Biometrika (2016), **103**, 3, *pp.* 513–530
Printed in Great Britain

doi: 10.1093/biomet/asw028
Advance Access publication 24 August 2016

On an additive partial correlation operator and nonparametric estimation of graphical models

By KUANG-YAO LEE

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, Connecticut 06520, U.S.A.
kuang-yao.lee@yale.edu

BING LI

Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, Pennsylvania 16802, U.S.A.
bxl9@psu.edu

AND HONGYU ZHAO

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, Connecticut 06520, U.S.A.
hongyu.zhao@yale.edu

Summary

We introduce an additive partial correlation operator as an extension of partial correlation to the nonlinear setting, and use it to develop a new estimator for nonparametric graphical models. Our graphical models are based on additive conditional independence, a statistical relation that captures the spirit of conditional independence without having to resort to high-dimensional kernels for its estimation. The additive partial correlation operator completely characterizes additive conditional independence, and has the additional advantage of putting marginal variation on appropriate scales when evaluating interdependence, which leads to more accurate statistical inference. We establish the consistency of the proposed estimator. Through simulation experiments and analysis of the DREAM4 Challenge dataset, we demonstrate that our method performs better than existing methods in cases where the Gaussian or copula Gaussian assumption does not hold, and that a more appropriate scaling for our method further enhances its performance.

Some key words: Additive conditional covariance operator; Additive conditional independence; Copula; Gaussian graphical model; Partial correlation; Reproducing kernel.

1. Introduction

We propose a new statistical object, the additive partial correlation operator, for estimating nonparametric graphical models. This operator is an extension of the partial correlation coefficient (Muirhead, 2005) to the nonlinear setting. It is akin to the additive conditional covariance operator of Li et al. (2014) but achieves better scaling, leading to enhanced estimation accuracy, when characterizing conditional independence in graphical models.

Let $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a p-dimensional random vector. Let $\mathcal{G} = \big(\mathcal{V}, \mathcal{E}\big)$ be an undirected graph, where $\mathcal{V} = \{1, \ldots, p\}$ represents the set of vertices corresponding to p random variables and $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V},\ i \neq j\}$ represents the set of undirected edges. For convenience we assume that $i > j$. A common approach to modelling an undirected graph is to associate separation with conditional independence; that is, node i and node j are separated if and only if X_i and X_j are independent given the rest of X. In symbols,

$$(i, j) \notin \mathcal{E} \iff X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}, \tag{1}$$

where $X_{-(i,j)}$ represents X with its ith and jth components removed. Intuitively, this means that nodes i and j are connected if and only if, after removing the effects of all the other nodes, X_i and X_j still depend on each other. In other words, nodes i and j are connected in the graph if and only if X_i and X_j have a direct relation. The statistical problem is to estimate $\mathcal{G}$ based on a sample of X.

One of the most commonly used statistical models for (1) is the Gaussian graphical model, which assumes that X satisfies (1) and is distributed as $N(\mu, \Sigma)$ for a nonsingular covariance matrix Σ. An appealing property of the multivariate Gaussian distribution is that conditional independence is completely characterized by the zero entries of the precision matrix. Specifically, let $\Omega = \Sigma^{-1}$ be the precision matrix and ω_{ij} its (i, j)th element. Then

$$X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)} \iff \omega_{ij} = 0. \tag{2}$$

Thus, under Gaussianity, estimation of $\mathcal{G}$ amounts to identifying the zero entries or, equivalently, the sparsity pattern of the precision matrix Ω. Many procedures have been developed to estimate the Gaussian graphical model. For example, Yuan & Lin (2007), Banerjee et al. (2008) and Friedman et al. (2008) considered penalized maximum likelihood estimation with L_1 penalties on Ω. Based on a relation between partial correlations and regression coefficients, Meinshausen & Bühlmann (2006) and Peng et al. (2009) proposed to select the neighbours of each node by solving multiple lasso problems (Tibshirani, 1996). Other recent advances include the work of Bickel & Levina (2008a, b), who used hard thresholding to determine the sparsity pattern, Lam & Fan (2009), who used the smoothly clipped absolute deviation penalty (Fan & Li, 2001), and Yuan (2010) and Cai et al. (2011), who used the Danzig selector (Candès & Tao, 2007).

Since Gaussianity could be restrictive in applications, many recent papers have considered extensions. The challenge is not only to relax Gaussianity but also to preserve the simplicity of the conditional independence structure imparted by the Gaussian distribution. One elegant solution is to assume a copula Gaussian model, under which the data can be transformed marginally to multivariate Gaussianity; see Liu et al. (2009, 2012), Xue & Zou (2012) and Harris & Drton (2013). The copula Gaussian model preserves the equivalence (2) for the transformed X, without requiring the X_i to be marginally Gaussian. Other work on non-Gaussian graphical models includes Fellinghauer et al. (2013) and Voorman et al. (2014). In their settings, a given node is associated with its neighbours via either a semiparametric or a nonparametric model.

Another extension is the additive semigraphoid model of Li et al. (2014), which is based on a new statistical relation called additive conditional independence. By generalizing the precision matrix to the additive precision operator and replacing the conditional independence in (2) by additive conditional independence, Li et al. (2014) showed that the equivalence (2) emerges at the linear operator level, at which no distributional assumption is needed.

The primary motivation for introducing additive conditional independence is to maintain nonparametric flexibility without employing high-dimensional kernels. The distribution of points in a Euclidean space becomes increasingly sparse as the dimension of the space increases. For a kernel estimator in such spaces to be effective, we need to increase the bandwidth; otherwise we may have very few observations within a local ball of radius equal to the bandwidth.

Increasing bandwidth, however, also increases bias. Therefore we face the dilemma of either increased bias or lack of data in each local region, a phenomenon known as the curse of dimensionality (Bellman, 1957). To avoid this problem while extracting useful information from high-dimensional data, one must impose some kind of additional structure, such as parametric models, sparsity or linear indices. The structure imposed by additive conditional independence is additivity, which allows us to employ only one-dimensional kernels, thus avoiding high dimensionality. The cost is that the graphical model is no longer characterized by conditional independence. Nonetheless, Li et al. (2014) have shown that additive conditional independence satisfies the semigraphoid axioms (Pearl & Verma, 1987; Pearl et al., 1989), a set of four fundamental properties of conditional independence.

To estimate the additive semigraphoid model, Li et al. (2014) proposed the additive conditional covariance and additive precision operators, which extend the conditional covariance and precision matrices and characterize additive conditional independence without distributional assumptions. In the classical setting, the conditional covariance $\mathrm{cov}(U, V \mid W)$ between two random variables U and V given a third random variable W describes the strength of dependence between U and V after removing the effect of W. However, it is confounded by statistical variations in $\mathrm{var}(U \mid W)$ and $\mathrm{var}(V \mid W)$, which have nothing to do with the conditional dependence. Partial correlation is designed to remove these effects, so that conditional dependence is retained. The additive partial correlation operator that we propose serves the same purpose in the nonlinear setting. We will also propose an estimator of the new operator, and establish its consistency along with that of the estimator of the additive conditional covariance operator, which was not proved in Li et al. (2014). Based on the additive partial correlation operator, we develop an estimator for the additive semigraphoid model and establish the consistency of this procedure.

All the proofs, as well as some additional propositions and numerical results, are presented in the Supplementary Material.

