{-1} = B_3A_2C_3 + o_P(1) = A_1^{-1/2}A_2A_3^{-1} + \ddot{o}_P(1),$$

which implies the first relation in (38). For the second relation in (38), we have, by (44),

$$\begin{aligned}
\|A_1^{-1/2}A_2(\hat{A}_3 + \epsilon_n I)^{-1}\|_{\mathrm{HS}} &\leq \|B_3A_2C_1\|_{\mathrm{HS}} + \|B_3A_2C_2\|_{\mathrm{HS}} + \|A_1^{-1/2}A_2A_3^{-1}\|_{\mathrm{HS}} \\
&= \|A_1^{-1/2}A_2A_3^{-1}\|_{\mathrm{HS}} + O_P(n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = O_P(1).
\end{aligned}$$

For the third relation in (38), we have

$$[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}]A_2 = B_1A_2 + B_2A_2.$$

Using Lemma 24, it is easy to see that

$$B_1A_2 = \dot{O}_P(\eta_n^{-1/2}n^{-1/2})(A_1 + \eta_n I)^{-3/2}A_2 = \dot{O}_P(\eta_n^{-1/2}n^{-1/2}),$$

where, for the last equality, we have used Lemma 21, which implies $\|(A_1 + \eta_n I)^{-3/2}A_2\|_{\mathrm{HS}} \leq \|A_1^{-3/2}A_2\|_{\mathrm{HS}}$. By (41),

$$\begin{aligned}
B_2A_2 &= \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1/2}A_1^{-3/2}A_2 \\
&= \dot{O}_P(\eta_n)A_1^{-3/2}(A_1 + \eta_n I)^{-1/2}A_2 \\
&= \dot{O}_P(\eta_n)A_1^{-2}A_2 = \ddot{O}_P(\eta_n).
\end{aligned}$$

Hence

$$[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}]A_2 = \ddot{O}_P(\eta_n^{-1/2}n^{-1/2} + \eta_n).$$

For the last relation in (38), we have

$$\begin{aligned}
&A_1^{-1/2}A_2[(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}]A_2^*A_1^{-1/2} \\
&= A_1^{-1/2}A_2A_3^{-1}[A_3 - \hat{A}_3 - \epsilon_n I)](\hat{A}_3 + \epsilon_n I)^{-1}A_2^*A_1^{-1/2} \\
&= A_1^{-1/2}A_2A_3^{-1}\dot{O}_P(n^{-1/2} + \epsilon_n)(\hat{A}_3 + \epsilon_n I)^{-1}A_2^*A_1^{-1/2}.
\end{aligned} \tag{45}$$

By the second relation in (38), $(\hat{A}_3 + \epsilon_n I)^{-1}A_2^*A_1^{-1/2} = \ddot{O}_P(1)$. Thus the last relation in (38) holds. □

Lemma 26 *Suppose*

(a) the conditions in Corollary 22 are satisfied for $\alpha = 1$*;*

(b) $\|\hat{A}_1 - A_1\|_{\rm OP} = O_P(n^{-1/2})$, $\|\hat{A}_2 - A_2\|_{\rm OP} = O_P(n^{-1/2})$*;*

(c) $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.

Then

$$\begin{aligned} &\|(\hat{A}_1 + \eta_n I)^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\rm HS} = O_P(1), \\ &\|A_1^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\rm HS} = O_P(1), \\ &\|[(\hat{A}_1 + \eta_n I)^{-1} - A_1^{-1}] A_2\|_{\rm HS} = O_P(\eta_n^{-1} n^{-1/2} + \eta_n), \\ &\|A_1^{-1} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\rm HS} = O_P(n^{-1/2} + \epsilon_n). \end{aligned} \tag{46}$$

Proof Reset B_1, B_2, and B_3 to

$$B_1 = (\hat{A}_1 + \eta_n I)^{-1} - (A_1 + \eta_n I)^{-1}, \quad B_2 = (A_1 + \eta_n I)^{-1} - A_1^{-1}, \quad B_3 = A_1^{-1},$$

and keep C_1, C_2, C_3 the same as before. Then

$$B_1 = \dot{O}_P(n^{-1/2} \eta_n^{-1})(A_1 + \eta_n I)^{-1}, \quad B_2 = \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1} A_1^{-1}.$$

Hence

$$\begin{aligned} &(\hat{A}_1 + \eta_n I)^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} \\ &= \textstyle\sum_{i=1}^{3} \sum_{j=1}^{3} B_i A_2 C_j \\ &= \ddot{O}_P(n^{-1/2} \eta_n^{-1} n^{-1/2} \epsilon_n^{-1} + n^{-1/2} \eta_n^{-1} \epsilon_n + n^{-1/2} \eta_n^{-1} \\ &\qquad + \eta_n n^{-1/2} \epsilon_n^{-1} + \eta_n \epsilon_n + \eta_n + n^{-1/2} \epsilon_n^{-1} + \epsilon_n) \\ &= \ddot{O}_P(n^{-1/2} \eta_n^{-1} + \eta_n + n^{-1/2} \epsilon_n^{-1} + \epsilon_n) \\ &= A_1^{-1} A_2 A_3 + \ddot{o}_P(1), \end{aligned}$$

where the last equality follows from condition (c). Hence the first relation in (46) holds. The second relation in (46) holds because

$$\begin{aligned} A_1^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} &= A_1^{-1} A_2 C_1 + A_1^{-1} A_2 C_2 + A_1^{-1} A_2 C_3 \\ &= \ddot{O}_P(n^{-1/2} \epsilon_n^{-1} + \epsilon_n) + A_1^{-1} A_2 A_3^{-1} = A_1^{-1} A_2 A_3^{-1} + \ddot{o}_P(1). \end{aligned}$$

The third relation in (46) holds because

$$[(\hat{A}_1 + \eta_n I)^{-1} - A_1^{-1}] A_2 = B_1 A_2 + B_2 A_2 = \ddot{O}_P(\eta_n^{-1} n^{-1/2} + \eta_n).$$

The proof of the fourth relation in (46) is similar to the derivation in (45). □

PROOF OF THEOREM 10. Note that, because $\mathrm{ran}(\Sigma_{XX}^{3/2}) \subseteq \mathrm{ran}(\Sigma_{XX})$, condition (b) implies that $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX}^{\alpha})$ is satisfied for both $\alpha = 1$ and $\alpha = 3/2$. Also, condition (d) is made to simplify the proof; it can be relaxed with a lengthier proof.

Denote the operators

$$(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_n I)^{-1}, \quad \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}, \quad (\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} + \eta_n I)^{-1/2}$$
$$\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}, \quad \Sigma_{X^{-(i,j)}X^{(i,j)}}, \quad \Sigma^{-1/2}_{X^{-(i,j)}X^{-(i,j)}}$$

by $\hat{A}$, $\hat{B}$, $\hat{C}$, A, B, C, respectively. In this notation, $\hat{f}^{ij}_1, \ldots, \hat{f}^{ij}_{d_{ij}}$ are the first d_{ij} eigenfunctions of the generalized eigenvalue problem

$$\begin{aligned} &\text{maximize} && \langle f, \hat{B}\hat{A}\hat{B}^* f\rangle_{-(i,j)} \\ &\text{subject to} && \langle f, \hat{C}^{-2} f\rangle_{-(i,j)} = 1, \quad \langle f, \hat{C}^{-2} f_r\rangle_{-(i,j)} = 0, \quad i = 1, \ldots, k-1. \end{aligned}$$

This means $\hat{f}^{ij}_r = \hat{C}\hat{\phi}^{ij}_r$, where $\hat{\phi}^{ij}_r$ is the rth eigenfunction of $\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C}$. We first derive the order of magnitude of the operator

$$\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C \tag{47}$$

in terms of the Hilbert Schmidt norm (another route is to derive this in terms of the operator norm, which is also sufficient for our purpose). By simple calculation,

$$\begin{aligned} \hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C =& \hat{C}(\hat{B} - B)\hat{A}(\hat{B} - B)^*\hat{C} + \hat{C}(\hat{B} - B)\hat{A}B^*\hat{C} \\ &+ \hat{C}B\hat{A}(\hat{B} - B)^*\hat{C} + \hat{C}B\hat{A}B^*\hat{C} - CBAB^*C. \end{aligned} \tag{48}$$

The reason for choosing this particular form of decomposition is to expose the finite-rank operator B, so that, for example, when combined with the operator C (an unbounded operator), BC is still a finite-rank operator. The Hilbert-Schmidt norm of the first term on the right is

$$\|\hat{C}(\hat{B} - B)\hat{A}(\hat{B} - B)^*\hat{C}\|_{\text{HS}} \le \|\hat{C}\|^2_{\text{OP}} \|\hat{B} - B\|^2_{\text{HS}} \|\hat{A}\|_{\text{OP}} = O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1}). \tag{49}$$

For the second term in (48), by Lemma 25, $\|\hat{C}B\hat{A}\|_{\text{HS}} = \|\hat{A}B^*\hat{C}\|_{\text{HS}} = O_P(1)$. Hence, by Lemmas 16 and 18,

$$\begin{aligned} \|\hat{C}(\hat{B} - B)\hat{A}B^*\hat{C}\|_{\text{HS}} =& \|\hat{C}B\hat{A}(\hat{B} - B)^*\hat{C}\|_{\text{HS}} \\ \le& \|\hat{C}\|_{\text{OP}}\, \|\hat{B} - B\|_{\text{HS}}\, \|\hat{A}B^*\hat{C}\|_{\text{HS}} = O_P(\eta_n^{-1/2}n^{-1/2}). \end{aligned} \tag{50}$$

The third term in (48) is the adjoint operator of the second term, so it has the same norm. The Hilbert-Schmidt norm of the last two terms in (48) is

$$\begin{aligned} &\|\hat{C}B\hat{A}B^*\hat{C} - CBAB^*C\|_{\text{HS}} \\ &\le \|\hat{C}B\hat{A}B^*(\hat{C} - C)\|_{\text{HS}} + \|(\hat{C} - C)B\hat{A}B^*C\|_{\text{HS}} \\ &\quad + \|CB(\hat{A} - A)B^*C\|_{\text{HS}} \\ &\le \|\hat{C}B\hat{A}\|_{\text{HS}}\, \|B^*(\hat{C} - C)\|_{\text{HS}} + \|(\hat{C} - C)B\|_{\text{HS}}\, \|\hat{A}B^*C\|_{\text{HS}} \\ &\quad + \|CB(\hat{A} - A)B^*C\|_{\text{HS}} \end{aligned} \tag{51}$$

By the first relation in (38), $\|\hat{C}B\hat{A}\|_{\text{HS}} = O_P(1)$; by the second, $\|\hat{A}B^*C\|_{\text{HS}} = O_P(1)$; by the third, $\|(\hat{C} - C)B\|_{\text{HS}} = O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n)$; by the fourth, $\|CB(\hat{A} - A)B^*C\|_{\text{HS}} = O_P(n^{-1/2} + \epsilon_n)$. Therefore,

$$\begin{aligned} \|\hat{C}B\hat{A}B^*\hat{C} - CBAB^*C\|_{\text{HS}} =& O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n + n^{-1/2} + \epsilon_n) \\ =& O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n). \end{aligned} \tag{52}$$

Combining (49), (50), and (52), we have

$$\|\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C\|_{\mathrm{HS}} = O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n).$$

Next, recall that $(\hat{\lambda}_1^{ij}, \hat{\phi}_1^{ij}), \ldots, (\hat{\lambda}_{d_{ij}}^{ij}, \hat{\phi}_{d_{ij}}^{ij})$ are the first d_{ij} pairs of eigenvalue and eigenfunction of $\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C}$, and let $(\lambda_1^{ij}, \phi_1^{ij}), \ldots, (\lambda_{d_{ij}}^{ij}, \phi_{d_{ij}}^{ij})$ be the first d_{ij} eigenvalue-eigenfunction pairs of $CBAB^*C$. By perturbation theory of linear operators, $|\hat{\lambda}_r^{ij} - \lambda_r^{ij}|$ is of the same order of magnitude as $\|\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C\|_{\mathrm{HS}}$, and, if condition (d) holds, then

$$\|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{(i,j)}(X)}$$

also has the same order of magnitude. That is, for each $r = 1, \ldots, d_{ij}$,

$$\begin{aligned} \hat{\lambda}_r^{ij} - \lambda_r^{ij} &= O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n) \\ \|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{(i,j)}(X)} &= O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n). \end{aligned} \tag{53}$$

