Published in final edited form as:

J Am Stat Assoc. 2014 September ; 109(507): 1188–1204. doi:10.1080/01621459.2014.882842.

On an Additive Semigraphoid Model for Statistical Networks With Application to Pathway Analysis

Bing Li¹, Hyonho Chun², and Hongyu Zhao³

Bing Li: bing@stat.psu.edu; Hyonho Chun: chunh@purdue.edu; Hongyu Zhao: hongyu.zhao@yale.edu

¹Professor of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802

²Assistant Professor of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907

³Professor of Biostatistics, Yale University, Suite 503, 300 George Street, New Haven, CT 06410

Abstract

We introduce a nonparametric method for estimating non-gaussian graphical models based on a new statistical relation called additive conditional independence, which is a three-way relation among random vectors that resembles the logical structure of conditional independence. Additive conditional independence allows us to use one-dimensional kernel regardless of the dimension of the graph, which not only avoids the curse of dimensionality but also simplifies computation. It also gives rise to a parallel structure to the gaussian graphical model that replaces the precision matrix by an additive precision operator. The estimators derived from additive conditional independence cover the recently introduced nonparanormal graphical model as a special case, but outperform it when the gaussian copula assumption is violated. We compare the new method with existing ones by simulations and in genetic pathway analysis.

Key words and phrases

Additive conditional independence; additive precision operator; copula; conditional independence; covariance operator; gaussian graphical model; nonparanormal graphical model; reproducing kernel Hilbert space

1 Introduction

The gaussian graphical model is one of the most important tools for statistical network analysis, partly because its estimation can be neatly formulated as identifying the zero entries in the precision matrix, which turns estimation of a graph into sparse estimation of a high-dimensional precision matrix. For recent advances in this area, see Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Bickel and Levina (2008), Peng, Wang, Zhou, and Zhu (2009), Guo, Levina, Michailidis, and Zhu (2009), Lam and Fan (2009).

Specifically, let $X = (X^1, \dots, X^p)^\top$ be a p -dimensional random vector distributed as $N(\mu, \Sigma)$, where Σ is positive definite. Let $\Gamma = \{1, \dots, p\}$, $E \subseteq \{(i, j) \in \Gamma \times \Gamma : i \neq j\}$, and $\Theta = \Sigma^{-1}$. Let θ_{ij} be the (i, j) th element of Θ . Then X follows a gaussian graphical model (GGM) with respect to the graph $\mathcal{G} = \{\Gamma, E\}$ iff

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

where $X^{-(i,j)}$ represents the set $\{X^k : k \neq i, j\}$. Under the multivariate gaussian assumption, the above relation is equivalent to

$$\theta_{ij}=0, \quad (i,j) \notin E. \quad (2)$$

Based on this equivalence the estimation of $\mathcal{G}$ reduces to identification of the 0 entries in Θ , which can be done by sparse estimation procedures such as thresholding (Bickel and Levina, 2008), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and adaptive LASSO (Zuo, 2006).

Since the multivariate gaussian assumption limits the applicability of the GGM, there have been intense developments in generalizing it to non-gaussian cases in the past few years, chiefly through the use of nonparametric copula transformations. See Liu, Lafferty, and Wasserman (2009), Liu, Han, Yuan, Lafferty, and Wasserman (2013), and Xue and Zou (2013). The nonparametric gaussian copula graphical model (GCGM) is a flexible and effective model that relaxes the marginal gaussian assumption but preserves the inter-dependence structure of a multivariate gaussian distribution, so that the equivalence between (1) and (2) still holds in an altered form. It assumes the existence of unknown injections $f_1, \dots, f_p$ such that $(f_1(X^1), \dots, f_p(X^p))^T$ is multivariate gaussian. The GCGM keeps the simplicity and elegance of the GGM but greatly expands its scope.