2. Additive conditional independence and graphical models

2·1. *Additive conditional independence*

Let $(\Omega, \mathcal{F}, P)$ be a probability space, Ω_X a subset of $\mathbb{R}^p$, and $X : \Omega \to \Omega_X$ a random vector. Let P_X be the distribution of X. Let X_i be the ith component of X, and let Ω_{X_i} be the support of X_i. For a subvector U of X, let P_U be the distribution of U and $L_2(P_U)$ the centred L_2 class

$$\{f : E\{f(U)\} = 0,\ E\{f^2(U)\} < \infty\}.$$

We assume that all functions in $L_2(P_U)$ have mean zero, because constants have no bearing on our construction. Additive conditional independence (Li et al., 2014) was introduced in terms of the $L_2(P_X)$ geometry. Suppose that $U = (U_1, \ldots, U_a)^{\mathrm{T}}$, $V = (V_1, \ldots, V_b)^{\mathrm{T}}$ and $W = (W_1, \ldots, W_c)^{\mathrm{T}}$ are subvectors of X. For a subvector such as U, let $\mathscr{L}_U$ be the additive family formed by functions in each $L_2(P_{U_i})$; that is,

$$\mathscr{L}_U = \{u_1 + \cdots + u_a : u_i \in L_2(P_{U_i}),\ i = 1, \ldots, a\}.$$

Note that $\mathscr{L}_X$ and $L_2(P_X)$ are different: the former consists of additive functions, while the latter has no such restriction. If $\mathscr{S}_1$ and $\mathscr{S}_2$ are subspaces of $\mathscr{L}_X$, we write $\mathscr{S}_1 \perp \mathscr{S}_2$ if $\mathrm{cov}\{f_1(X), f_2(X)\} = \langle f_1, f_2 \rangle_{L_2} = 0$ for all $f_1 \in \mathscr{S}_1$ and $f_2 \in \mathscr{S}_2$, where $\langle \cdot\,, \cdot \rangle_{L_2}$ denotes the inner product in $L_2(P_X)$. If $\mathscr{S}_1 \subseteq \mathscr{S}_2$, we denote by $\mathscr{S}_2 \ominus \mathscr{S}_1$ the set of functions $f \in \mathscr{S}_2$ such that $f \perp \mathscr{S}_1$. Also, let $\mathscr{S}_1 + \mathscr{S}_2$ denote the subspace $\{f_1 + f_2 : f_1 \in \mathscr{S}_1,\ f_2 \in \mathscr{S}_2\}$.

Definition 1. *We say that U and V are additively independent conditional on W if and only if*

$$(\mathcal{L}_U + \mathcal{L}_W) \ominus \mathcal{L}_W \perp (\mathcal{L}_V + \mathcal{L}_W) \ominus \mathcal{L}_W. \tag{3}$$

We denote this relation by $U \perp\!\!\!\perp_{\mathrm{A}} V \mid W$.

Li et al. (2014) showed that the three-way relation $U \perp\!\!\!\perp_{\mathrm{A}} V \mid W$ satisfies the four semigraphoid axioms (Pearl & Verma, 1987; Pearl et al., 1989), which are features abstracted from probabilistic conditional independence suitable for describing a graph.

Based on (3), Li et al. (2014) proposed the following graphical model. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be as defined in § 1.

Definition 2. *We say that X follows an additive semigraphoid model with respect to $\mathcal{G}$ if*

$$(i, j) \notin \mathcal{E} \iff X_i \perp\!\!\!\perp_{\mathrm{A}} X_j \mid X_{-(i,j)}.$$

Li et al. (2014) developed theoretical results and estimation methods for the additive semigraphoid model using the additive L_2 space $\mathcal{L}_X$.

2·2. *Additive reproducing kernel Hilbert spaces*

Rather than use L_2 geometry, here we use reproducing kernel Hilbert space geometry to derive our new operator and related methods. This is mainly because many asymptotic tools for linear operators have recently been developed in the reproducing kernel Hilbert space setting (Fukumizu et al., 2007; Bach, 2008). The advantage of this alternative formulation will become clear in § 4. Let $\kappa_{X_i} : \Omega_{X_i} \times \Omega_{X_i} \to \mathbb{R}$ be a positive-definite kernel. For convenience, we assume Ω_{X_i} and κ_{X_i} to be the same for all $i = 1, \ldots, p$ and write the common kernel function as κ. Let $\mathcal{H}_{X_i}$ be the reproducing kernel Hilbert space of functions of X_i based on the kernel κ; that is, $\mathcal{H}_{X_i}$ is the space spanned by $\{\kappa(\cdot, x_i) : x_i \in \Omega_{X_i}\}$, with its inner product given by $\langle \kappa(\cdot, s_i), \kappa(\cdot, t_i)\rangle_{\mathcal{H}_{X_i}} = \kappa(s_i, t_i)$. In our theoretical developments, we require that all the functions in $\mathcal{H}_{X_i}$ be square-integrable, which is guaranteed by the following assumption:

Assumption 1. $E\{\kappa(X_i, X_i)\} < \infty$ for $i = 1, \ldots, p$.

This condition is satisfied by most of the commonly used kernels, including the radial basis function.

Let $S = (S_1, \ldots, S_k)$ be a subvector of X. The additive reproducing kernel Hilbert space $\mathcal{H}_S$ of functions of S is defined as follows.

Definition 3. *The space $\mathcal{H}_S$ is the direct sum $\bigoplus_{i=1}^k \mathcal{H}_{S_i}$ in the sense that*

$$\mathcal{H}_S \equiv \bigoplus_{i=1}^{k} \mathcal{H}_{S_i} = \{f_1 + \cdots + f_k : f_1 \in \mathcal{H}_{S_1}, \ldots, f_k \in \mathcal{H}_{S_k}\},$$

with inner product $\langle f_1 + \cdots + f_k, g_1 + \cdots + g_k\rangle_{\mathcal{H}_S} = \langle f_1, g_1\rangle_{\mathcal{H}_{S_1}} + \cdots + \langle f_k, g_k\rangle_{\mathcal{H}_{S_k}}$.

Equivalently, $\mathcal{H}_S$ can be viewed as the reproducing kernel Hilbert space generated by the additive kernel $\kappa_S : \Omega_S \times \Omega_S \to \mathbb{R}$, $(s, t) \mapsto \sum_{i=1}^k \kappa(s_i, t_i)$, where $s = (s_1, \ldots s_k)^{\mathrm{T}}$ and $t = (t_1, \ldots, t_k)^{\mathrm{T}}$.

2·3. *Other notation*

For two Hilbert spaces $\mathscr{H}$ and $\mathscr{G}$, we let $\mathscr{B}(\mathscr{H}, \mathscr{G})$ denote the class of all bounded operators from $\mathscr{H}$ to $\mathscr{G}$ and $\mathscr{B}_2(\mathscr{H}, \mathscr{G})$ the class of all Hilbert–Schmidt operators. When $\mathscr{H} = \mathscr{G}$, we denote these classes simply by $\mathscr{B}(\mathscr{H})$ and $\mathscr{B}_2(\mathscr{H})$. The symbols $\|\cdot\|$ and $\|\cdot\|_{\text{HS}}$ stand for the operator and Hilbert–Schmidt norms. For $f \in \mathscr{H}$ and $g \in \mathscr{G}$, the tensor product $g \otimes f$ is the mapping $\mathscr{H} \to \mathscr{G}$, $h \mapsto g\langle f, h\rangle_{\mathscr{H}}$. For two matrices or Euclidean vectors A and B, $A \otimes B$ denotes their Kronecker product. The symbol I stands for the identity mapping in a functional space, whereas I_n means the $n \times n$ identity matrix. The symbol 1_n stands for the vector of length n whose entries are all ones. For an operator $T \in \mathscr{B}(\mathscr{H}, \mathscr{G})$, $\text{null}(T)$ represents the null space of T and $\text{ran}(T)$ the range of T; that is,

$$\text{null}(T) = \{h \in \mathscr{H} : T(h) = 0\}, \quad \text{ran}(T) = \{Th : h \in \mathscr{H}\}.$$

Also, $\overline{\text{ran}}(T)$ stands for the closure of $\text{ran}(T)$.

3. Additive partial covariance operator

3·1. *The additive conditional covariance operator*

The additive conditional covariance operator was proposed by Li et al. (2014) in terms of the $L_2(P_X)$ geometry; here we redefine it in the reproducing kernel Hilbert space geometry.

For subvectors U and V of X, by the Riesz representation theorem there exists a unique operator $\Sigma_{UV} : \mathscr{H}_V \to \mathscr{H}_U$ such that (Conway, 1994, p. 31)

$$\langle f, \Sigma_{UV} g\rangle_{\mathscr{H}_U} = E\{\langle f, \kappa_U(\cdot\,, U)\rangle_{\mathscr{H}_U} \langle g, \kappa_V(\cdot\,, V)\rangle_{\mathscr{H}_V}\} \quad (f \in \mathscr{H}_U,\ g \in \mathscr{H}_V).$$

We define Σ_{UU} and Σ_{VV} similarly. The nonadditive versions of these operators were introduced by Baker (1973) and Fukumizu et al. (2004, 2009). Moreover, by Baker (1973), for any (i, j) there exists a unique operator

$$C_{U_i U_j} \in \mathscr{B}\{\overline{\text{ran}}(\Sigma_{U_i U_i}),\ \overline{\text{ran}}(\Sigma_{U_j U_j})\}$$

such that $\Sigma_{U_i U_j} = \Sigma_{U_i U_i}^{1/2} C_{U_i U_j} \Sigma_{U_j U_j}^{1/2}$. The operator $C_{U_i U_j}$ is the correlation operator between U_i and U_j. Let $D_{UU} = \text{diag}(\Sigma_{U_i U_i} : i = 1, \ldots, a)$ denote the $a \times a$ diagonal matrix of operators whose diagonal entries are the operators $\Sigma_{U_i U_i}$, and let $C_{UU} \in \mathscr{B}(\mathscr{H}_U)$ be the $a \times a$ matrix of operators whose (i, j)th element is $C_{U_i U_j}$. Then it is obvious that $\Sigma_{UU} = D_{UU}^{1/2} C_{UU} D_{UU}^{1/2}$. Notice that $C_{U_i U_i}$ is the identity operator. Define operators such as C_{UW}, C_{WV} and C_{WW} in a similar way. We make the following assumption about the entries of C_{WW}.