By construction,

$$\hat{f}_r^{ij} = \hat{C}\hat{\phi}_r = \hat{\lambda}_r^{ij}\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\hat{\phi}_r^{ij}. \tag{54}$$

We now derive the order of magnitude of

$$\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C\|_{\mathrm{HS}}.$$

Similar to (48),

$$\begin{aligned} &\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C \\ &= \hat{C}^2(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}^2(\hat{B}-B)\hat{A}B^*\hat{C} \\ &\quad + \hat{C}^2B\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}^2B\hat{A}B^*\hat{C} - CBAB^*C \end{aligned}$$

where the first term on the right, similar to (49), is of the order

$$\|\hat{C}^2(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C}\|_{\mathrm{HS}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1}). \tag{55}$$

By the first relation in (38), $\|\hat{A}B^*\hat{C}\|_{\mathrm{HS}} = O_P(1)$, and hence

$$\begin{aligned} \|\hat{C}^2(\hat{B}-B)\hat{A}B^*\hat{C}\|_{\mathrm{HS}} &\le \|\hat{C}^2\|_{\mathrm{OP}}\,\|\hat{B}-B\|_{\mathrm{HS}}\,\|\hat{A}B^*\hat{C}\|_{\mathrm{HS}} \\ &= O_P(\eta_n^{-1}n^{-1/2}). \end{aligned} \tag{56}$$

By the first relation in (46), $\|\hat{C}^2B^*\hat{A}\|_{\mathrm{HS}} = O_P(1)$, and hence

$$\begin{aligned} \|\hat{C}^2B\hat{A}(\hat{B}-B)^*\hat{C}\|_{\mathrm{HS}} &\le \|\hat{C}^2B\hat{A}\|_{\mathrm{HS}}\,\|\hat{B}-B\|_{\mathrm{HS}}\|\hat{C}\|_{\mathrm{OP}} \\ &= O_P(n^{-1/2}\eta_n^{-1/2}). \end{aligned} \tag{57}$$

Similar to (51), we have

$$\begin{aligned} &\|\hat{C}^2B\hat{A}B^*\hat{C} - C^2BAB^*C\|_{\mathrm{HS}} \\ &\le \|\hat{C}^2B\hat{A}\|_{\mathrm{HS}}\,\|B^*(\hat{C}-C)\|_{\mathrm{HS}} \\ &\quad + \|(\hat{C}^2 - C^2)B\|_{\mathrm{HS}}\,\|\hat{A}B^*C\|_{\mathrm{HS}} + \|C^2B(\hat{A}-A)B^*C\|_{\mathrm{HS}} \end{aligned}$$

Applying Lemma 25 and Lemma 26 to the right-hand side above, we obtain

$$\begin{aligned}&\|\hat{C}^2 B\hat{A}B^*\hat{C} - C^2BAB^*C\|_{\mathrm{HS}} \\ &= O_P(1)O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n) + O_P(\eta_n^{-1}n^{-1/2} + \eta_n) \\ &\quad + O_P(n^{-1/2} + \epsilon_n) \\ &= O_P(\eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n)\end{aligned} \tag{58}$$

Combining (55) through (58), we have

$$\begin{aligned}&\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C \\ &= \ddot{O}_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{-1/2}n^{-1/2} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n) \\ &= \ddot{O}_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n).\end{aligned} \tag{59}$$

Finally, let us derive the convergence rate of $\hat{U}^{ij}$. By (54), we have

$$\begin{aligned}\hat{f}_r^{ij} - f_r^{ij} &= (\hat{\lambda}_r^{ij} - \lambda_r^{ij})\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\hat{\phi}_r^{ij} \\ &\quad + \lambda_r^{ij}(\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C)\hat{\phi}_r^{ij} + \lambda_r^{ij}C^2BAB^*C(\hat{\phi}_r^{ij} - \phi_r^{ij}).\end{aligned}$$

Hence

$$\begin{aligned}\|\hat{f}_r^{ij} - f_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \leq{}& |\hat{\lambda}_r^{ij} - \lambda_r^{ij}|\,\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\|_{\mathrm{OP}}\,\|\hat{\phi}_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \\ &+ \lambda_r^{ij}\,\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C\|_{\mathrm{OP}}\,\|\hat{\phi}_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \\ &+ \lambda_r^{ij}\|C^2BAB^*C\|_{\mathrm{OP}}\,\|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)}.\end{aligned}$$

By (53) and (59), the right-hand side is of the order

$$\begin{aligned}&O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n) \\ &+ O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n) \\ &+ O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n + \epsilon_n) \\ &= O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n)\end{aligned} \tag{60}$$

Because

$$\|\hat{U}_{ij} - U_{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = (\textstyle\sum_{r=1}^{d_{ij}}\|\hat{f}_r^{ij} - f_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)})^{1/2},$$

$\|\hat{U}_{ij} - U_{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}$ has the same order of magnitude as (60). □

Appendix G. Proof of Theorem 11

Using the notation defined in (29), we have

$$\begin{aligned}&\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} \\ &= \|\hat{\Sigma}_{(X^iU^{ij})(X^jU^{ij})} - \Sigma_{(X^iU^{ij})(X^jU^{ij})}\|_{\mathrm{HS}} + \|\tilde{A}\tilde{B}\tilde{C} - ABC\|_{\mathrm{HS}}\end{aligned} \tag{61}$$

By Lemma 20, the first term is of the order $O_P(n^{-1/2})$. Similar to (31),

$$\tilde{A}\tilde{B}\tilde{C} - ABC = (\tilde{A} - A)\tilde{B}\tilde{C} + A(\tilde{B} - B)\tilde{C} + AB(\tilde{C} - C). \tag{62}$$

By Lemma 20, $\tilde{A} - A = \ddot{O}_P(n^{-1/2})$; by (37), $\tilde{B}\tilde{C} = \dot{O}_P(1)$. Hence the first term is of the order $\ddot{O}_P(n^{-1/2})$. The second term is

$$A(\tilde{B} - B)\tilde{C} = AB(B^{-1} - \tilde{B}^{-1})\tilde{B}\tilde{C} = \ddot{O}_P(\delta_n + n^{-1/2}). \tag{63}$$

It is easy to see that the third term on the right-hand side of (62) is also of the order $\ddot{O}_P(\delta_n + n^{-1/2})$. Hence

$$\tilde{A}\tilde{B}\tilde{C} - ABC = \ddot{O}_P(n^{-1/2} + \delta_n) = \ddot{O}_P(\delta_n),$$

where the last equality holds because $n^{-1/2} \prec b_n \preceq \delta_n$.

Appendix H. Proof of Theorem 14

When ϵ_n, η_n, and ϵ_n take the given form, the convergence rate in (17) becomes

$$b_n \asymp n^{3b/2+a-1} + n^{b-1/2} + n^{-b} + n^{-a} \asymp \max(n^{3b/2+a-1}, n^{b-1/2}, n^{-b}, n^{-a})$$

We need to minimize b_n over the set

$$C = \{(a,b) :\ a < \tfrac{1}{2},\ b < \tfrac{1}{2},\ \tfrac{3b}{2} + a - 1 < 0\}.$$

Equivalently, we need to minimize

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, b - \tfrac{1}{2}, -b, -a)$$

over C. Our strategy is to minimize $f(a,b)$ over $(0, \frac{1}{2}) \times (0, \frac{1}{2})$ and then check the minimizers (there are more than one) belong to C.

Because $b - \frac{1}{2} \geq -b$ iff $b \geq \frac{1}{4}$, we have

$$f(a,b) = \begin{cases} \max(\frac{3b}{2} + a - 1, b - \frac{1}{2}, -a) & b \geq \frac{1}{4} \\ \max(\frac{3b}{2} + a - 1, -b, -a) & b < \frac{1}{4} \end{cases}$$

Furthermore, for $b \geq \frac{1}{4}$,

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, b - \tfrac{1}{2}, -a) = \begin{cases} -a & a \in (0, \frac{1}{2} - b) \\ b - \frac{1}{2} & a \in [\frac{1}{2} - b, \frac{1}{2} - \frac{b}{2}] \\ \frac{3}{2}b + a - 1 & a \in (\frac{1}{2} - \frac{b}{2}, \frac{1}{2}) \end{cases}$$

which implies

$$\min_{0<a<\frac{1}{2}} f(a,b) = b - \tfrac{1}{2} \Rightarrow \min_{b \geq \frac{1}{2}} \min_{0<a<\frac{1}{2}} f(a,b) = \tfrac{1}{3} - \tfrac{1}{2} = -\tfrac{1}{6}$$

For $b < \frac{1}{4}$,

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, -b, -a) = \begin{cases} -a & a \in (0, b) \\ -b & a \in [b, \frac{1}{2}) \end{cases}$$

which implies

$$\min_{0<a<\frac{1}{2}} f(a,b) = -b \Rightarrow f(a,b) > -\tfrac{1}{4} \text{ for all } b \in (0, \tfrac{1}{4}),\ a \in (0, \tfrac{1}{2}).$$

Thus $f(a,b)$ reaches its minimum $-\frac{1}{4}$ when $b = \frac{1}{4}$, $a \in [\frac{1}{2} - b, \frac{1}{2} - \frac{b}{2}] = [\frac{1}{4}, \frac{3}{8}]$. Finally, it is easy to check that this set is contained in C.

Appendix I. Asymptotic analysis under high-dimensional setting

In this section we consider the scenario where the dimension of p_n of X goes to infinity with n. This asymptotic regime is significantly different from the fixed-p case, because, under some conditions, the unscaled covariance operator Σ_{UV} in (18) tends to 0 as $p_n \to \infty$. In this case, the convergence rate of $\|\hat{\Sigma}_{UV} - \Sigma_{UV}\|_{\text{HS}}$ is no longer meaningful unless it is compared with the magnitude of Σ_{UV}.

Specifically, suppose we use the Gaussian radial basis function kernel, and let $U_1, \ldots, U_n$ be an i.i.d. sample of a generic random vector $U \in \mathbb{R}^{pn}$. Commonly used choices of the shape parameter γ in the radial basis function (11) are based on some types of the center point of the distances

$$\{\|U_i - U_j\| : i, j = 1, \ldots, n, i \neq j\}.$$

For example, the default choice of γ in Kernlab (Karatzoglou et al., 2004) is $\gamma = 1/\tau^2$ where τ is a number between the 10th and 90th percentiles of the above set; Fukumizu et al. (2009) uses the median of the above set, whereas Lee et al. (2013) takes τ^2 to be the average of $\|U_i - U_j\|^2$. At the population level, the choice of γ of Lee et al. (2013) amounts to taking τ^2 to be $E\|U - \tilde{U}\|^2$, where $\tilde{U}$ an independent copy of U.

To make further progress possible we need to give a prototype on the dependence structure of U, the kernel, and its tuning parameter, which we summarize in the following assumption.

Assumption 7 *We make the following assumption for the $p_n \to \infty$ asymptotic regime:*

1. *(sparsity) There is a subset A_{p_n} of $\{1, \ldots, p_n\}$ such that*
   - *(a) $\{X^i : i \in A_{p_n}\}$ are independent;*
   - *(b) $\{X^i : i \in A_{p_n}\} \perp\!\!\!\perp \{X^i : i \in A_{p_n}^c\}$;*
   - *(c) The cardinality of $A_{p_n}^c$ is bounded as $p_n \to \infty$.*
2. *(kernel) For any subvector U of X, the kernel κ_U is the Gaussian radial basis function with shape parameter $\gamma = 1/\tau^2$, and τ^2 proportional to $E\|U - \tilde{U}\|^2$, where $\tilde{U}$ is an independent copy of U.*
3. *(identical marginal distributions) $X^1, \ldots, X^{pn}$ are identically distributed with variance σ^2 and finite fourth moment.*

The first assumption is a sparsity assumption: it implies that the number of edges does not diverge to infinity. While this is by no means the only possible scenario under which we can obtain asymptotic result similar to that given below, it helps us to gauge the magnitude of $E\|U - \tilde{U}\|^2$ with minimal complication. The third assumption is a simplifying assumption: without it the theory still holds but the notation for the proofs will be more complicated. The second assumption can also be relaxed to non-Gaussian radial basis function's.