However, the gaussian copula assumption can be violated even for commonly used interactions. For example, if (U, V) is a random vector where $V = U^2 + \varepsilon$, and U and ε are i.i.d. $N(0, 1)$, then no transformation can render $(f_1(U), f_2(V))$ bivariate gaussian. In fact, the copula transformations in this case are: $f_1(u) = u, f_2(v) = \Phi^{-1} \circ F_V(v)$, where F_V is the convolution of $N(0, 1)$ and χ_1^2 and Φ is the c.d.f. of $N(0, 1)$. Figure 1 plots $(U, \Phi^{-1} \circ F_V(V))$ for a sample of 100 observed pairs, which shows a strong non-gaussian pattern even though U and $\Phi^{-1} \circ F_V(V)$ are marginally gaussian. Under such circumstances the gaussian copula assumption no longer holds, and our goal is to develop a broader class of graphical models that cover such cases.

In developing the new graphical models we would like to preserve an appealing feature of the GCGM; that is, the nonparametric operation is acted upon one-dimensional random variables individually, rather than a high-dimensional random vector jointly. This feature allows us to avoid high-dimensional smoothing, which is the source of the curse of dimensionality (Bellman, 1957). However, if we insist on using conditional independence (1) as the criterion for constructing graphical models, then we are inevitably lead to a fully fledged nonparametric procedure involving smoothing over the entire p -dimensional space, and the nice structures of the gaussian graphical model such as (2) would be lost. This motivates us to introduce a new statistical relation — *additive conditional independence*, or ACI, to replace conditional independence as our criterion. Although ACI is different from conditional independence, it satisfies the axioms for a semi-graphoid (Pearl and Verma, 1987) shared by conditional independence, which captures the essence of a graph.

Furthermore, under the gaussian copula assumption, ACI reduces to conditional independence. The graphical model defined by ACI — which we refer to as the Additive Semi-graphoid Model (ASG) — preserves a parallel structure to (2)

The paper is organized follows. In section 2 we outline the axiomatic system for a semi-graphoid logical relation, and introduce the notion of ACI as well as the ASG model it induces. In section 3 we investigate the relation between ACI and conditional independence under the gaussian copula assumption. In sections 4 and 5, we introduce the additive conditional covariance operator and additive precision operator, and study their relations with ACI. In sections 6 and 7 we introduce two estimators for the ASG model based on these operators. In section 8 we propose cross validation procedures to determine the tuning parameters. In sections 9 and 10 we compare the ASG model with GGM and GCGM by simulation and in an applied setting of genetic pathway inference. We conclude in section 11 with suggested future research directions. All proofs are presented in the Appendix.

2 Semi-graphoid axioms and additive conditional independence

Before introducing additive conditional independence, we first restate the axiomatic system developed by Pearl and Verma (1987) and Pearl, Geiger, and Verma (1989), in a slightly different form. Let 2^Γ be the class of all subsets of Γ . We call any subset $\mathcal{R} \subseteq 2^\Gamma \times 2^\Gamma \times 2^\Gamma$ a *three-way relation* on Γ . This terminology follows the tradition of two-way relations in mathematics (see, for example, Kelley, 1955, page 6).

Definition 1 (Semi-graphoid Axioms)

A three-way relation $\mathcal{R}$ is called a *semi-graphoid* if it satisfies the following conditions:

1. (symmetry) $(A, C, B) \in \mathcal{R} \Rightarrow (B, C, A) \in \mathcal{R}$;
2. (decomposition) $(A, C, B \cup D) \in \mathcal{R} \Rightarrow (A, C, B) \in \mathcal{R}$;
3. (weak union) $(A, C, B \cup D) \in \mathcal{R} \Rightarrow (A, C \cup B, D) \in \mathcal{R}$;
4. (contraction) $(A, C \cup B, D) \in \mathcal{R}, (A, C, B) \in \mathcal{R} \Rightarrow (A, C, B \cup D) \in \mathcal{R}$.