Assumption 2. For $i \neq j$, $C_{W_i W_j}$ is a compact operator.

In the Supplementary Material, we show that C_{WW} is invertible and its inverse is bounded. We are now ready to define the additive conditional covariance operator.

Definition 4. *Suppose that Assumptions 1 and 2 hold. Then the operator*

$$\Sigma_{UV|W} = \Sigma_{UV} - D_{UU}^{1/2} C_{UW} C_{WW}^{-1} C_{WV} D_{VV}^{1/2}$$

is called the additive conditional covariance operator of (U, V) *given* W.

Again, this definition also accommodates operators such as $\Sigma_{UU|W}$ and $\Sigma_{VV|W}$.

3·2. *The additive partial correlation operator*

We now introduce the additive partial correlation operator and establish its population-level properties. A straightforward way to define the additive partial correlation operator might be as

$$\Sigma_{UU|W}^{-1/2}\Sigma_{UV|W}\Sigma_{VV|W}^{-1/2}, \tag{4}$$

but caution is needed here because $\Sigma_{UU|W}$ and $\Sigma_{VV|W}$ are Hilbert–Schmidt operators and their eigenvalues tend to zero, so that there is no guarantee that (4) will be well-defined. The following theorem, which echoes Theorem 1 of Baker (1973), shows that (4) is well-defined under minimal conditions.

THEOREM 1. *Suppose that Assumptions 1 and 2 hold. Then there exists a unique operator $\Theta_{UV|W} \in \mathscr{B}(\mathscr{H}_V, \mathscr{H}_U)$ such that:*

(i) $\Theta_{UV|W} = P_{\mathscr{R}_U}\Theta_{UV|W}P_{\mathscr{R}_V}$, *where* $\mathscr{R}_U = \overline{\mathrm{ran}}(\Sigma_{UU|W}^{1/2})$, $\mathscr{R}_V = \overline{\mathrm{ran}}(\Sigma_{VV|W}^{1/2})$, *and* $P_{\mathscr{A}}$ *denotes the projection onto a subspace* $\mathscr{A}$ *in* $L_2(P_X)$*;*
(ii) $\Sigma_{UV|W} = \Sigma_{UU|W}^{1/2}\Theta_{UV|W}\Sigma_{VV|W}^{1/2}$*;*
(iii) $\|\Theta_{UV|W}\| \leqslant 1$.

Theorem 1 justifies the following definition.

DEFINITION 5. *Under Assumptions 1 and 2, the operator $\Theta_{UV|W}$ in Theorem 1 is called the additive partial correlation operator.*

The additive partial correlation operator is defined via a reproducing kernel Hilbert space, whereas additive conditional independence is characterized via $\mathscr{L}_X$. In the Supplementary Material we show that when the kernel function is sufficiently rich that it is a characteristic kernel (Fukumizu et al., 2008, 2009), projections onto $\mathscr{L}_X$ can be well approximated by elements in reproducing kernel Hilbert spaces. Specifically, this requires the following assumption.

Assumption 3. Each $\mathscr{H}_{X_i}$ is a dense subset of $L_2(P_{X_i})$ up to a constant; that is, for each $f \in L_2(P_{X_i})$ there is a sequence $\{f_n\}$ in $\mathscr{H}_{X_i}$ such that $\mathrm{var}\{f_n(X) - f(X)\} \to 0$ as $n \to \infty$.

We are now ready to state the first main result: one can use the additive conditional covariance or additive partial correlation operator to characterize additive conditional independence.

THEOREM 2. *If Assumptions 1–3 hold, then*

$$\Sigma_{UV|W} = 0 \iff \Theta_{UV|W} = 0 \iff U \perp\!\!\!\perp_{\mathrm{A}} V \mid W.$$

3·3. *Estimators*

Here we define sample estimators of $\Sigma_{UV|W}$ and $\Theta_{UV|W}$. Let $(X^1, Y^1), \ldots, (X^n, Y^n)$ be independent copies of (X, Y). Let E_n represent the sample average: $E_n\{f(X, Y)\} = n^{-1}\sum_{\ell=1}^n f(X^\ell, Y^\ell)$. We define the estimate of $\Sigma_{X_iX_j}$ by replacing the expectation with the sample average E_n; that is,

$$\hat{\Sigma}_{X_jX_i} = E_n\left[\{\kappa(\cdot\,, X_j) - \hat{\mu}_{X_j}\} \otimes \{\kappa(\cdot\,, X_i) - \hat{\mu}_{X_i}\}\right], \quad \hat{\mu}_{X_i} = n^{-1}\sum_{\ell=1}^{n}\kappa(\cdot\,, X_i^\ell).$$

Let $\hat{\Sigma}_{XX}$ be the $p \times p$ matrix of operators whose (i, j)th entry is $\hat{\Sigma}_{X_i X_j}$, and let $\hat{\Sigma}_{UU}$, $\hat{\Sigma}_{UV}$ and so forth be the submatrices corresponding to subvectors U and V. Let $\{\epsilon_n\}$ be a sequence of positive constants converging to zero. We define the estimator of $\Sigma_{UV|W}$ as

$$\hat{\Sigma}^{(\epsilon_n)}_{UV|W} = \hat{\Sigma}_{UV} - \hat{\Sigma}_{UW}(\hat{\Sigma}_{WW} + \epsilon_n I)^{-1}\, \hat{\Sigma}_{WV}. \tag{5}$$

Let $\{\delta_n\}$ be another sequence of positive constants converging to zero. We define the estimator of $\Theta_{UV|W}$ as

$$\hat{\Theta}^{(\epsilon_n,\delta_n)}_{UV|W} = (\hat{\Sigma}^{(\epsilon_n)}_{UU|W} + \delta_n I)^{-1/2}\, \hat{\Sigma}^{(\epsilon_n)}_{UV|W}(\hat{\Sigma}^{(\epsilon_n)}_{VV|W} + \delta_n I)^{-1/2}. \tag{6}$$

The tuning parameters ϵ_n and δ_n in (5) and (6) play roles similar to that of the penalty in ridge regression (Hoerl & Kennard, 1970). Technically, they ensure the invertibility of the relevant linear operators and the consistency of the estimators. In practice, they often bring efficiency gains in high dimensions due to their shrinkage effects. Interestingly, as we will see in the next section, δ_n needs to converge to zero more slowly than ϵ_n in order for $\hat{\Theta}^{(\epsilon_n,\delta_n)}_{UV|W}$ to be consistent.

4. Consistency and convergence rate

We first establish consistency of $\hat{\Sigma}^{(\epsilon_n)}_{UV|W}$. Besides serving as an intermediate step for proving the consistency of $\hat{\Theta}^{(\epsilon_n,\delta_n)}_{UV|W}$, the consistency of $\hat{\Sigma}^{(\epsilon_n)}_{UV|W}$ is of interest in its own right, because it was not proved in Li et al. (2014), where it was originally proposed under L_2 geometry. To derive the convergence rate, we need an additional assumption.