In addition to the above structural assumption, we also need the following technical assumption.

Assumption 8 *Let U and V represent $X^{-(i,j)}$ and $X^{(i,j)}$, respectively, $(\tilde{U}, \tilde{V})$ an independent copy of (U, V), and $\bar{V}$ a copy of V that is independent of $(U, V, \tilde{U}, \tilde{V})$. Let U^t, V^t and so on be the t-th component of U, V. Let $B_{ij} = \{1, \ldots, p_n\} \setminus \{i, j\}$ and let*

$$\begin{aligned} S_{p_n} &= (p_n - 2)^{-1} \textstyle\sum_{t \in B_{ij}} [(U^t - \tilde{U}^t)^2 - 2\sigma^2], \\ T_{p_n} &= (p_n - 2)^{-1} \textstyle\sum_{t \in B_{ij}} [(U^t - \bar{U}^t)^2 - 2\sigma^2]. \end{aligned} \tag{64}$$

We assume that the sequences

$$\begin{aligned} &\{e^{-S_{p_n}} (\sqrt{p_n} S_{p_n})^2 : n = 1, 2, \ldots\}, \\ &\{e^{-S_{p_n}} [\sqrt{p_n}(S_{p_n} + T_{p_n})]^2 : n = 1, 2, \ldots\} \end{aligned}$$

are uniformly integrable.

As will be clear in the proof of Theorem 27, the quantities

$$e^{-S_{p_n}} (\sqrt{p_n} S_{p_n})^2 \quad \text{and} \quad e^{-S_{p_n}} [\sqrt{p_n}(S_{p_n} + T_{p_n})]^2$$

are of the order $O_P(1)$ as $p_n \to \infty$. This assumption is used to guarantee the boundedness of certain expectation sequences.

The next theorem shows that $\|\Sigma_{UU}\|_{\text{HS}}$ and $\|\Sigma_{UV}\|_{\text{HS}}$ are of the order $O(p_n^{-1/2})$.

Theorem 27 *Under Assumptions 7 and 8 we have, as $p_n \to \infty$,*

$$(a) \quad \|\Sigma_{X^{-(i,j)} X^{-(i,j)}}\|_{\text{HS}} \asymp p_n^{-1/2}, \quad (b) \quad \|\Sigma_{X^{(i,j)} X^{-(i,j)}}\|_{\text{HS}} \asymp p_n^{-1/2}.$$

Before proving this theorem, we first prove three lemmas. For convenience, we abbreviate p_n by p, keeping in mind that it goes to infinity with the sample size.

Lemma 28 *Suppose Assumption 7 holds and $\tilde{X}$ is an independent copy of X. Then, for any $i, j \in \{1, \ldots, p\}$,*

$$\begin{aligned} &(a) \quad E(\|X^{-(i,j)} - \tilde{X}^{-(i,j)}\|^2) \asymp p \\ &(b) \quad E[(\gamma \|X^{-(i,j)} - \tilde{X}^{-(i,j)}\|^2 - 2\sigma^2)^2] \asymp p^{-1}. \end{aligned}$$

Proof To prove the first relation, let W^t represent the t-th component of the $(p-2)$-vector $X^{-(i,j)} - \tilde{X}^{-(i,j)}$, $t \in \{1, \ldots, p\} \setminus \{i, j\} \equiv B_{ij}$. Then

$$E(\|X^{-(i,j)} - \tilde{X}^{-(i,j)}\|^2) = \sum_{t \in B_{ij}} E[(W^t)^2] = (p - 2) 2\sigma^2 \asymp p.$$

Thus the first relation holds.

To prove the second relation we take $\gamma \asymp p^{-1}$ – without loss of generality, take $\gamma = (p-2)^{-1}$. Then,

$$\gamma \| X^{-(i,j)} - \tilde{X}^{-(i,j)} \|^2 - 2\sigma^2 = (p-2)^{-1} \sum_{t \in B_{ij}} [(W^t)^2 - 2\sigma^2].$$

Let $H^t = (W^t)^2 - 2\sigma^2$. Then

$$\begin{aligned} & E[(\gamma \| X^{-(i,j)} - \tilde{X}^{-(i,j)} \|^2 - 2\sigma^2)^2] \\ & = (p-2)^{-2} \sum_{t,s \in B_{ij}} E(H^s H^t) \\ & = (p-2)^{-2} \sum_{t,s \in B_{ij} \cap A_p} E(H^s H^t) + (p-2)^{-2} \sum_{t,s \in B_{ij} \cap A_p^c} E(H^s H^t). \end{aligned}$$

Since the cardinality of A_p^c is bounded, the second term on the right is of the order $(p-2)^{-2} O(1) = O(p^{-2})$. Because $\{H^t : t \in A_p\}$ are i.i.d., and $\text{card}(A_p) \asymp p$, the first term on the right-hand side is

$$(p-2)^{-2} \sum_{t,s \in B_{ij} \cap A_p} E[(H^t)^2] \asymp (p-2)^{-2} E[(H^t)^2] \text{card}(A_p) \asymp p^{-1},$$

which proves the second relation. □

In the following, if two random elements A and B have the same distribution, then we write $A \overset{D}{=} B$.

Lemma 29 *Suppose U and V are random vectors and κ_1 and κ_2 their respective kernels. Then*

$$\begin{aligned} \|\Sigma_{UV}\|_{\text{HS}}^2 = & E[\kappa_1(U, \tilde{U}) \kappa_2(V, \tilde{V})] - 2E[\kappa_1(U, \tilde{U}) \kappa_2(V, \bar{V})] \\ & + E[\kappa_1(U, \tilde{U})] E[\kappa_2(V, \tilde{V})]. \end{aligned} \tag{65}$$

where $(\tilde{U}, \tilde{V}) \perp\!\!\!\perp (U, V) \perp\!\!\!\perp \bar{V}$, $(\tilde{U}, \tilde{V}) \overset{D}{=} (U, V)$, and $\bar{V} \overset{D}{=} V$.

Proof By definition,

$$\Sigma_{UV} = E[(\kappa_1(\cdot, U) - \mu_U) \otimes (\kappa_2(\cdot, V) - \mu_V)].$$

Let $F = \kappa_1(\cdot, U) - \mu_U$ and $G = \kappa_2(\cdot, V) - \mu_V$, and let $(\tilde{F}, \tilde{G})$ denote the counterpart of (F, G) with U and V replaced by $\tilde{U}$ and $\tilde{V}$. Then

$$\begin{aligned} \|\Sigma_{UV}\|_{\text{HS}}^2 & = \langle E(F \otimes G), E(F \otimes G) \rangle_{\text{HS}} \\ & = E \langle F \otimes G, \tilde{F} \otimes \tilde{G} \rangle_{\text{HS}} \\ & = E(\langle F, \tilde{F} \rangle_{\mathscr{H}_U} \langle G, \tilde{G} \rangle_{\mathscr{H}_V}). \end{aligned}$$

Note that

$$\begin{aligned} \langle F, \tilde{F} \rangle_{\mathscr{H}_U} & = \kappa_1(U, \tilde{U}) - \langle \kappa_1(\cdot, U), \mu_U \rangle_{\mathscr{H}_U} - \langle \mu_U, \kappa_1(\cdot, \tilde{U}) \rangle_{\mathscr{H}_U} + \langle \mu_U, \mu_U \rangle_{\mathscr{H}_U} \\ & \equiv \kappa_1(U, \tilde{U}) - \tau_1(U) - \tau_1(\tilde{U}) + c_U, \end{aligned}$$

where $\tau_1(u) = \langle \kappa(\cdot, u), \mu_U \rangle_{\mathscr{H}_U}$ and $c_U = \langle \mu_U, \mu_U \rangle_{\mathscr{H}_U}$. Similarly,

$$\langle G, \tilde{G} \rangle_{\mathscr{H}_V} = \kappa_2(V, \tilde{V}) - \tau_2(V) - \tau_2(\tilde{V}) + c_V.$$

So

$$\|\Sigma_{UV}\|_{\mathrm{HS}}^2 = E\{[\kappa_1(U, \tilde{U}) - \tau_1(U) - \tau_1(\tilde{U}) + c_U][\kappa_2(V, \tilde{V}) - \tau_2(V) - \tau_2(\tilde{V}) + c_V]\}.$$

Put $\lambda_1(U, \tilde{U}) = \tau_1(U) + \tau_1(\tilde{U})$ and $\lambda_2(V, \tilde{V}) = \tau_2(V) + \tau_2(\tilde{V})$. Note that

$$E\kappa_1(U, \tilde{U}) = c_U,\ E\kappa_2(V, \tilde{V}) = c_V,\ E\lambda_1(U, \tilde{U}) = 2c_U,\ E\lambda_2(V, \tilde{V}) = 2c_V.$$

Use these relations to make the decomposition

$$\begin{aligned}
\|\Sigma_{UV}\|_{\mathrm{HS}}^2 &= E\{[\kappa_1(U, \tilde{U}) - \lambda_1(U, \tilde{U}) + c_U][\kappa_2(V, \tilde{V}) - \lambda_2(V, \tilde{V}) + c_V]\} \\
&= E[\kappa_1(U, \tilde{U})\kappa_2(V, \tilde{V})] - E[\kappa_1(U, \tilde{U})\lambda_2(V, \tilde{V})] \\
&\quad - E[\lambda_1(U, \tilde{U})\kappa_2(V, \tilde{V})] + E[\lambda_1(U, \tilde{U})\lambda_2(V, \tilde{V})] - c_U c_V.
\end{aligned} \tag{66}$$

The term $E[\lambda_1(U, \tilde{U})\lambda_2(V, \tilde{V})]$ on the right-hand side is

$$\begin{aligned}
E[(\tau_1(U) + \tau_1(\tilde{U}))(\tau_2(V) + \tau_2(\tilde{V}))] &= 2E[\tau_1(U)\tau_2(V)] + 2E[\tau_1(U)]E[\tau_2(V)] \\
&= 2E[\tau_1(U)\tau_2(V)] + 2c_U c_V
\end{aligned}$$

Furthermore,

$$E[\tau_1(U)\tau_2(V)] = E[\langle \kappa_1(\cdot, U), \mu_U \rangle_{\mathscr{H}_U} \langle \kappa_2(\cdot, V), \mu_V \rangle_{\mathscr{H}_V}] = E[\kappa_1(U, \tilde{U})\kappa_2(V, \bar{V})].$$

The term $E[\kappa_1(U, \tilde{U})\lambda_2(V, \tilde{V})]$ on the right-hand side of (66) is

$$\begin{aligned}
E[\kappa_1(U, \tilde{U})\lambda_2(V, \tilde{V})] &= E[\kappa_1(U, \tilde{U})(\tau_2(V) + \tau_2(\tilde{V})] \\
&= 2E[\kappa_1(U, \tilde{U})\tau_2(V)] = 2E[\kappa_1(U, \tilde{U})\kappa_2(V, \bar{V})]
\end{aligned}$$

Thus we have the desired equality. as desired. □

In the next lemma and theorem, $U, V, \tilde{U}, \tilde{V}, \bar{V}$ are random vectors defined in Assumption 8, and κ_1 and κ_2 are the Gaussian radial basis function kernels for U and V with shape parameters γ_1 and γ_2.