Assumption 4. There is an operator $T_{WU} \in \mathscr{B}_2(\mathscr{H}_U, \mathscr{H}_W)$ such that

$$C_{WU} D_{UU}^{1/2} = C_{WW} D_{WW}^{1/2} T_{WU}.$$

The operator T_{WU} also appeared in Lee et al. (2016), where it was called the regression operator because it can be written in the form $\Sigma_{WW}^{-1}\Sigma_{WU}$, resembling the coefficient vector in linear regression. Assumption 4 is essentially a smoothness condition: it requires that the main components in the relation between W and U be sufficiently concentrated on the low-frequency components of the covariance operator Σ_{WW}, in the following sense. If Σ_{WW} is invertible, then Assumption 4 requires $\Sigma_{WW}^{-1}\Sigma_{WU}$ to be a compact operator. Since, under mild conditions, Σ_{WW} is a Hilbert–Schmidt operator (Fukumizu et al., 2007), Σ_{WW}^{-1} is an unbounded operator. Intuitively, in order for $\Sigma_{WW}^{-1}\Sigma_{WU}$ to be compact, the range space of Σ_{WU} should be sufficiently concentrated on the eigenspaces of Σ_{WW} corresponding to its large eigenvalues, or the low-frequency components. As a simple special case of this scenario, in Lee et al. (2016, Proposition 1) it was shown that Assumption 4 is satisfied if the range of Σ_{WU} is a finite-dimensional reducing subspace of Σ_{WW}. This is true, for example, when the polynomial kernel of finite order is used. For kernels inducing infinite-dimensional spaces, Assumption 4 holds if there exist only finitely many eigenfunctions of Σ_{WW} that carry nontrivial correlations with any function in $\mathscr{H}_U$. Of course, these sufficient conditions can be relaxed with careful examination.

We state the consistency of the additive conditional covariance and additive partial correlation operators in the following two theorems, which require different rates for the ridge parameters. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \prec b_n$ if and only if $a_n/b_n \to 0$, and we write $b_n \succ a_n$ if and only if $a_n \prec b_n$.

THEOREM 3. *If Assumptions 1, 2 and 4 are satisfied and $n^{-1/2} \prec \epsilon_n \prec 1$, then*

$$\|\hat{\Sigma}^{(\epsilon_n)}_{UV|W} - \Sigma_{UV|W}\|_{\rm HS} = O_{\rm p}(\epsilon_n^{-1} n^{-1/2} + \epsilon_n^{1/2}), \quad n \to \infty.$$

THEOREM 4. *If Assumptions 1, 2 and 4 are satisfied and*

$$n^{-1/2} \prec \epsilon_n \prec 1, \quad \delta_n \succ (\epsilon_n^{-1} n^{-1/2} + \epsilon_n^{1/2})^{2/3}, \tag{7}$$

then $\|\hat{\Theta}^{(\epsilon_n,\delta_n)}_{UV|W} - \Theta_{UV|W}\|_{\rm HS} \to 0$ in probability as $n \to \infty$.

We return now to the estimation of the additive semigraphoid graphical model in Definition 2. The estimators of the additive conditional covariance operator and additive partial correlation operator lead to the following thresholding methods for estimating the additive semigraphoid model:

$$\hat{\mathcal{E}}_{\rm ACCO} = \{(i, j) : \|\hat{\Sigma}^{(\epsilon_n)}_{X_i X_j | X_{-(i,j)}}\|_{\rm HS} \geqslant \tau_{\rm ACCO}\},$$
$$\hat{\mathcal{E}}_{\rm APCO} = \{(i, j) : \|\hat{\Theta}^{(\epsilon_n,\delta_n)}_{X_i X_j | X_{-(i,j)}}\|_{\rm HS} \geqslant \tau_{\rm APCO}\},$$

where $\tau_{\rm ACCO}$ and $\tau_{\rm APCO}$ are thresholding constants for the additive conditional covariance operator and additive partial correlation operator, respectively. By combining Theorems 2, 3 and 4, it is easy to show that $\hat{\mathcal{E}}_{\rm ACCO}$ and $\hat{\mathcal{E}}_{\rm APCO}$ are consistent estimators of the true edge set $\mathcal{E}$, in the following sense.

THEOREM 5. *Suppose that Assumptions 1–4 hold and X satisfies the additive semigraphoid model in Definition 2 with respect to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Suppose further that ϵ_n and δ_n are positive sequences satisfying (7). Then, for sufficiently small $\tau_{\rm ACCO}$ and $\tau_{\rm APCO}$, as $n \to \infty$,*

$$\text{pr}(\hat{\mathcal{E}}_{\rm ACCO} = \mathcal{E}) \to 1, \quad \text{pr}(\hat{\mathcal{E}}_{\rm APCO} = \mathcal{E}) \to 1.$$

The foregoing asymptotic development is under the assumption that p is fixed as $n \to \infty$. We believe it should be possible to prove the consistency in the setting where $p \to \infty$ as $n \to \infty$, perhaps along the lines of Bickel & Levina (2008a, b). We leave this to future research.

5. IMPLEMENTATION OF ESTIMATION OF GRAPHICAL MODELS

5·1. *Coordinate representation*

The estimators in (5) and (6) are defined in operator form. To compute them, we need to represent the operators as matrices. In the subsequent development we describe this process in the context of estimating the graphical models. We adopt the system of notation for coordinate representation from Horn & Johnson (1985); see also Li et al. (2012). Let $\mathcal{H}$ be a generic n-dimensional Hilbert space with spanning system $\mathcal{B} = \{h_1, \ldots, h_n\}$. For any $f \in \mathcal{H}$, there is a vector $[f]_{\mathcal{B}} \in \mathbb{R}^n$ such that $f = \sum_{i=1}^n ([f]_{\mathcal{B}})_i h_i$. The vector $[f]_{\mathcal{B}}$ is called the coordinate of f with respect to $\mathcal{B}$. Suppose that $\mathcal{H}'$ is another Hilbert space, spanned by $\mathcal{B}' = \{h'_1, \ldots, h'_m\}$, and A is a linear operator from $\mathcal{H}$ to $\mathcal{H}'$. Then the coordinate of A relative to $\mathcal{B}$ and $\mathcal{B}'$ is the matrix $([Ah_1]_{\mathcal{B}'}, \ldots, [Ah_n]_{\mathcal{B}'})$, denoted by ${}_{\mathcal{B}'}[A]_{\mathcal{B}}$. If $\mathcal{H}''$ is a third finite-dimensional Hilbert space, with spanning system $\mathcal{B}''$, and $A' : \mathcal{H}' \to \mathcal{H}''$ is a linear operator, then ${}_{\mathcal{B}''}[A'A]_{\mathcal{B}} = ({}_{\mathcal{B}''}[A']_{\mathcal{B}'})({}_{\mathcal{B}'}[A]_{\mathcal{B}})$. When there is no ambiguity regarding the spanning system used, we abbreviate ${}_{\mathscr{B}'}[A]_{\mathscr{B}}$ to $[A]$, $[f]_{\mathscr{B}}$ to $[f]$, and so on. One can also show that $[A^{\alpha}] = [A]^{\alpha}$ for any $\alpha > 0$. In the rest of this section, square brackets $[\,\cdot\,]$ will be reserved exclusively for the coordinate notation.

5·2. *Norms of the estimated additive partial correlation operator*

For each $i = 1, \ldots, p$, let X_i^r and Y_i^r be the ith components of the vectors X^r and Y^r, respectively. Consider the reproducing kernel Hilbert space

$$\mathscr{H}_{X_i}^{(0)} = \text{span}\{\kappa(\cdot\,, X_i^1), \ldots, \kappa(\cdot\,, X_i^n)\}, \quad \langle \kappa(\cdot\,, X_i^r), \kappa(\cdot\,, X_i^s)\rangle = \kappa(X_i^r, X_i^s).$$

Let $\hat{\mu}_{X_i} = E_n\{\kappa(\cdot\,, X_i)\}$ be the Riesz representation of the linear functional $\Omega_{X_i} \to \mathbb{R}$, $f \mapsto E_n\{f(X)\}$, and let $\phi_i^r = \kappa(\cdot\,, X_i^r) - \hat{\mu}_{X_i}$. For our purposes, it suffices to consider the subspace $\text{span}\{\phi_i^r : r = 1, \ldots, n\}$ of $\mathscr{H}_{X_i}^{(0)}$, because it is the range of operators such as $\hat{\Sigma}_{UV|W}$ and $\hat{\Theta}_{UV|W}$. For this reason, we define this subspace to be $\mathscr{H}_{X_i}$.

Let $K_{X_i} = \{\kappa(X_i^s, X_i^t)\}_{s,t=1}^n$ be the Gram kernel matrix. Let $Q = I_n - 1_n 1_n^{\mathrm{T}}/n$, which is the projection onto the orthogonal complement of $\text{span}(1_n)$ in $\mathbb{R}^n$. Let $G_{X_i} = QK_{X_i}Q$, and let $G_{X_{-(i,j)}}$ be the $n(p-2) \times n$ matrix obtained by removing the ith and jth blocks from the $np \times n$ matrix $(G_{X_1}, \ldots, G_{X_p})^{\mathrm{T}}$. Let $\text{diag}(G_{X_{-(i,j)}})$ be the $\{n(p-2) \times n(p-2)\}$-dimensional block-diagonal matrix whose diagonal blocks are the $p-2$ blocks of $G_{X_{-(i,j)}}$, each of dimension $n \times n$. To avoid complicated notation, throughout this subsection we write the estimated operators $\hat{\Sigma}^{(\epsilon_n)}_{X_iX_j|X_{-(i,j)}}$ and $\hat{\Theta}^{(\epsilon_n,\delta_n)}_{X_iX_j|X_{-(i,j)}}$ as simply $\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}$ and $\hat{\Theta}_{X_iX_j|X_{-(i,j)}}$.