Lemma 30 *If Assumptions 7 and 8 are satisfied, then*

$$\begin{aligned}
&(a) \quad E[\kappa_1(U, \tilde{U})] - e^{2\sigma^2} \asymp p^{-1} \\
&(b) \quad E[\kappa_1(U, \tilde{U})\kappa_1(U, \bar{U})] - e^{4\sigma^2} \asymp p^{-1} \\
&(c) \quad E[\kappa_1(U, \tilde{U})^2] - e^{4\sigma^2} \asymp p^{-1} \\
&(d) \quad E\{[\kappa_1(U, \tilde{U}) - e^{2\sigma^2}]\kappa_2(V, \tilde{V})\} \asymp p^{-1} \\
&(e) \quad E\{[\kappa_1(U, \tilde{U}) - e^{2\sigma^2}]\kappa_2(V, \bar{V})\} \asymp p^{-1}
\end{aligned} \tag{67}$$

Proof Proof of (a), (b), and (c). Using the notation in Assumption 8, we have

$$E[\kappa_1(U,\tilde{U})] = e^{2\sigma^2} e^{S_p}.$$

By Chebychev's inequality and part (b) of Lemma 28, $S_p \xrightarrow{P} 0$. By Skorohod's representation theorem, there is a sequence $\tilde{S}_p$ such that $\tilde{S}_p \overset{D}{=} S_p$ and $\tilde{S}_p \to 0$ almost surely. By Taylor expansion,

$$E(e^{-S_p}) = E(e^{-\tilde{S}_p}) = 1 - E(\tilde{S}_p) + E(e^{\xi_p}\tilde{S}_p^2)/2, \tag{68}$$

where ξ_p is a random number between 0 and $-\tilde{S}_p$. This means, if $\tilde{S}_p > 0$, then $\xi_p \leq 0$, and if $\tilde{S}_p \leq 0$, $\xi_p \leq -\tilde{S}_p$. Consequently, $e^{\xi_p} \leq 1 + e^{-\tilde{S}_p}$. Also, by construction $E(\tilde{S}_p) = 0$. Let $c > 0$. Then, by (68) and the above discussion,

$$\begin{aligned} |E(e^{-S_p}) - 1| &\leq E(\tilde{S}_p^2)/2 + E(e^{-\tilde{S}_p}\tilde{S}_p^2)/2 \\ &\leq E(\tilde{S}_p^2)/2 + e^c E(\tilde{S}_p^2)/2 + E[e^{-\tilde{S}_p}\tilde{S}_p^2 I(|S_p| > c)]/2 \\ &\asymp p^{-1} + E[e^{-\tilde{S}_p}\tilde{S}_p^2 I(|S_p| > c)]/2, \end{aligned}$$

where the third line follows from Lemma 28. Since $pe^{-\tilde{S}_p}\tilde{S}_p^2 I(|S_p| > c)$ converges to 0 almost surely, and the sequence is uniformly integrable, we have

$$pE[e^{-\tilde{S}_p}\tilde{S}_p^2 I(|S_p| > c)] \to 0 \;\Rightarrow\; E[e^{-\tilde{S}_p}\tilde{S}_p^2 I(|S_p| > c)] = o(p^{-1}).$$

This proves (a) in (67). The proofs of (b) and (c) are similar.

Proof of (d) and (e). By Taylor expansion,

$$\begin{aligned} E[(\kappa_1(U,\tilde{U}) - e^{2\sigma^2})\kappa_2(V,\tilde{V})] &= E[e^{2\sigma^2}(e^{-S_p} - 1)\kappa_2(V,\tilde{V})] \\ &= E[(-S_p + e^{\xi_p}S_p^2/2)\kappa_2(V,\tilde{V})] \\ &= -E[S_p\kappa_2(V,\tilde{V})] + E[e^{\xi_p}S_p^2\kappa_2(V,\tilde{V})]/2, \end{aligned}$$

where ξ_p is a number between 0 and $-S_p$. The first term on the right is

$$E[S_p\kappa_2(V,\tilde{V})] = (p-2)^{-1}\textstyle\sum_{t\in B_{ij}} E\{[(U^t - \tilde{U}^t)^2 - 2\sigma^2]\kappa_2(V,\tilde{V})\}$$

Note that the components $U - \tilde{U}$ in A_p are independent of $(V,\tilde{V})$ regardless of the positions of $\{i,j\}$. Furthermore, $B_{ij} \cap A_p^c$ has bounded number of terms. Hence, in the decomposition

$$\begin{aligned} &\textstyle\sum_{t\in B_{ij}\cap A_p} E\{[(U^t - \tilde{U}^t)^2 - 2\sigma^2]\kappa_2(V,\tilde{V})\} \\ &\quad + \textstyle\sum_{t\in B_{ij}\cap A_p^c} E\{[(U^t - \tilde{U}^t)^2 - 2\sigma^2]\kappa_2(V,\tilde{V})\}, \end{aligned}$$

the first sum is 0; the second sum, which has bounded number of terms, is of the order $O(1)$. Therefore

$$E[S_p\kappa_2(V,\tilde{V})] = O(p^{-1}).$$

Also, because κ_2 is bounded by 1, we have

$$E[e^{\xi_p}S_p^2\kappa_2(V,\tilde{V})] \leq E(e^{\xi_p}S_p^2).$$

As was already shown, the right-hand side $\asymp p^{-1}$. This proves (d). Relation (e) can be prove similarly. □

Proof of Theorem 27 Proof of (a). By Lemma 28, $E(\|U-\tilde{U}\|^2) \asymp p^{-1}$. So, without loss of generality, we take $\gamma_1 = (p-2)^{-1}$. By Lemma 29,

$$\|\Sigma_{UU}\|^2_{\text{HS}} = E[\kappa_1(U,\tilde{U})^2] - 2E[\kappa_1(U,\tilde{U})\kappa_1(U,\bar{U})] + E^2[\kappa_1(U,\tilde{U})], \tag{69}$$

Using the notation in Assumption 8, we have

$$E[\kappa_1(U,\tilde{U})] = e^{2\sigma^2} E(e^{-S_p}) = e^{2\sigma^2}[1+1-E(e^{-S_p})].$$

Applying (a), (b), and (c) of Lemma 30, we have

$$\begin{aligned}\|\Sigma_{UU}\|^2_{\text{HS}} = &\{E[\kappa_1(U,\tilde{U})^2] - e^{4\sigma^2} + e^{4\sigma^2}\} \\ &- 2\{E[\kappa_1(U,\tilde{U})\kappa_1(U,\bar{U})] - e^{4\sigma^2} + e^{4\sigma^2}\} \\ &+ \{E[\kappa_1(U,\tilde{U})] - e^{-2\sigma^2} + e^{2\sigma^2}\}^2 \asymp p^{-1},\end{aligned}$$

which proves (a).

Proof of (b). Since the dimension of V is 2, we have $E(\|V-\tilde{V}\|^2) \asymp 1$. Thus, without loss of generality, we take $\gamma_1 = (p-2)^{-1}$, $\gamma_2 = 1$. By Lemma 29,

$$\begin{aligned}\|\Sigma_{UV}\|^2_{\text{HS}} = &E[\kappa_1(U,\tilde{U})\kappa_2(V,\tilde{V})] \\ &- 2E[\kappa_1(U,\tilde{U})\kappa_2(V,\bar{V})] + E[\kappa_1(U,\tilde{U})]E[\kappa_2(V,\tilde{V})].\end{aligned}$$

Applying (a), (d) and (e) of Lemma 30, we have

$$\begin{aligned}\|\Sigma_{UV}\|^2_{\text{HS}} = &E\{[\kappa_1(U,\tilde{U}) - e^{2\sigma^2}]\kappa_2(V,\tilde{V})\} + e^{2\sigma^2}E[\kappa_2(V,\tilde{V})] \\ &- 2E\{[\kappa_1(U,\tilde{U}) - e^{2\sigma^2}]\kappa_2(V,\bar{V})\} - 2e^{2\sigma^2}E[\kappa_2(V,\bar{V})] \\ &+ E[\kappa_1(U,\tilde{U}) - e^{2\sigma^2}]E[\kappa_2(V,\tilde{V})] + e^{2\sigma^2}E[\kappa_2(V,\tilde{V})] \asymp p^{-1},\end{aligned}$$

which completes the proof. □

Theorem 27 motivates us to redefine the covariance operator and its sample estimate as

$$\begin{aligned}\Sigma_{UV}(p_n) = &p_n^{1/2}E[(\kappa(\cdot,U)-\mu_U)\otimes(\kappa(\cdot,V)-\mu_V)], \\ \hat{\Sigma}_{UV}(p_n) = &p_n^{1/2}E_n[(\kappa(\cdot,U)-\hat{\mu}_U)\otimes(\kappa(\cdot,V)-\hat{\mu}_V)],\end{aligned}$$

where (U,V) is either $(X^{-(i,j)}, X^{-(i,j)})$ or $(X^{-(i,j)}, X^{(i,j)})$. Since $X^{(i,j)}$ is of the fixed dimension 2, we do not need to re-scale $\Sigma_{X^{(i,j)}X^{(i,j)}}$. The next theorem gives the convergence rates of the rescaled covariance operators $\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}(p_n)$ and $\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(p_n)$.

Theorem 31 *If Assumptions 7 and 8, then, as $p_n \to \infty$,*

$$\begin{aligned}&(a) \quad \|\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}(p_n) - \Sigma_{X^{-(i,j)}X^{-(i,j)}}(p_n)\|_{\text{HS}} = O_P(p_n/n), \\ &(b) \quad \|\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(p_n) - \Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)\|_{\text{HS}} = O_P(p_n/n).\end{aligned}$$

Proof Let $U, V, \tilde{U}, \tilde{V}, \bar{V}, \kappa_1, \kappa_2, \gamma_1, \gamma_2, F, G, \tau_1, c_U, c_V$ be as defined in the proof of Theorem 27.

Proof of (a). By definition,

$$\begin{aligned}\hat{\Sigma}_{UU} - \Sigma_{UU} &= E_n[(F - E_n F) \otimes (F - E_n F)] \\ &= E_n(F \otimes F) - E_n(F) \otimes E_n(F).\end{aligned}$$

By the triangular inequality,

$$\begin{aligned}\|\hat{\Sigma}_{UU} - \Sigma_{UU}\|_{\text{HS}} &\leq \|E_n[F \otimes F - E(F \otimes F)]\|_{\text{HS}} + \|E_n(F) \otimes E_n(F)\|_{\text{HS}} \\ &\leq \|E_n(Z)\|_{\text{HS}} + \|E_n(F) \otimes E_n(F)\|_{\text{HS}},\end{aligned}$$

where $Z = F \otimes F - E(F \otimes F)$. By Chebychev's inequality,

$$P(\|E_n(Z)\|_{\text{HS}} > K) \leq K^{-2} E[\|E_n(Z)\|_{\text{HS}}^2].$$

Because $Z_1, \ldots, Z_n$ are i.i.d. random operators with mean 0, we have

$$E[\|E_n(Z)\|_{\text{HS}}^2] = n^{-1} E\|Z\|_{\text{HS}}^2.$$

We next derive magnitude of $E\|Z\|_{\text{HS}}^2$. Note that

$$\begin{aligned}E\|Z\|_{\text{HS}}^2 &= E\langle F \otimes F - E(F \otimes F), F \otimes F - E(F \otimes F)\rangle_{\text{HS}} \\ &= E\langle F \otimes F, F \otimes F\rangle_{\text{HS}} - \langle E(F \otimes F), E(F \otimes F)\rangle_{\text{HS}} \\ &= E\langle F \otimes F, F \otimes F\rangle_{\text{HS}} + O(p^{-1}) \\ &= E(\langle F, F\rangle_{\mathscr{H}_U}^2) + O(p^{-1}),\end{aligned}$$

where the third equality follows from Theorem 9. The inner product $\langle F, F\rangle_{\mathscr{H}_U}$ is calculated as

$$\langle F, F\rangle_{\mathscr{H}_U} = \langle \kappa_1(\cdot, U) - \mu_U, \kappa_1(\cdot, U) - \mu_U\rangle_{\mathscr{H}_U} = 1 - 2\tau_1(U) + c_U,$$

where, for the second equality, we used $\kappa_1(U, U) = 1$. Hence

$$\langle F, F\rangle_{\mathscr{H}_U}^2 = 1 - 4\tau_1(U) + 2c_U + 4\tau_1(U)^2 - 4\tau_1(U)c_U + c_U c_U.$$

Taking expectation of the above quantity and evoking the relation $E[\tau_1(U)] = c_U$, we have

$$E\langle F \otimes F, F \otimes F\rangle_{\text{HS}} = 1 - 2c_U + 4E[\tau_1(U)^2] - 3c_U^2.$$

By Lemma 30, we have

$$c_U = E\kappa_1(U, \tilde{U}) = e^{2\sigma^2} + O(p^{-1}), \quad E[\tau_1(U)^2] = E[\kappa_1(U, \tilde{U})^2] = e^{4\sigma^2} + O(p^{-1}).$$

Hence

$$E\langle F \otimes F, F \otimes F\rangle_{\text{HS}} = 1 - 2e^{2\sigma^2} + e^{4\sigma^2} + O(p^{-1}).$$

Note that the right-hand side (without $O(p^{-1})$ term) is always positive if $\sigma_u^2 > 0$ and $\sigma_v^2 > 0$. So we have

$$\|\hat{\Sigma}_{UU}(p) - \Sigma_{UU}(p)\|_{\text{HS}}^2 \asymp p/n.$$

It follows that

$$P(\|\hat{\Sigma}_{UU}(p) - \Sigma_{UU}(p)\|_{\mathrm{HS}} > K) \leq \frac{C}{K^2}(p/n)$$

for a constant $C > 0$. Denoting the right-hand side by ϵ, we have

$$P(\|\hat{\Sigma}_{UU}(p) - \Sigma_{UU}(p)\|_{\mathrm{HS}} > \sqrt{Cn/p}/\sqrt{\epsilon}) \leq \epsilon,$$

which implies $\|\hat{\Sigma}_{UU}(p) - \Sigma_{UU}(p)\|_{\mathrm{HS}} = O_P(\sqrt{p/n})$.