By straightforward calculations, details of which are given in the Supplementary Material, we have the following coordinate representations:

$$\begin{gathered}
\big[\hat{\Sigma}_{X_iX_j}\big] = n^{-1}G_{X_j}, \quad \big[\hat{\Sigma}_{X_iX_i}\big] = n^{-1}G_{X_i}, \quad \big[\hat{\Sigma}_{X_iX_{-(i,j)}}\big] = n^{-1}G_{X_{-(i,j)}}, \\
\big[\hat{\Sigma}_{X_{-(i,j)}X_i}\big] = n^{-1}(1_{p-2} \otimes G_{X_i}), \quad \big[\hat{\Sigma}_{X_{-(i,j)}X_{-(i,j)}}\big] = n^{-1}\{1_{p-2} \otimes G^{\mathrm{T}}_{X_{-(i,j)}}\}.
\end{gathered} \tag{8}$$

Let $H_{X_{-(i,j)}} = G^{\mathrm{T}}_{X_{-(i,j)}}[G_{X_{-(i,j)}}G^{\mathrm{T}}_{X_{-(i,j)}} + \epsilon_n \,\text{diag}\{G_{X_{-(i,j)}}\}]^{\dagger}\, G_{X_{-(i,j)}}$, where † indicates the Moore–Penrose inverse. Therefore, $[\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}]$ is equal to $n^{-1}\{I_n - H_{X_{-(i,j)}}\}G_{X_j}$ and is denoted by $R_{X_j|X_{-(i,j)}}$. Similarly, $[\hat{\Sigma}_{X_jX_i|X_{-(i,j)}}] = R_{X_i|X_{-(i,j)}}$. Then we can compute $\|\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}\|^2_{\mathrm{HS}}$ via

$$\|\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}\|^2_{\mathrm{HS}} = \text{tr}\big\{R_{X_i|X_{-(i,j)}}R_{X_j|X_{-(i,j)}}\big\}. \tag{9}$$

In the Supplementary Material, we also derive an explicit formula for calculating $\|\hat{\Theta}_{X_iX_j|X_{-(i,j)}}\|^2_{\mathrm{HS}}$. Let $\tilde{R}_{X_i|X_{-(i,j)}} = \{R_{X_i|X_{-(i,j)}} + \delta_n I_n\}^{-1}R_{X_i|X_{-(i,j)}}$ and $\tilde{R}_{X_j|X_{-(i,j)}} = \{R_{X_j|X_{-(i,j)}} + \delta_n I_n\}^{-1}R_{X_j|X_{-(i,j)}}$. Then we have

$$\|\hat{\Theta}_{X_iX_j|X_{-(i,j)}}\|^2_{\mathrm{HS}} = \text{tr}\big\{\tilde{R}_{X_i|X_{-(i,j)}}\tilde{R}_{X_j|X_{-(i,j)}}\big\}. \tag{10}$$

The following result links the additive partial correlation operator with the partial correlation when a linear kernel is considered.

Corollary 1. *Let $\kappa(x_s, x_t) = 1 + x_sx_t$. Then, as $n \to \infty$, $\|\hat{\Theta}_{X_iX_j|X_{-(i,j)}}\|_{\mathrm{HS}}$ converges in probability to the absolute value of the partial correlation between X_i and X_j given $X_{-(i,j)}$.*

5·3. *Reduced kernel and generalized crossvalidation*

To make our method readily applicable to relatively large networks with thousands of nodes, we now propose, as alternatives to (9) and (10), simplified algorithms for estimating $\|\Sigma_{X_iX_j|X_{-(i,j)}}\|_{\mathrm{HS}}$ and $\|\Theta_{X_iX_j|X_{-(i,j)}}\|_{\mathrm{HS}}$. Lower-frequency eigenfunctions of kernels often play dominant roles, and the numbers of statistically significant eigenvalues of kernel matrices are often much smaller than n; see, for example, Lee & Huang (2007) and Chen et al. (2010). By employing only the dominant low-frequency eigenfunctions, we can greatly reduce the amount of computation without incurring much loss of accuracy. Let the eigendecomposition of the kernel matrix G_{X_i} be written as

$$G_{X_i} = V_{X_i}\Lambda_{X_i}V_{X_i}^{\mathrm{T}} + \tilde{V}_{X_i}\tilde{\Lambda}_{X_i}\tilde{V}_{X_i}^{\mathrm{T}}, \tag{11}$$

where $V_{X_i}\Lambda_{X_i}V_{X_i}^{\mathrm{T}}$ corresponds to the first m_i eigenvalues of G_{X_i} and $\tilde{V}_{X_i}\tilde{\Lambda}_{X_i}\tilde{V}_{X_i}^{\mathrm{T}}$ corresponds to the last $n - m_i$ eigenvalues. Instead of the original bases $\{\phi_i^r\}_{r=1}^n$, we now work with $\mathrm{span}\{\psi_i^1, \ldots, \psi_i^{m_i}\}$, where $(\psi_i^1, \ldots, \psi_i^{m_i})^{\mathrm{T}} = \Lambda_{X_i}^{-1/2}V_{X_i}^{\mathrm{T}}\phi_i$ and will be written simply as ψ_i.

Let $M_i = (V_{X_i}\Lambda_{X_i}^{1/2})^{\mathrm{T}}$, let $M_{\mathrm{V}} = (M_1, \ldots, M_p)$, let $M_{-(i,j)}$ be the matrix obtained by removing M_i and M_j from M_{V}, and let $N_{-(i,j)}(\epsilon_n) = M_{-(i,j)}\{M_{-(i,j)}^{\mathrm{T}}M_{-(i,j)} + \epsilon_n I_{m_{-(i,j)}}\}^{-1}M_{-(i,j)}^{\mathrm{T}}$, where $m_{-(i,j)} = \sum_{k \neq i,j} m_k$. Using derivations similar to (8) and (10), we find the coordinate representation of the additive conditional covariance operator with respect to the new basis $\{\psi_i\}$ as

$$_{\psi_i}[\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}]_{\psi_j} = n^{-1}M_i^{\mathrm{T}}\{I_n - N_{-(i,j)}(\epsilon_n)\}M_j, \tag{12}$$

which is denoted by $O_{i,j|-(i,j)}$. Correspondingly, the Hilbert–Schmidt norms of the additive conditional covariance operator and the additive partial correlation operator can be computed via

$$\begin{aligned} \|\hat{\Sigma}_{X_iX_j|X_{-(i,j)}}\|_{\mathrm{HS}}^2 &= \|O_{i,j|-(i,j)}\|_{\mathrm{F}}^2, \\ \|\hat{\Theta}_{X_iX_j|X_{-(i,j)}}\|_{\mathrm{HS}}^2 &= \|\{O_{i,i|-(i,j)} + \delta_n I_{m_i}\}^{-1/2}O_{i,j|-(i,j)}\{O_{j,j|-(i,j)} + \delta_n I_{m_j}\}^{-1/2}\|_{\mathrm{F}}^2, \end{aligned} \tag{13}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius matrix norm. In (12) we need to invert an $m_{-(i,j)} \times m_{-(i,j)}$ matrix $M_{-(i,j)}^{\mathrm{T}}M_{-(i,j)} + \epsilon_n I_{m_{-(i,j)}}$, which could be large if p is large. However, as shown in Proposition 4 of Li et al. (2014), calculation of this matrix can be reduced to the eigendecomposition of an $n \times n$ matrix.

For the choice of m_i, we follow Fan et al. (2011) and determine it adaptively according to the sample size n. Specifically, we take

$$m_i = 3 \times [n^{1/5}] = O(n^{1/5}). \tag{14}$$

We use the reduced kernel bases consistently for all the simulations and the real-data analysis in § 6. Based on our experience, using the reduced bases not only cuts the computation time substantially but also gives very high accuracy compared with using the full bases.

Next, we introduce a generalized crossvalidation procedure to choose the thresholds τ_{ACCO} and τ_{APCO}. Our process roughly follows Li et al. (2014). Given $\tau > 0$, let $\hat{\mathcal{E}}(\tau)$ be the estimated graph by either criterion in (13), and define the neighbours of node $i \in \mathcal{V}$ as $\mathcal{V}_i(\tau) = \{j \in \mathcal{V} : (i, j) \in \hat{\mathcal{E}}(\tau)\}$. Our strategy is to regress each node on its neighbours and obtain the residuals; the generalized crossvalidation criterion is then used to minimize the total prediction error. Specifically,

τ is determined by minimizing

$$\textsc{gcv}(\tau) = \sum_{i=1}^{p} \frac{\|\{I_n - N_{\mathcal{V}_i(\tau)}(\epsilon_{n,i})\}M_i\|^2}{[1 - \mathrm{tr}\{N_{\mathcal{V}_i(\tau)}(\epsilon_{n,i})\}/n]^2}, \tag{15}$$

where the $\epsilon_{n,i}$ are chosen differently for each node, as shown in the next subsection.

5·4. *Algorithm*

The following algorithm summarizes the estimating procedure for the additive semigraphoid model based on the estimated additive partial correlation operator and the estimated additive conditional covariance operator.

Step 1. For each $i = 1, \ldots, p$, standardize $(X_i^1, \ldots, X_i^n)$ such that $E_n(X_i) = 0$ and $\mathrm{var}_n(X_i) = 1$.

Step 2. Select the kernel κ, for example as the radial basis function $\kappa(X_i^s, X_i^t) = \exp(-\gamma_i |X_i^s - X_i^t|^2)$ where γ_i is the bandwidth parameter. As in Lee et al. (2013), we recommend choosing γ_i via

$$\gamma_i = \binom{n}{2}^2 \left(\sum_{s<t} |X_i^s - X_i^t| \right)^{-2}.$$

Step 3. Use the selected κ and γ_i to compute the kernel matrix K_{X_i}, its centred counterpart G_{X_i}, and the eigendecomposition (11) for $i = 1, \ldots, p$. Choose m_i according to (14).

Step 4. Determine the tuning parameters ϵ_n, δ_n and $\epsilon_{n,i}$ to be the fractions of the largest singular values of relevant matrices to be penalized. That is, let

$$\begin{aligned}
\epsilon_n &= c_1 \times \max\left[\sigma_{\max}\{M_{-(i,j)} M_{-(i,j)}^{\mathrm{T}}\} : 1 \leqslant i < j \leqslant p\right],\\
\delta_n &= c_2 \times \max\left(\max[\sigma_{\max}\{O_{i,i|-(i,j)}\}, \sigma_{\max}\{O_{j,j|-(i,j)}\}] : 1 \leqslant i < j \leqslant p\right),\\
\epsilon_{n,i} &= c_3 \times \sigma_{\max}\{M_{\mathcal{V}_i(\tau)} M_{\mathcal{V}_i(\tau)}^{\mathrm{T}}\} \quad (i = 1, \ldots, p),
\end{aligned}$$

where $\sigma_{\max}$ denotes the largest singular value of a matrix. The constants c_1, c_2 and c_3 control the smoothing effects; we fix $c_2 = 0{\cdot}01$ and let c_1 and c_3 be $n^{-1/5}$ based on a criterion similar to that used in Step 3. Finally, to further simplify the computation, we can approximate ϵ_n by $\epsilon_n = \sigma_{\max}(M_{\mathcal{V}} M_{\mathcal{V}}^{\mathrm{T}})$.

Step 5. For each $i > j$, calculate $\|\hat{\Sigma}_{X_i X_j | X_{-(i,j)}}\|_{\textsc{hs}}$ or $\|\hat{\Theta}_{X_i X_j | X_{-(i,j)}}\|_{\textsc{hs}}$ using (9) and (10) or their fast versions given in (13).

Step 6. Compute the thresholds that minimize (15), and determine the graph using either of the two criteria. For example, if $\hat{\tau}_{\textsc{apco}}$ is the best threshold, then remove (i, j) from the edge set $\mathcal{E}$ if $\|\hat{\Theta}_{X_i X_j | X_{-(i,j)}}\|_{\textsc{hs}} < \hat{\tau}_{\textsc{apco}}$.

6. Numerical study

6·1. *Additive and high-dimensional settings*

By means of simulated examples, we compare the additive partial correlation operator with the additive conditional covariance operator of Li et al. (2014) and the methods of Yuan & Lin (2007), Liu et al. (2009), Fellinghauer et al. (2013) and Voorman et al. (2014). The additive partial correlation operator is able to identify the graph whose underlying distribution does not satisfy the Gaussian or copula Gaussian assumption. To demonstrate this feature, we generate dependent random variables that do not have Gaussian or copula Gaussian distributions using the structural equation models of Pearl (2009). Specifically, given an edge set $\mathcal{E}$, we generate $(X_1, \ldots, X_p)$ sequentially via

$$X_i = f_i[\{X_j : (i, j) \in \mathcal{E}\},\ \varepsilon_i] \quad (i = 1, \ldots, p),$$

where $f_i(\cdot)$ is the link function and $\varepsilon_1, \ldots, \varepsilon_p$ are independent and identically distributed standard Gaussian variables. If f_i is linear, the joint distribution is Gaussian; otherwise, the joint distribution may be neither Gaussian nor copula Gaussian.

We consider the following graphical models based on three choices of f_i.

Model I: $f_i[\{x_j : (i, j) \in \mathcal{E}\}, \varepsilon] = \sum_{(i,j)\in \mathrm{E}} x_j + \varepsilon_i$.
Model II: $f_i[\{x_j : (i, j) \in \mathcal{E}\}, \varepsilon_i] = \sum_{(i,j)\in \mathrm{E}} (1 + |x_j|)^2 + \varepsilon_i$.
Model III: $f_i[\{x_j : (i, j) \in \mathcal{E}\}, \varepsilon_i] = \{\sum_{(i,j)\in \mathrm{E}} x_j\}\, \varepsilon_i$.

The sample sizes are taken to be $n = 50$ and 100, and the number of nodes is $p = 200$.

We use the hub structure to generate the underlying graphs and the corresponding edge sets $\mathcal{E}$. Hubs are commonly observed in networks such as gene regulatory networks and citation networks; see Newman (2003). Specifically, given a graph of size p, ten independent hubs are generated so that each module is of degree $p/10 - 1$. For each of the $3 \times 2 = 6$ combinations, we generate 100 samples and produce the averaged receiver operating characteristic curves and the areas under these curves. To draw the curves, we need to compute the false positive and true positive rates. Suppose $\hat{\mathcal{E}}$ is an estimate of $\mathcal{E}$; then the formal definitions of these two measures are

$$\mathrm{TP} = \frac{\sum_{1\leqslant j<i\leqslant p} I\{(i, j) \in \mathcal{E}, (i, j) \in \hat{\mathcal{E}}\}}{\sum_{1\leqslant j<i\leqslant p} I\{(i, j) \in \mathcal{E}\}}, \qquad \mathrm{FP} = \frac{\sum_{1\leqslant j<i\leqslant p} I\{(i, j) \notin \mathcal{E}, (i, j) \in \hat{\mathcal{E}}\}}{\sum_{1\leqslant j<i\leqslant p} I\{(i, j) \notin \mathcal{E}\}}. \tag{16}$$

The receiver operating characteristic curves are plotted in Fig. 1.

For all the comparisons in § § 6·1 and 6·2, we use the radial basis function for both the additive conditional covariance and the additive partial correlation operators. For Model I, we see that the methods of Yuan & Lin (2007) and Liu et al. (2009) perform better than the nonparametric methods. This is not surprising as Gaussianity holds under Model I, and because both methods use the L_1 penalty, which is more efficient than thresholding. Nevertheless, the performance of the additive partial correlation operator is not far behind. For example, the areas under the receiver operating characteristic curves for the additive partial correlation operator have an average of 0·98 for the two curves in Model I, only slightly smaller than the average of the areas under the curves for the methods of Yuan & Lin (2007) and Liu et al. (2009), which is 1·00.