Proof of (b). Similar to part (a), we have

$$\|\hat{\Sigma}_{UV} - \Sigma_{UV}\|_{\mathrm{HS}} \leq \|E_n(R)\|_{\mathrm{HS}} + \|E_n(F) \otimes E_n(G)\|_{\mathrm{HS}},$$

where $R = F \otimes G - E(F \otimes G)$. By Chebychev's inequality,

$$P(\|E_n(R)\|_{\mathrm{HS}} > K) \leq K^{-2}E[\|E_n(R)\|_{\mathrm{HS}}^2] = n^{-1}K^{-2}E\|R\|_{\mathrm{HS}}^2.$$

By the second relation in Theorem 27,

$$E\|R\|_{\mathrm{HS}}^2 = E\|F \otimes G\|_{\mathrm{HS}}^2 - \|\Sigma_{UV}\|_{\mathrm{HS}}^2 = E(\|F\|_{\mathscr{H}_U}^2 \|G\|_{\mathscr{H}_V}^2) + O(p^{-1}).$$

Similar to part (a), we can show that

$$E\langle F \otimes F, G \otimes G\rangle_{\mathrm{HS}} = 1 - c_V - c_U + 4E[\kappa_1(U, \tilde{U})\kappa_2(V, \bar{V})] - 3c_U c_V.$$

Note that, here, c_V doesn't depend on p. By Lemma 30,

$$\begin{aligned} c_U &= E\kappa_1(U, \tilde{U}) = e^{2\sigma^2} + O(p^{-1}), \\ E[\kappa_1(U, \tilde{U})\kappa_2(V, \bar{V})] &= e^{2\sigma^2} c_V + O(p^{-1}). \end{aligned}$$

Hence

$$E\langle F \otimes F, G \otimes G\rangle_{\mathrm{HS}} = 1 - c_V - e^{2\sigma^2} + e^{2\sigma^2} c_V + O(p^{-1}).$$

So we have $\|\hat{\Sigma}_{UV}(p) - \Sigma_{UV}(p)\|_{\mathrm{HS}}^2 \asymp p/n$, which implies (b). □

Next, we consider the re-scaled eigenvalue problem (8) with $\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}$ and $\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}$ replaced by $\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}(p_n)$ and $\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(p_n)$, while keeping $\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}}$ intact as the dimension of $X^{(i,j)}$ is fixed at 2. Let $\hat{U}^{ij}$ be the same random vector as defined at the end of Section 4.1 except it corresponds to the re-scaled version of the eigenvalue problem (8). Then we have the following convergence rate for $\hat{U}^{ij}$. We will use m_n to abbreviate p_n/n.

Theorem 32 *Suppose*

(a) Assumptions 1, 7, and 8 are satisfied;

(b) *$\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)$ is a finite-rank operator with*

$$
\begin{aligned}
&\mathrm{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)) \subseteq \mathrm{ran}(\Sigma^2_{X^{-(i,j)}X^{-(i,j)}}(p_n)),\\
&\mathrm{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}(p_n)) \subseteq \mathrm{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});
\end{aligned}
$$

the rank d_{ij} of $\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)$ does not depend on n so long as $p_n \geq \max(i,j)$;

(c) *$m_n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$;*

(d) *for each $r = 1, \ldots, d_{ij}$, $\lambda_1^{ij} > \cdots > \lambda_{d_{ij}}^{ij}$.*

Then

$$
\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1} + \eta_n^{-1}m_n^{-1/2} + \eta_n + \epsilon_n).
$$

To prove this theorem, we first prove two lemmas which are modified versions of Lemmas 25 and 26. We will only highlight the differences from the earlier proofs without repeating the similar parts.

Lemma 33 *Suppose*

(a) *the conditions in Corollary 22 are satisfied for $\alpha = 2$, in addition, the rank of A_2 does not depend on p,*

(b) *as $n \to \infty$ (and hence $p \to \infty$),*

$$
\begin{aligned}
&A_1 \asymp 1, A_2 \asymp 1, A_3 \asymp 1,\\
&\|\hat{A}_1 - A_1\|_{\mathrm{HS}} = O_P(m_n^{-1/2}), \quad \|\hat{A}_2 - A_2\|_{\mathrm{HS}} = O_P(m_n^{-1/2}),\\
&\|\hat{A}_3 - A_3\|_{\mathrm{HS}} = O_P(n^{-1/2});
\end{aligned}
$$

(c) *$m_n^{-1} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.*

Then

$$
\begin{aligned}
&\|(\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\mathrm{HS}} = O_P(1),\\
&\|A_1^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\mathrm{HS}} = O_P(1),\\
&\|[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}] A_2\|_{\mathrm{HS}} = O_P(\eta_n^{-1/2} m_n^{-1/2} + \eta_n),\\
&\|A_1^{-1/2} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\mathrm{HS}} = O_P(n^{-1/2} + \epsilon_n).
\end{aligned}
\tag{70}
$$

Proof Let $B_1, B_2, B_3, C_1, C_2, C_3$ be as defined in the proof of Lemma 25. Then following the proof of that lemma we can show

$$
\begin{aligned}
B_1 &= \dot{O}_P(\eta_n^{-1/2} m_n^{-1/2})(A_1 + \eta_n I)^{-3/2}\\
B_2 &= \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1/2} A_1^{-3/2}\\
C_1 &= (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2}\epsilon_n^{-1})\\
C_2 &= A_3^{-1}(A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n).
\end{aligned}
\tag{71}
$$

Hence

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2}\sum_{j=1}^{2} B_i A_2 C_j \\
&= \dot{O}_P(\eta_n^{-1/2} m_n^{-1/2})(A_1+\eta_n I)^{-3/2} A_2 (A_3+\epsilon_n I)^{-1} \dot{O}_P(n^{-1/2}\epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n^{-1/2} m_n^{-1/2})(A_1+\eta_n I)^{-3/2} A_2 A_3^{-1} (A_3+\epsilon_n I)^{-1} \dot{O}_P(\epsilon_n) \\
&+ \dot{O}_P(\eta_n)(A_1+\eta_n I)^{-1/2} A_1^{-3/2} A_2 (A_3+\epsilon_n I)^{-1} \dot{O}_P(n^{-1/2}\epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n)(A_1+\eta_n I)^{-1/2} A_1^{-3/2} A_2 A_3^{-1} (A_3+\epsilon_n I)^{-1} \dot{O}_P(\epsilon_n).
\end{aligned} \tag{72}$$

Because

$$A_1^{-3/2} A_2 A_3^{-1}, \quad A_1^{-3/2} A_2 A_3^{-2}, \quad A_1^{-2} A_2 A_3^{-1}, \quad A_1^{-2} A_2 A_3^{-2}$$

are finite-rank operators with their ranks not dependent on p, by Lemma 21 and Corollary 22, the four operators in the middle of the four terms in (72) all have finite Hilbert-Schmidt norms which do not depend on n. Thus

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2}\sum_{j=1}^{2} B_i A_2 C_j \\
&= \ddot{O}_P(\eta_n^{-1/2}\epsilon_n^{-1} m_n^{-1/2} n^{-1/2} + \eta_n^{-1/2} m_n^{-1/2}\epsilon_n + \eta_n n^{-1/2}\epsilon_n^{-1} + \eta_n\epsilon_n) = \ddot{o}_P(1),
\end{aligned} \tag{73}$$

where the last equality follows from condition (c). Let R be the index set defined in the proof of Lemma 25. Then, by (72) it is easy to see that

$$\textstyle\sum_{(i,j)\in R} B_i A_2 C_j = \ddot{O}_P(\eta_n^{-1/2} m_n^{-1/2} + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = \ddot{o}_P(1),$$

where the last equality follows from condition (c). The first relation in (70) can then be proved following the corresponding steps in the proof of Lemma 25. By (71),

$$\|A_1^{-1/2} A_2 (\hat{A}_3+\epsilon_n I)^{-1}\|_{\mathrm{HS}} = \|A_1^{-1/2} A_2 A_3^{-1}\|_{\mathrm{HS}} + O_P(n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = O_P(1),$$

which is the second relation in (70). Similarly, by (71),

$$[(\hat{A}_1+\eta_n I)^{-1/2} - A_1^{-1/2}]A_2 = B_1 A_2 + B_2 A_2 = \ddot{O}_P(\eta_n^{-1/2} m_n^{-1/2} + \eta_n),$$

which is the third relation in (70). The last relation in (70) is proved exactly as that of Lemma 25.
□

Lemma 34 *Suppose*

(a) the conditions in Corollary 22 are satisfied for $\alpha = 1$, in addition, the rank of A_2 does not depend on p,

(b) as $n \to \infty$ (and hence $p \to \infty$),

$$\begin{aligned}
&A_1 \asymp 1, A_2 \asymp 1, A_3 \asymp 1, \\
&\|\hat{A}_1 - A_1\|_{\mathrm{HS}} = O_P(m_n^{-1/2}), \quad \|\hat{A}_2 - A_2\|_{\mathrm{HS}} = O_P(m_n^{-1/2}), \\
&\|\hat{A}_3 - A_3\|_{\mathrm{HS}} = O_P(n^{-1/2});
\end{aligned}$$

(c) $m_n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.