For Models II and III, under which neither Gaussianity nor copula Gaussianity is satisfied, the methods of Yuan & Lin (2007) and Liu et al. (2009) do not perform well. In contrast, both the additive conditional covariance and the additive partial correlation operators still perform remarkably well. Moreover, the receiver operating characteristic curves of the additive partial correlation operator are consistently better than those of the additive conditional covariance operator for Models I and II and for sample sizes 50 and 100, indicating the benefit of a better scaling by the additive partial correlation operator. We also observe that the performance of the method of Fellinghauer et al. (2013) is not very stable. Since their method is based on random forests, it may be affected by the curse of dimensionality that a fully nonparametric approach tends to suffer from. The method of Voorman et al. (2014) is implemented using the R package (R Development Core Team, 2016) spacejam, whose default basis is the cubic polynomial. It shows improvements

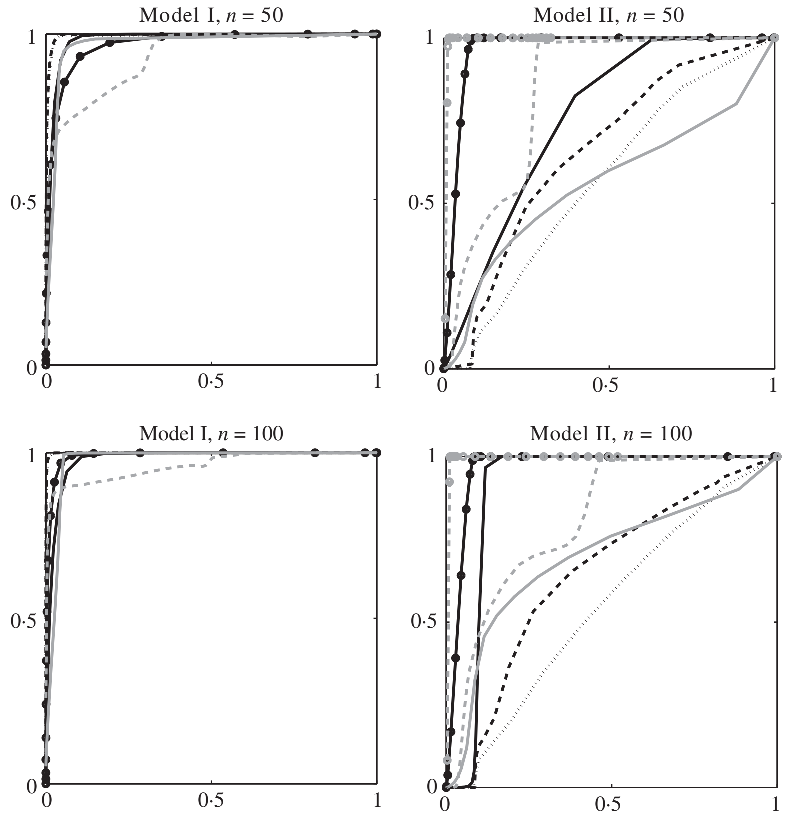

Fig. 1. Receiver operating characteristic curves for different estimators: the additive partial correlation operator (—●—); the additive conditional covariance operator (black ——); the method of Yuan & Lin (2007) (black - - - -); the method of Liu et al. (2009) (· · · · · ·); the method of Fellinghauer et al. (2013) (grey ——); and the method of Voorman et al. (2014) with default basis (grey - - - -). The two middle panels also display the method of Voorman et al. (2014) with the correct basis $\{|x|, x^2\}$ for Model II (grey - -●- -). In each panel the horizontal axis shows the false positive rate and the vertical axis the true positive rate.

over the methods of Yuan & Lin (2007) and Liu et al. (2009), but does not perform as well as the additive partial correlation operator. To investigate the effect of the choice of basis on the method of Voorman et al. (2014), we compute its receiver operating characteristic curve for Model II using the correct basis $\{|x|, x^2\}$. Notably, this method with the correct basis performs the best under Model II among all the competing methods. Results for smaller graphs are presented in the Supplementary Material.

6·2. *Nonadditive and low-dimensional settings*

We also investigate a setting where the relationships between nodes are nonadditive and the dimension of the graph is relatively low, which favours a fully nonparametric method such as the method of Fellinghauer et al. (2013). Specifically, we consider

$$\text{Model IV:}\quad X_1 = \varepsilon_1,\quad X_2 = X_1 + \varepsilon_2,\quad X_3 = \sin(2X_1X_2) + \varepsilon_3,\quad X_4 = \cos(2X_1X_2X_3) + \varepsilon_4,$$
$$X_5 = X_3 + \varepsilon_5, X_6 = X_4 + \varepsilon_6,$$

where $(\varepsilon_1, \ldots, \varepsilon_6)$ are independent and identically distributed standard Gaussian variables.

Our goal is to recover the graph determined by the set of conditional independence relations $X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}$ whenever $(i, j) \notin \mathcal{E}$. Under Model IV, the edge set is

$\mathcal{E} = \{(2, 1), (3, 1), (3, 2), (4, 1), (4, 2), (4, 3), (5, 3), (6, 4)\}$. The graphical model based on pairwise conditional independence cannot fully describe the interdependence in X, because it cannot capture three-way or multi-way conditional dependence. A fully descriptive approach in such situations would be to use a hypergraph (Lauritzen, 1996, p. 21). Nevertheless, the pairwise conditional independence graphical model is well-defined and helps to illustrate the difference between an additive and a fully nonparametric model.

Taking $n = 1000$, we compute the receiver operating characteristic curves for 100 replicates, which are presented in the Supplementary Material. Since the model is nonlinear, we only compare the additive partial correlation operator with the additive conditional covariance operator and the methods of Voorman et al. (2014) and Fellinghauer et al. (2013). The method of Fellinghauer et al. (2013) performs the best, because it allows the conditional expectation of each node to be a nonadditive function of its neighbouring nodes. On the other hand, the additive partial correlation operator still performs reasonably well. This indicates that, in spite of its additive formulation, the additive partial correlation operator is capable of identifying conditional independence even in nonadditive models.

6·3. *Effects of the choices of kernels, ridge parameters and number of eigenfunctions*

In this subsection we study the performance of the additive partial correlation operator with different choices of kernel. We investigate six types of kernel: the radial basis function, the rational quadratic kernel with parameters 200 and 400, the linear kernel, the quadratic kernel, and the Laplacian kernel. The choice of parameters for the rational quadratic kernel follows Li et al. (2014). For each model, ten replicates are generated using $p = 200$ and $n = 100$. The averaged receiver operating characteristic curves for the six kernels are presented in the Supplementary Material. The results suggest that all the nonlinear kernels give comparable performance across Models I, II and III. As expected, the linear kernel fails for Models II and III.

Next, we investigate the sensitivity of the proposed estimator to the tuning parameters ϵ_n and δ_n. We take 20 equally spaced grid points in each of the ranges $(10^{-1}\hat{\epsilon}_n, 10\hat{\epsilon}_n)$ and $(10^{-1}\hat{\delta}_n, 10\hat{\delta}_n)$, with $\hat{\epsilon}_n$ and $\hat{\delta}_n$ computed via the empirical formulas in § 5·4. Then, for each of the $20 \times 20 = 400$ combinations, a receiver operating characteristic curve is produced and its area under the curve is computed. The means of the areas for Models I, II and III are 0·995, 0·956 and 0·971, respectively, with standard deviations of 0·004, 0·004 and 0·011. These values indicate that the performance of the proposed estimator is reasonably robust with respect to the choice of tuning parameters. The actual receiver operating characteristic curves for different combinations of tuning parameters are plotted in the Supplementary Material.

We also investigated the effect of using different numbers of eigenfunctions. For each of Models I–III and with $(p, n) = (200, 100)$, we increase m_i from 1 to $n - 1$ and, for each fixed m_i, produce a receiver operating characteristic curve and compute the area under the curve. The areas under the curves are reported in the Supplementary Material. The results show that the effect of using a different number of eigenfunctions varies across the three models, which is to be expected as they have different complexities. Specifically, a single eigenfunction achieves the largest area under the curve for the linear model, but for the nonlinear models the optimal areas are achieved when more eigenfunctions are used. We also see that our choices of m_i are not far from the best choice for all three models.