Then

$$\begin{aligned}
&\|(\hat{A}_1+\eta_n I)^{-1} A_2 (\hat{A}_3+\epsilon_n I)^{-1}\|_{\mathrm{HS}} = O_P(1),\\
&\|A_1^{-1} A_2 (\hat{A}_3+\epsilon_n I)^{-1}\|_{\mathrm{HS}} = O_P(1),\\
&\|[(\hat{A}_1+\eta_n I)^{-1} - A_1^{-1}] A_2\|_{\mathrm{HS}} = O_P(\eta_n^{-1} m_n^{-1/2} + \eta_n),\\
&\|A_1^{-1} A_2 [(\hat{A}_3+\epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\mathrm{HS}} = O_P(n^{-1/2} + \epsilon_n).
\end{aligned} \tag{74}$$

Proof Let B_1, B_2, B_3, C_1, C_2, and C_3 be as defined in Lemma 26. Then

$$B_1 = \dot{O}_P(m_n^{-1/2}\eta_n^{-1})(A_1+\eta_n I)^{-1}, \quad B_2 = \dot{O}_P(\eta_n)(A_1+\eta_n I)^{-1} A_1^{-1}.$$

By these, the last two relations in (71), as well as condition (c), we have

$$\begin{aligned}
&(\hat{A}_1+\eta_n I)^{-1} A_2 (\hat{A}_3+\epsilon_n I)^{-1}\\
&= \textstyle\sum_{i=1}^{3}\sum_{j=1}^{3} B_i A_2 C_j\\
&= A_1^{-1} A_2 A_3 + \ddot{O}_P(m_n^{-1/2}\eta_n^{-1} n^{-1/2}\epsilon_n^{-1} + m_n^{-1/2}\eta_n^{-1}\epsilon_n + m_n^{-1/2}\eta_n^{-1}\\
&\qquad + \eta_n n^{-1/2}\epsilon_n^{-1} + \eta_n\epsilon_n + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n)\\
&= A_1^{-1} A_2 A_3 + \ddot{O}_P(m_n^{-1/2}\eta_n^{-1} + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n)\\
&= A_1^{-1} A_2 A_3 + \ddot{o}_P(1),
\end{aligned}$$

proving the first relation in (74). The second and third relations in (74) are proved as follows:

$$\begin{aligned}
A_1^{-1} A_2 (\hat{A}_3+\epsilon_n I)^{-1} &= A_1^{-1} A_2 C_1 + A_1^{-1} A_2 C_2 + A_1^{-1} A_2 C_3\\
&= \ddot{O}_P(n^{-1/2}\epsilon_n^{-1} + \epsilon_n) + A_1^{-1} A_2 A_3^{-1}\\
&= A_1^{-1} A_2 A_3^{-1} + \ddot{o}_P(1)\\
[(\hat{A}_1+\eta_n I)^{-1} - A_1^{-1}] A_2 &= B_1 A_2 + B_2 A_2 = \ddot{O}_P(\eta_n^{-1} m_n^{-1/2} + \eta_n).
\end{aligned}$$

The proof of the fourth relation in (74) is similar to (45). □

Proof of Theorem 32. Again, since this theorem is a modified version of Theorem 10, we only highlight the differences from the proof of that theorem. Denote the operators

$$\begin{gathered}
(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_n I)^{-1}, \quad \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(p), \quad [\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}(p) + \eta_n I]^{-1/2}\\
\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}, \quad \Sigma_{X^{-(i,j)}X^{(i,j)}}(p), \quad [\Sigma_{X^{-(i,j)}X^{-(i,j)}}(p)]^{-1/2}
\end{gathered}$$

by $\hat{A}$, $\hat{B}$, $\hat{C}$, A, B, C, respectively. Follow the proof of Theorem 10 until (49), and use Theorem 31 to replace (49) by

$$\|\hat{C}(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C}\|_{\mathrm{HS}} = O_P(\eta_n^{-1}\epsilon_n^{-1}m_n^{-1}). \tag{75}$$

Use Theorem 31 and Lemma 33 to replace (50) by

$$\|\hat{C}(\hat{B}-B)\hat{A}B^*\hat{C}\|_{\mathrm{HS}} = O_P(\eta_n^{-1/2}m_n^{-1/2}). \tag{76}$$

Continue to follow the proof of Theorem 10 until (52). Evoke Lemma 33 to replace (52) by

$$\begin{aligned}\|\hat{C}B\hat{A}B^{*}\hat{C}-CBAB^{*}C\|_{\mathrm{HS}}&=O_P(\eta_n^{-1/2}m_n^{-1/2}+\eta_n+n^{-1/2}+\epsilon_n)\\&=O_P(\eta_n^{-1/2}m_n^{-1/2}+\eta_n+\epsilon_n).\end{aligned}\tag{77}$$

Combine (75), (76), and (77) to obtain

$$\|\hat{C}\hat{B}\hat{A}\hat{B}^{*}\hat{C}-CBAB^{*}C\|_{\mathrm{HS}}=O_P(\eta_n^{-1}\epsilon_n^{-1}m_n^{-1}+\eta_n^{-1/2}m_n^{-1/2}+\eta_n+\epsilon_n).$$

Continue to follow the proof of Theorem 10, using the above updated rate, to obtain

$$\begin{aligned}\hat{\lambda}_r^{ij}-\lambda_r^{ij}&=O_P(\eta_n^{-1}\epsilon_n^{-1}m_n^{-1}+\eta_n^{-1/2}m_n^{-1/2}+\eta_n+\epsilon_n),\\\|\hat{\phi}_r^{ij}-\phi_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)}&=O_P(\eta_n^{-1}\epsilon_n^{-1}m_n^{-1}+\eta_n^{-1/2}m_n^{-1/2}+\eta_n+\epsilon_n).\end{aligned}\tag{78}$$

Continue to follow the proof of Theorem 10 until (55), and use Theorem 31 to replace (55) by

$$\|\hat{C}^2(\hat{B}-B)\hat{A}(\hat{B}-B)^{*}\hat{C}\|_{\mathrm{HS}}=O_P(\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1}).\tag{79}$$

Use Theorem 31 and the first relation in (70) to replace (56) by

$$\|\hat{C}^2(\hat{B}-B)\hat{A}B^{*}\hat{C}\|_{\mathrm{HS}}=O_P(\eta_n^{-1}m_n^{-1/2}).\tag{80}$$

Use Theorem 31 and the first relation in (74) to replace (57) by

$$\|\hat{C}^2B\hat{A}(\hat{B}-B)^{*}\hat{C}\|_{\mathrm{HS}}=O_P(m_n^{-1/2}\eta_n^{-1/2}).\tag{81}$$

Continue to follow the proof of Theorem 10 until (58), and use Lemmas 33 and 34 to replace (58) by

$$\|\hat{C}^2B\hat{A}B^{*}\hat{C}-C^2BAB^{*}C\|_{\mathrm{HS}}=O_P(\eta_n^{-1}m_n^{-1/2}+\eta_n+\epsilon_n).\tag{82}$$

Combine (79) through (82) to obtain

$$\hat{C}^2\hat{B}\hat{A}\hat{B}^{*}\hat{C}-C^2BAB^{*}C=\ddot{O}_P(\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1}+\eta_n^{-1}m_n^{-1/2}+\eta_n+\epsilon_n).\tag{83}$$

Continue to follow the proof of Theorem 10 until (84), and use (78) and (83) to replace (60) by

$$\|\hat{f}_r^{ij}-f_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)}=O_P(\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1}+\eta_n^{-1}m_n^{-1/2}+\eta_n+\epsilon_n).\tag{84}$$

Since d_{ij} is bounded as $n\to\infty$, the above implies the asserted result. □

From this point onwards the diverging p_n no longer plays a role in the asymptotics because the kernels $\kappa_{XU}^{i,ij}$ and κ_U^{ij} in the second step of SGM has fixed dimensions that do not depend on n. In other words, Theorems 7, 8, 9, and 11 still apply, but this time to the new convergence rate b_n given in Theorem 32, which leads to the following the convergence rate of the conjoined conditional covariance operator in the $p_n\to\infty$ setting.

Theorem 35 *Suppose the following conditions hold:*

(a) (First-level kernel) $E[\kappa(S,S)] < \infty$ *is satisfied for* $\kappa = \kappa_X^{(i,j)}$ *and* $\kappa = \kappa_X^{-(i,j)}$*;*

(b) (First-level operator) $\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)$ *is a finite-rank operator with rank* d_{ij} *and*

$$\begin{aligned}\mathrm{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)) &\subseteq \mathrm{ran}(\Sigma^2_{X^{-(i,j)}X^{-(i,j)}}(p_n)),\\ \mathrm{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}(p_n)) &\subseteq \mathrm{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});\end{aligned}$$

d_{ij} *does not depend on* n *so long as* $p_n \geq \max(i,j)$*; all the* d_{ij} *nonzero eigenvalues of*

$$\Sigma_{X^{(i,j)}X^{-(i,j)}}(p_n)[\Sigma_{X^{-(i,j)}X^{-(i,j)}}(p_n)]^{-1}\Sigma_{X^{-(i,j)}X^{(i,j)}}(p_n)$$

are distinct, and the eigenvalues gaps do not tend to 0;

(c) (First-level tuning parameters) $m_n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$, $\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1} + \eta_n^{-1}m_n^{-1/2} + \eta_n^{1/2} + \epsilon_n \prec 1$*;*

(d) (Second-level kernel) $E[\kappa(S,S)] < \infty$ *is satisfied for* $\kappa = \kappa_U^{ij}$, $\kappa_{XU}^{i,ij}$, *and* $\kappa_{XU}^{j,ij}$*; furthermore, they are transparent kernels;*

(e) (Second-level operators) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(f) (Second-level tuning parameter) $\delta_n \asymp \eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1} + \eta_n^{-1}m_n^{-1/2} + \eta_n + \epsilon_n$.

Then

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\mathrm{HS}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}m_n^{-1} + \eta_n^{-1}m_n^{-1/2} + \eta_n + \epsilon_n). \tag{85}$$

Appendix J. Joint distribution satisfying condition (3)

It is relatively easy to find joint distributions of $X = (X^1, \ldots, X^p)^\intercal$ that satisfy the (3) for every pair of nodes (i,j) with $\mathcal{G}_{X^{-(i,j)}}$ being a proper sub-σ-field of $\sigma(X^{-(i,j)})$. The multivariate Gaussian distribution is an obvious (but trivial) example, because the conditional distribution of $X^{(i,j)}$ given $X^{-(i,j)}$ depends on a linear function of $X^{-(i,j)}$. In this case, U^{ij} is of dimension 1. Another example is the copula Gaussian distribution; that is, there exist injective functions $c_1, \ldots, c_p$ such that $(c_1(X^1), \ldots, c_p(X^p)) = (C^1, \ldots, C_p)$ is multivariate Gaussian. In this case the conditional distribution of $C^{(i,j)}$ given $C^{-(i,j)}$ is a linear function of $C^{-(i,j)}$. This implies that there exists a 1-dimensional U^{ij} such that $X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)}|U^{ij}$. There are also abundant examples satisfying (3) that are unrelated to multivariate Gaussian distribution. For example, consider a multivariate distribution determined by the graph where each pair of vertices can have at most r neighbors. In this case, for each pair (i,j), there exists a U^{ij} of dimension at most r that satisfies (3).

Appendix K. Equivalent condition for $\mathrm{ran}(B) \subseteq \mathrm{ran}(A)$

Let $\mathscr{H}$ and $\mathscr{K}$ be separable Hilbert spaces, let $A : \mathscr{H} \to \mathscr{H}$ a compact and self adjoint operator, and let $B : \mathscr{K} \to \mathscr{H}$ be compact operator. Let $\{(\lambda_i, u_i) : i = 1, 2, \ldots\}$ be the eigenvalue-eigenfunction sequence of A, with $|\lambda_1| \geq |\lambda_2| \geq \cdots$, and let $\{(\tau_i, v_i, w_i) : i = 1, 2, \ldots\}$ be a sequence where τ_i is the ith largest singular value of B, v_i is the corresponding left eigenfunction of

B, and w_i is the corresponding right eigenfunction of B, with $\tau_1 \geq \tau_2 \geq \cdots$. The next proposition gives a necessary and sufficient condition for $\mathrm{ran}(B) \subseteq \mathrm{ran}(A)$ in terms of their eigenvalues, eigenfunctions, singular values, and singular functions. Without loss of generality, assume that $\{u_i\}$ and $\{v_i\}$ are orthonormal bases of $\mathscr{H}$ and $\{w_i\}$ is an orthonormal basis of $\mathscr{K}$. In the following, in a series such as $\sum_{i=1}^{\infty} a_i b_i$, we allow a_i or b_i to be ∞, and treat as if ∞ were a number. In particular, we adopt the convention $0 \cdot \infty = 0$ and $c \cdot \infty = \infty$ if $c \neq 0$.

Proposition 36 *If the assumptions in the last paragraph are satisfied, then the following statements hold true:*

1. $\mathrm{ran}(B) = \{f \in \mathscr{H} : \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty\}$*;*
2. $\mathrm{ran}(B) \subseteq \mathrm{ran}(A)$ *if and only if, for any* $f \in \mathscr{H}$*, and as* $n \to \infty$*,*

$$\sum_{i=1}^{n} \lambda_i^{-2} \langle f, u_i \rangle_{\mathscr{H}}^2 = O\left(\sum_{i=1}^{n} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2\right).$$

In part 1, when there are only a finite number (say n_0) of nonzero τ_i, we have $\tau_i^{-2} = \infty$ for $i > n_0$. Therefore, by the stated convention about ∞, the inequality $\sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty$ holds if and only if $\langle f, v_i \rangle_{\mathscr{H}} = 0$ for all $i > n_0$. To provide further intuition, we give a necessary and sufficient condition for Assumption 2 in Proposition 36, which implies that following examples satisfy and violate Assumption 2, respectively:

1. Assumption 2 is satisfied if $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ has finite number of nonzero singular values, and the corresponding left singular functions are contained in the subspace spanned by a finite number of eigenfunctions of $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$;
2. Assumption 2 is violated if the left singular functions of $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ are aligned with the eigenfunctions of $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$, and the singular values of $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ converge to 0 at a slower rate than the eigenvalues of $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$.