6·4. *Exploring the generalized crossvalidation procedure*

In this subsection we investigate the performance and computational cost of the generalized crossvalidation procedure introduced in § 5·3, and compare it with the method of Voorman et al. (2014) using two different selection criteria: the Akaike information criterion and the Bayesian

Table 1. *Comparison of the tuning procedures for* (a) *the additive partial correlation operator with generalized crossvalidation and* (b) *the method of Voorman et al.* 2014 *with the Akaike information criterion;* TP *and* FP (%) *are defined in* (16), *and* DIS *is defined in* (17)*; larger* TP, *smaller* FP *and lower* DIS *indicate better performance*

	$p = 50$			$p = 100$			$p = 200$		
	TP	FP	DIS	TP	FP	DIS	TP	FP	DIS
(a)	97·1	10·9	0·11	98·1	17·8	0·18	98·7	22·1	0·22
(b)	58·8	24·7	0·61	90·0	71·6	0·72	78·6	51·5	0·56

Table 2. *Comparison of computing times for* (a) *the additive partial correlation operator with generalized crossvalidation and* (b) *the method of Voorman et al.* 2014 *with the Akaike information criterion. All experiments were conducted on an Intel Xeon E5520 with* 2·26 GHz *CPU*

(n, p)	(50, 1000)		(100, 1000)		(50, 5000)		(100, 5000)	
	Minutes	DIS	Minutes	DIS	Minutes	DIS	Minutes	DIS
(a)	2·3	0·35	8·6	0·23	58·4	0·59	213·3	0·54
(b)	47·7	0·87	122·8	0·68	553·4	0·97	1572·8	0·96

information criterion. Three measures are used to evaluate the comparisons: the true positive and false positive rates in (16), and a synthetic score defined as

$$\text{DIS} = \{\text{FP}^2 + (1 - \text{TP})^2\}^{1/2}. \tag{17}$$

Table 1 shows the averages of these criteria over 100 replicates using Model III with $p = 50$, 100, 200 and $n = 100$. We omit the result obtained from the method of Voorman et al. (2014) using the Bayesian information criterion, because the Akaike information criterion for the same method always performs better in this setting. Our procedure consistently picks up the thresholds located around the best scenario. In comparison, the method of Voorman et al. (2014) with the Akaike information criterion does not perform as well as our estimator.

We also compare the computational costs of the two methods in estimating larger networks with $p = 1000$ or 5000. For the tuning parameters, 40 grid points are used for both the additive partial correlation operator and the method of Voorman et al. (2014). The results are reported in Table 2. With $p = 1000$, our algorithm takes only minutes to complete, and for $p = 5000$ it is still reasonably efficient. In terms of estimation accuracy, our method has smaller DIS than the method of Voorman et al. (2014). The complexity of the additive partial correlation operator grows as p^2. However, for handling graphs with thousands of nodes, our method is faster than the regression-based approaches.

6·5. *Application to the DREAM4 Challenges data*

We apply the six methods to a dataset from the DREAM4 Challenges project (Marbach et al., 2010). The goal of this study is to infer network structure from gene expression data. The topologies of the graphs are obtained by extracting subgraphs from real biological networks. The gene expression levels are generated based on a system of ordinary differential equations governing the dynamics of the biological interactions between the genes. There are five networks of size 100 to be estimated in this dataset. For each network, we stack up observations from three different experimental conditions, wild-type, knockdown and knockout, so that the overall sample size

Table 3. *Areas under the receiver operating characteristic curves for the DREAM*4 *Challenges dataset, obtained from* (a) *the additive partial correlation operator,* (b) *the additive conditional covariance operator,* (c) *the method of Voorman et al. (2014),* (d) *the method of Fellinghauer et al.* 2013*,* (e) *the method of Liu et al.* 2009*,* (f) *the method of Yuan & Lin* 2007*, and* (g) *the championship method*

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Network 1	0·86	0·67	0·79	0·73	0·61	0·74	0·91
Network 2	0·81	0·62	0·70	0·64	0·57	0·70	0·81
Network 3	0·83	0·70	0·77	0·68	0·64	0·73	0·83
Network 4	0·83	0·71	0·76	0·71	0·61	0·72	0·83
Network 5	0·77	0·66	0·70	0·73	0·61	0·70	0·75

n is 201. Then, the estimated graphs are produced using the additive partial correlation operator, the additive conditional covariance operator of Li et al. (2014), and the methods of Voorman et al. (2014), Fellinghauer et al. (2013), Yuan & Lin (2007) and Liu et al. (2009). The areas under the receiver operating characteristic curves are reported in Table 3, and the actual receiver operating characteristic curves are displayed in the Supplementary Material. We see that the additive partial correlation operator consistently performs best among the six estimators.

The original DREAM4 project was open to public challenges, so it is reasonable to compare our results with those submitted by the participating teams. In column (g) of Table 3 we show the areas under the receiver operating characteristic curves obtained from the method of the championship team. The additive partial correlation operator yields the best areas under the curves for four of the five networks; in particular, it performs better than the method of the championship team for Network 5. As mentioned in Marbach et al. (2010), the best-performing approach used a combination of multiple models, including ordinary differential equations. Our operator replicates the most competitive results without employing any prior information on the model setting, which demonstrates the benefit of relaxing the distributional assumption; moreover, its additive structure does not seem to hamper its accuracy in this application.

7. Concluding remarks

In establishing the consistency of the additive conditional covariance operator and the additive partial correlation operator, we have developed a theoretical framework that is not limited to the current setting; it can be applied to other problems where additive conditional independence and linear operators are involved. Moreover, the idea of characterizing conditional independence by small values of the additive partial correlation operator has ramifications beyond those explored in this paper. For instance, the penalty in the proposed additive partial correlation operator is based on hard thresholding, but other penalties, such as the lasso-type penalties, may be more efficient in dealing with sparsity in the estimation of operators. We leave these extensions and refinements to future research.

Acknowledgement

We are grateful to three referees for their constructive comments and helpful suggestions. Bing Li’s research was supported in part by the U.S. National Science Foundation; Hongyu Zhao’s research was supported in part by both the National Science Foundation and the National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs of the theoretical results and additional plots for the numerical studies.

References

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9**, 1179–225.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Trans. Am. Math. Soc.* **186**, 273–89.

Banerjee, O., Ghaoui, El, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton: Princeton University Press.

Bickel, P. J. & Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.

Bickel, P. J. & Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

Chen, P.-C., Lee, K.-Y., Lee, T.-J., Lee, Y.-J. & Huang, S.-Y. (2010). Multiclass support vector classification via coding and regression. *Neurocomputing* **73**, 1501–12.

Conway, J. B. (1994). *A Course in Functional Analysis*. New York: Springer, 2nd ed.

Fan, J., Feng, Y. & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Statist. Assoc.* **106**, 544–57.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fellinghauer, B., Bühlmann, P., Ryffelb, M., Rheinc, M. & Reinhardta, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comp. Statist. Data Anal.* **64**, 132–52.

Friedman, J. H., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Fukumizu, K., Bach, F. R. & Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.* **8**, 361–83.

Fukumizu, K., Bach, F. R. & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5**, 73–99.

Fukumizu, K., Bach, F. R. & Jordan, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37**, 1871–905.

Fukumizu, K., Gretton, A., Sun, X. & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Adv. Neural Info. Proces. Syst.* **20**, 489–96.

Harris, N. & Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14**, 3365–83.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Horn, R. A. & Johnson, C. R. (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.

Lam, C. & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–78.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Lee, K.-Y., Li, B. & Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Statist.* **41**, 221–49.

Lee, K.-Y., Li, B. & Zhao, H. (2016). Variable selection via additive conditional independence. *J. R. Statist. Soc.* B to appear, doi:10.1111/rssb.12150.

Lee, Y.-J. & Huang, S.-Y. (2007). Reduced support vector machines: A statistical theory. *IEEE Trans. Neural Networks* **18**, 1–13.

Li, B., Chun, H. & Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Am. Statist. Assoc.* **107**, 152–67.

Li, B., Chun, H. & Zhao, H. (2014). On an additive semi-graphoid model for statistical networks with application to pathway analysis. *J. Am. Statist. Assoc.* **109**, 1188–204.

Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–326.

Liu, H., Lafferty, J. & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295–328.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D. & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Nat. Acad. Sci.* **107**, 6286–91.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

MUIRHEAD, R. J. (2005). *Aspects of Multivariate Statistical Theory*. New York: Wiley, 2nd ed.

NEWMAN, M. (2003). The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.

PEARL, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2nd ed.

PEARL, J., GEIGER, D. & VERMA, T. (1989). Conditional independence and its representations. *Kybernetika* **25**, 33–44.

PEARL, J. & VERMA, T. (1987). The logic of representing dependencies by directed graphs. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, vol. 1. AAAI Press, pp. 374–9.

PENG, J., WANG, P., ZHOU, N. & ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Assoc.* **104**, 735–46.

R DEVELOPMENT CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

VOORMAN, A., SHOJAIE, A. & WITTEN, D. (2014). Graph estimation with joint additive models. *Biometrika* **101**, 85–101.

XUE, L. & ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40**, 2541–71.

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11**, 2261–86.

YUAN, M. & LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

[*Received January* 2015. *Revised June* 2016]