Proof *1.* First, consider the cases where there are $n_0 < \infty$ nonzero τ_i. In this case, it is easy to see that

$$\mathrm{ran}(B) = \mathrm{span}(v_1, \ldots, v_{n_0}\} = \{f \in \mathscr{H} : \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty\}.$$

Next, assume all τ_i's are nonzero, and let

$$\mathscr{S} = \left\{ f \in \mathscr{H} : \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty \right\},$$

and f a member of $\mathscr{S}$. Since $\sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty$, the function

$$h = \sum_{i=1}^{\infty} \tau_i^{-1} \langle f, v_i \rangle_{\mathscr{H}} w_i$$

is a well defined member of $\mathscr{K}$. Furthermore,

$$\begin{aligned} Bh &= \sum_{i=1}^{\infty} \tau_i \langle h, w_i \rangle_{\mathscr{K}} v_i \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tau_i \tau_j^{-1} \langle f, v_j \rangle_{\mathscr{H}} \langle w_j, w_i \rangle_{\mathscr{K}} v_i \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tau_i \tau_j^{-1} \langle f, v_j \rangle_{\mathscr{H}} \langle w_j, w_i \rangle_{\mathscr{K}} v_i \\ &= \sum_{i=1}^{\infty} \langle f, v_i \rangle_{\mathscr{H}} v_i = f. \end{aligned}$$

Thus f is a member of $\operatorname{ran}(B)$, which proves $\mathscr{S} \subseteq \operatorname{ran}(B)$.

If $f \in \operatorname{ran}(B)$, then there is an $h \in \mathscr{K}$ such that

$$f = Bh = \sum_{i=1}^{\infty} \tau_i \langle h, w_i \rangle_{\mathscr{K}} v_i.$$

Hence

$$\begin{aligned} \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 &= \sum_{i=1}^{\infty} \tau_i^{-2} \left(\sum_{j=1}^{\infty} \tau_j \langle h, w_j \rangle_{\mathscr{K}} \langle v_j, v_i \rangle_{\mathscr{H}} \right)^2 \\ &= \sum_{i=1}^{\infty} \tau_i^{-2} \tau_i^2 \langle h, w_i \rangle_{\mathscr{K}}^2 \\ &= \sum_{i=1}^{\infty} \langle h, w_i \rangle_{\mathscr{K}}^2 = \|h\|_{\mathscr{K}}^2 < \infty. \end{aligned}$$

Thus f is a member of $\mathscr{S}$.

2. Applying part 1 of this proposition to operators A and B, we have

$$\begin{aligned} \operatorname{ran}(A) &= \left\{ f : \sum_{i=1}^{\infty} \lambda_i^{-2} \langle f, u_i \rangle_{\mathscr{H}}^2 < \infty \right\}, \\ \operatorname{ran}(B) &= \left\{ f : \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2 < \infty \right\}. \end{aligned}$$

Let

$$\begin{aligned} a_n &= \sum_{i=1}^{n} \lambda_i^{-2} \langle f, u_i \rangle_{\mathscr{H}}^2, \quad & b_n &= \sum_{i=1}^{n} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2, \\ a &= \sum_{i=1}^{\infty} \lambda_i^{-2} \langle f, u_i \rangle_{\mathscr{H}}^2, \quad & b &= \sum_{i=1}^{\infty} \tau_i^{-2} \langle f, v_i \rangle_{\mathscr{H}}^2. \end{aligned}$$

Then $\lim_{n \to \infty} a_n = a$ and $\lim_{n \to \infty} b_n = b$. Hence the following statements are equivalent:

1. $b < \infty \Rightarrow a < \infty$;
2. $b_n = O(a_n)$,

which proves the second assertion. □

Appendix L. Coordinate representation of SGM algorithm

L.1 Coordinate mapping

We first briefly describe how to represent operators as matrices and functions as vectors using coordinate mapping. A full description of coordinate mapping can be found in Sections 12.3 and 12.4 of Li (2018b). Let $\mathscr{H}_1$ and $\mathscr{H}_2$ be finite-dimensional Hilbert spaces with spanning sets $\mathscr{B}_1 = \{h_{11}, \ldots, h_{1m_1}\}$ and $\mathscr{B}_2 = \{h_{21}, \ldots h_{2m_2}\}$. Here, we allow the vectors in the spanning sets to be linearly dependent. Any function $f \in \mathscr{H}_1$ can be represented as a linear combination of vectors in $\mathscr{B}_1$; we call the $\mathbb{R}^{m_1}$-vector of linear coefficients the coordinate of f with respect to $\mathscr{B}_1$ and denote the coordinate by $[f]_{\mathscr{B}_1}$. If $A : \mathscr{H}_1 \to \mathscr{H}_2$ is a linear operator, then Af is a member of $\mathscr{H}_2$ and has a coordinate $[Af]_{\mathscr{B}_2}$ with respect to the spanning set $\mathscr{B}_2$ of $\mathscr{H}_2$. There is always a matrix $M \in \mathbb{R}^{m_2 \times m_1}$ such that $[Af]_{\mathscr{B}_2} = M[f]_{\mathscr{B}_1}$, and we call this matrix the coordinate of A with respect to $\mathscr{B}_1$-$\mathscr{B}_2$, and denote it by ${}_{\mathscr{B}_2}[A]_{\mathscr{B}_1}$.

The following proposition will be used in the subsequent discussions. It concerns a single finite-dimensional Hilbert space and its spanning set $\mathscr{B} = \{h_1, \ldots, h_m\}$. Let $A : \mathscr{H} \to \mathscr{H}$ be a self-adjoint operator.

Proposition 37 *Let $G_{\mathscr{B}} = \{\langle h_a, h_b \rangle_{\mathscr{H}}\}_{a,b=1}^n$ be the Gram matrix of the set $\mathscr{B}$, $I : \mathscr{H} \to \mathscr{H}$ the identity mapping, and $\epsilon > 0$ a constant. Then*

1. $\|A\|_{\mathrm{HS}} = \|G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2}\|_{\mathrm{F}}$, *where $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm of a linear operator, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm of a matrix, and $G_{\mathscr{B}}^{\dagger 1/2}$ is the Moore-Penrose inverse of the matrix $G_{\mathscr{B}}^{1/2}$.*
2. ${}_{\mathscr{B}}[(A + cI)^{-1}]_{\mathscr{B}} = G_{\mathscr{B}}^{\dagger 1/2}\{G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2} + cQ_{\mathscr{B}}\}^{\dagger}G_{\mathscr{B}}^{1/2}$, *where $Q_{\mathscr{B}}$ is the projection on to* $\mathrm{span}\{[h_1]_{\mathscr{B}}, \ldots, [h_m]_{\mathscr{B}}\}$.

The proof, which is omitted, can be done using Theorem 8 of Li and Solea (2018a). Note that part 2 of the proposition can be equivalently written as

$$ {}_{\mathscr{B}}[(A + \epsilon I)^{-1}]_{\mathscr{B}} = G_{\mathscr{B}}^{\dagger 1/2}\{G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2} + \epsilon I_n\}^{\dagger}G_{\mathscr{B}}^{1/2}, $$

because $G_{\mathscr{B}}^{\dagger 1/2}G_{\mathscr{B}}^{1/2} = Q_{\mathscr{B}}$.

L.2 Matrix representation for eigenvalue problem (8)

Let $K_{X^{-(i,j)}}$, $G_{X^{-(i,j)}}$, Q, and $\mathscr{H}_X^{-(i,j)}$ be the objects defined in Section 4 of the manuscript. Then, it can be easily verified that the number

$$ \begin{aligned} &\langle \kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})], \\ &\qquad\qquad \kappa_X^{-(i,j)}(\cdot, X_b^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\rangle_{-(i,j)} \end{aligned} $$

is the (a, b)th entry of $G_{X^{-(i,j)}}$. In other words, $G_{X^{-(i,j)}}$ is the Gram matrix of the spanning set of $\mathscr{H}_X^{-(i,j)}$:

$$\{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] : \; a = 1, \ldots, n\} \equiv \mathscr{C}.$$

By the same argument $G_{X^{(i,j)}}$ is the Gram matrix of the spanning set of $\mathscr{H}_X^{(i,j)}$:

$$\{\kappa_X^{(i,j)}(\cdot, X_a^{(i,j)}) - E_n[\kappa_X^{(i,j)}(\cdot, X^{(i,j)})] : \; a = 1, \ldots, n\} \equiv \mathscr{B}.$$

By Lemma 12.3, part 4, in Li (2018b),

$$\begin{aligned}&\langle f, \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f\rangle_{-(i,j)} \\ &\quad = [f]_{\mathscr{C}}{}^{\mathsf{T}} G_{X^{-(i,j)}}\, [\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f]_{\mathscr{C}}.\end{aligned}$$

By parts 1 and 3 of the same lemma,

$$\begin{aligned}&[\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f]_{\mathscr{C}} \\ &\quad = ({}_{\mathscr{C}}[\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}]_{\mathscr{B}})({}_{\mathscr{B}}[(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}]_{\mathscr{B}})({}_{\mathscr{B}}[\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}}]_{\mathscr{C}})[f]_{\mathscr{C}}.\end{aligned}$$

By Theorem 12.1 of Li (2018b),

$$\begin{aligned}{}_{\mathscr{C}}[\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}]_{\mathscr{B}} &= G_{X^{(i,j)}}, \\ {}_{\mathscr{B}}[\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}}]_{\mathscr{B}} &= G_{X^{(i,j)}}, \\ {}_{\mathscr{B}}[\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} &= G_{X^{-(i,j)}}.\end{aligned}$$

By Proposition 37,

$$\begin{aligned}{}_{\mathscr{B}}[(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}]_{\mathscr{B}} &= G_{X^{(i,j)}}^{\dagger 1/2}(G_{X^{(i,j)}}^{1/2} G_{X^{(i,j)}} G_{X^{(i,j)}}^{\dagger 1/2} + \epsilon_X^{(i,j)} Q_{\mathscr{B}})^{\dagger} G_{X^{(i,j)}}^{1/2} \\ &= Q_{\mathscr{B}}(G_{X^{(i,j)}} +_X^{(i,j)} I_n)^{-1} Q_{\mathscr{B}}.\end{aligned}$$

Hence

$$\begin{aligned}&[\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f]_{\mathscr{C}} \\ &\quad = G_{X^{(i,j)}}(G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}}[f]_{\mathscr{C}}\end{aligned}$$

and consequently,

$$\begin{aligned}&\langle f, \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}\hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f\rangle_{-(i,j)} \\ &\qquad\qquad = [f]_{\mathscr{C}}{}^{\mathsf{T}} G_{X^{-(i,j)}}\, G_{X^{(i,j)}}(G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}}[f]_{\mathscr{C}}.\end{aligned}$$

Similarly,

$$\langle f, \hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} f\rangle_{-(i,j)} = [f]_{\mathscr{C}}{}^{\mathsf{T}} G_{X^{-(i,j)}}^2 [f]_{\mathscr{C}}.$$

Thus the operator-level generalized eigenvalue problem (8) can be rewritten in the following matrix form

$$\begin{aligned}\text{maximize} \quad & b^{\mathsf{T}} G_{X^{-(i,j)}} G_{X^{(i,j)}}(G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} b \\ \text{subject to} \quad & b^{\mathsf{T}} G_{X^{-(i,j)}}^2 b = 1.\end{aligned}$$

Setting $G_{X^{-(i,j)}} b = a$ and solving this equation for a with Tychonoff regularization, we have $b = (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} a$. Thus, at the kth step, a is simply the kth eigenvector of

$$(G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} G_{X^{-(i,j)}} \\ G_{X^{(i,j)}} (G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1}.$$

Let $a_1, \ldots, a_{d_{ij}}$ be first d_{ij} eigenfunctions of the above matrix, and $b^r = (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} a^r$ for $r = 1, \ldots, d_{ij}$. The eigenfunctions $f_1^{ij}, \ldots, f_{d_{ij}}^{ij}$ of the problem (8) are then

$$f_r^{ij} = \sum_{a=1}^{n} b_a^r \{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\}.$$

L.3 Matrix representation of $\|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}}$

By Theorem 12.1 of Li (2018b), the coordinate representations of the estimated covariance operators in (10) are

$$\begin{aligned} {}_{\mathscr{B}_{XU}^{i,ij}}[\hat{\Sigma}_{(X^i U^{ij})(X^j U^{ij})}]_{\mathscr{B}_{XU}^{j,ij}} &= G_{X^j U^{ij}}, & {}_{\mathscr{B}_{XU}^{i,ij}}[\hat{\Sigma}_{(X^i U^{ij}) U^{ij}}]_{\mathscr{B}_U^{ij}} &= G_{U^{ij}}, \\ {}_{\mathscr{B}_U^{ij}}[\hat{\Sigma}_{U^{ij} (X^j U^{ij})}]_{\mathscr{B}_{XU}^{j,ij}} &= G_{X^j U^{ij}}, & {}_{\mathscr{B}_U^{ij}}[\hat{\Sigma}_{U^{ij} U^{ij}}]_{\mathscr{B}_U^{ij}} &= G_{U^{ij}}. \end{aligned}$$

Applying the above relations and part 2 of Proposition 37, we obtain the coordinate representation of the conjoined conditional covariance operator for each (i, j) as

$${}_{\mathscr{B}_{XU}^{i,ij}}[\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | U^{ij}}]_{\mathscr{B}_{XU}^{j,ij}} = G_{X^j U^{ij}} - G_{U^{ij}} (G_{U^{ij}} + \epsilon_U^{(i,j)} Q)^{\dagger} G_{X^j U^{ij}}.$$

By part 1 of Proposition 37, the Hilbert Schmidt norm of the above operator is the Frobenius norm

$$\left\| G_{X^i U^{ij}}^{1/2} G_{X^j U^{ij}}^{1/2} - G_{X^i U^{ij}}^{1/2} G_{U^{ij}} (G_{U^{ij}} + \epsilon_U^{(i,j)} Q)^{\dagger} G_{X^j U^{ij}}^{1/2} \right\|_{\mathrm{F}}.$$

Appendix M. Additional simulation for estimating threshold ρ

Appendix N. Miscellaneous

Nonlinear sufficient dimension reduction is a particularly natural framework for reducing dimension in a statistical graphical model: the following example illustrates the conceptual difficulty to use linear sufficient dimension reduction. Suppose X has four components X^1, X^2, X^3, X^4 and the linear sufficient dimension reduction relation is imposed on (X^3, X^4) by

$$\begin{pmatrix} X^3 \\ X^4 \end{pmatrix} = \begin{pmatrix} (X^1 + X^2)^2 \\ \sin(X^1 + X^2) \end{pmatrix} + \begin{pmatrix} \epsilon^3 \\ \epsilon^4 \end{pmatrix}$$

where $(\epsilon^3, \epsilon^4) \perp\!\!\!\perp X$. Then, there might not exist constants β_1, β_2 and a functions f_1, f_2 such that

$$\begin{pmatrix} X^1 \\ X^2 \end{pmatrix} = \begin{pmatrix} f_1(\beta_1 X^3 + \beta_2 X^4) \\ f_2(\beta_1 X^3 + \beta_2 X^4) \end{pmatrix} + \begin{pmatrix} \epsilon^1 \\ \epsilon^2 \end{pmatrix}$$

where $(\epsilon^1, \epsilon^2) \perp\!\!\!\perp X$. In other works, linear sufficient dimension reduction is not rich enough to be imposed on every pair of nodes without causing inconsistency. This is not a problem for nonlinear sufficient dimension reduction, as it imposes no specific form on the conditional distributions of $X^{(i,j)}$ given $X^{-(i,j)}$.

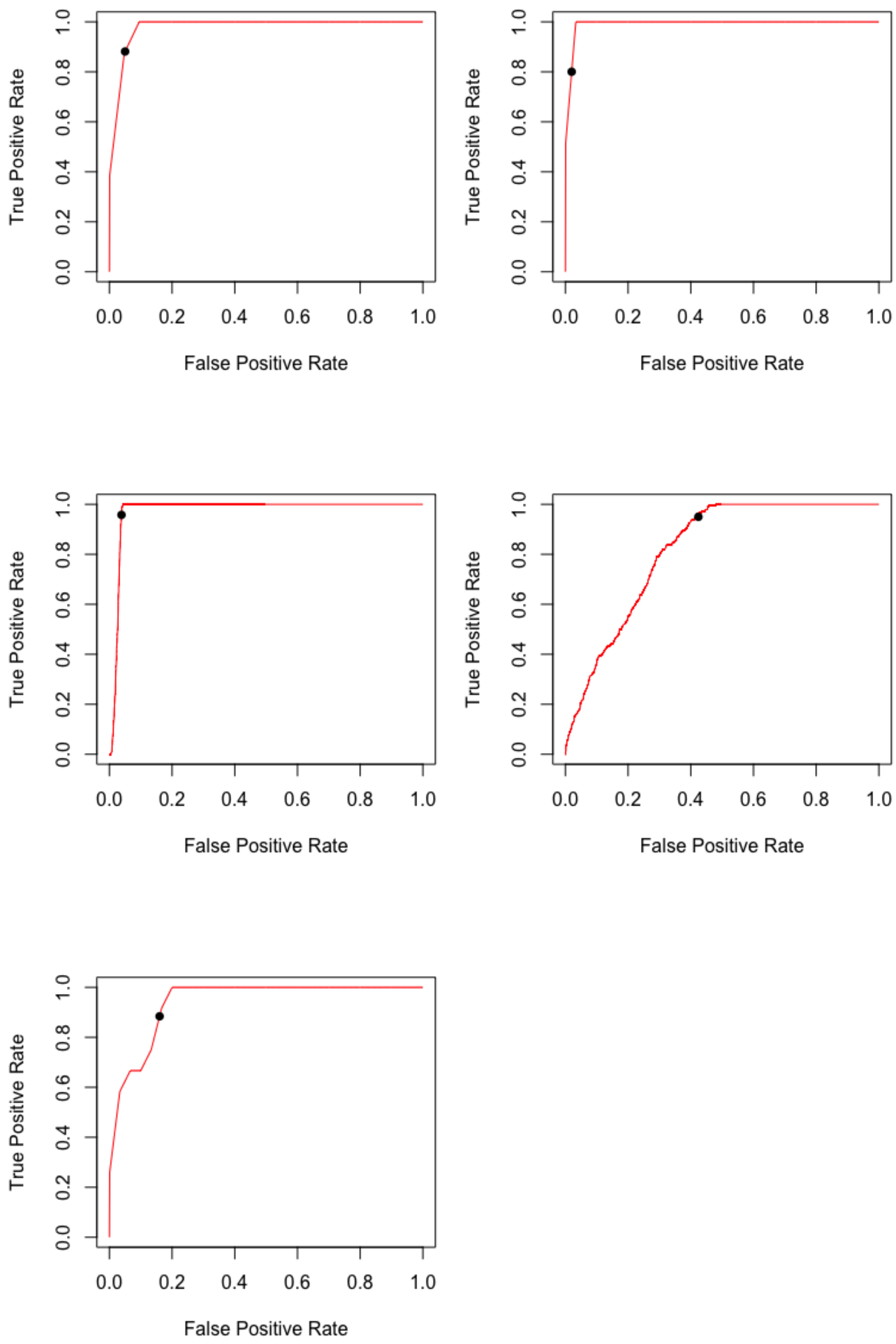

Figure 7: Threshold determination for simulation studies. Upper panels: Model I with $n = 1000$ (left) and Model II with $n = 1000$ (right); middle panels: Model III with $n = 50$ (left) and Model IV with $n = 50$ (right); bottom panel: Model V with $n = 100$. The red curves are the receiver operating characteristic curves; and the black dots are the positions of the thresholds determined generalized cross validation.

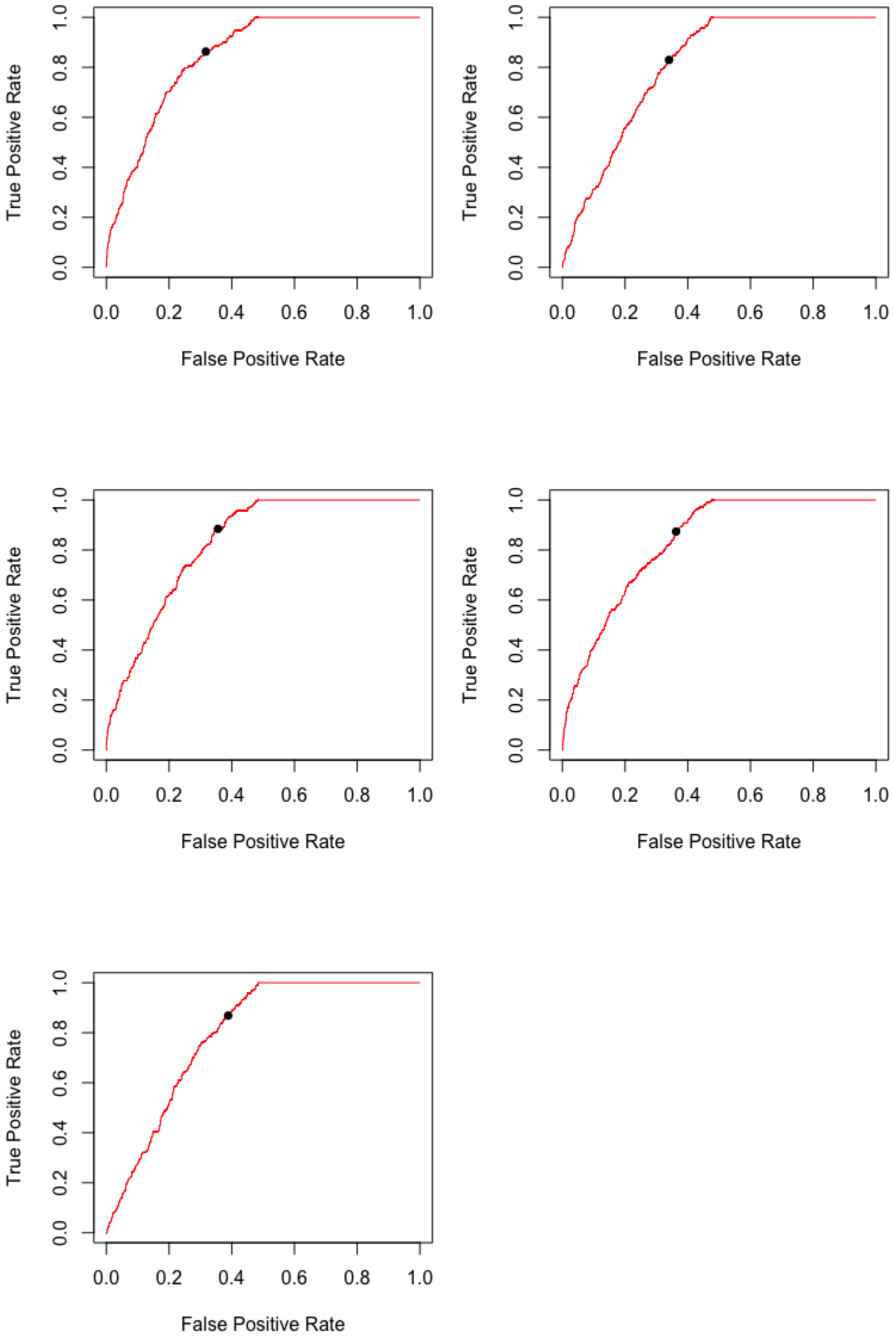

Figure 8: Threshold determination for application. Upper panels: Network 1 (left) and Network 2 (right); middle panels: Network 3 (left) and Network 4 (right); bottom panel: Network 5. The red curves are the receiver operating characteristic curves; and the black dots are the positions of the thresholds determined generalized cross validation.