These axioms are extracted from conditional independence to convey the general idea of “ B is irrelevant for understanding A once C is known”, or “ C separates A and B ”. Conditional independence is a special case of the semi-graphoid three-way relation. More specifically, let $X = \{X^i : i \in \Gamma\}$ be a random vector and, for any $A \in \Gamma$, let X^A be the $\{X^i : i \in A\}$. Let $\mathcal{R}$ be the class of (A, C, B) such that $X^A \perp\!\!\!\perp X^B | X^C$ (X^A and X^B are conditionally independent given X^C), then it can be shown that $\mathcal{R}$ indeed satisfies the semi-graphoid axioms (Dawid, 1979).

Because, like conditional independence, a semi-graphoid relation $(A, C, B) \in \mathcal{R}$ also conveys the idea that C separates A and B , it seems reasonable to use the semi-graphoid relation itself as a criterion for constructing a graph. Indeed, the very statement “there is no direct link between node i and node j ” means, precisely, that “ $\{i, j\}^c$ separates i and j ”. In other words, it is reasonable to use the logical statement $(\{i\}, \{i, j\}^c, \{j\}) \in \mathcal{R}$ to determine the absence of edge between i and j . Using the semi-graphoid relation in place of

probabilistic conditional independence can free us from some awkward restrictions that come with a fully fledged probability model.

Let $X = (X^1, \dots, X^p)^\top$ be a random vector. For any subvector $U = (U^1, \dots, U^r)^\top$ of X , let $L_2(P_U)$ be the class of functions of U such that $E f(U) = 0$, $E f^2(U) < \infty$. For each U^i , let $\mathcal{A}_{U^i}$ denote a subset of $L_2(P_{U^i})$. Let $\mathcal{A}_U$ denote the additive family

$$\mathcal{A}_{U^1} + \dots + \mathcal{A}_{U^r} = \{f_1 + \dots + f_r : f_1 \in \mathcal{A}_{U^1}, \dots, f_r \in \mathcal{A}_{U^r}\}.$$

For two subspaces $\mathcal{A}$ and $\mathcal{B}$ of $L_2(P_X)$, let $\mathcal{A} \ominus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^\perp$, where orthogonality is in terms of the $L_2(P_X)$ -inner product. In the following, an unadorned inner product $\langle \cdot, \cdot \rangle$ always represents the $L_2(P_X)$ -inner product.

Definition 2

Let U , V , and W be subvectors of X . We say that U and V are additively conditionally independent (ACI) given W iff

$$(\mathcal{A}_U + \mathcal{A}_W) \ominus \mathcal{A}_W \perp (\mathcal{A}_V + \mathcal{A}_W) \ominus \mathcal{A}_W, \quad (3)$$

where $\perp$ indicates orthogonality in the $L_2(P_X)$ -inner product. We write relation (3) as $U \perp\!\!\!\perp_A V|W$.

Strictly speaking, ACI should be stated with respect to the triplet of functional families $(\mathcal{A}_U, \mathcal{A}_V, \mathcal{A}_W)$, but in most cases we can ignore this qualification without loss of clarity. Lauritzen (1996, page 31) mentioned a relation similar to (3) where $\mathcal{A}_U$, $\mathcal{A}_V$, and $\mathcal{A}_W$ are Euclidean subspaces, but absent from that construction are the underlying random vectors U , V , W , which is crucial to our development.

The next theorem gives seven equivalent conditions for ACI, each of which can be used as its definition. As we will see in the proof of Theorem 2, these various forms are immensely useful in exploring the ramifications of ACI. In the following, for an operator P on a Hilbert space $\mathcal{H}$ let $\ker P$ and $\text{ran} P$ denote the kernel and range of P , and let $\overline{\text{ran} P}$ be the closure of $\text{ran} P$.

Theorem 1

Let U , V , W be subvectors of X . The following statements are equivalent:

1. $(\mathcal{A}_U + \mathcal{A}_W) \ominus \mathcal{A}_W \perp (\mathcal{A}_V + \mathcal{A}_W) \ominus \mathcal{A}_W$;
2. $\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W \perp \mathcal{A}_{V \cup W} \ominus \mathcal{A}_W$;
3. $(I - P_{\mathcal{A}_W}) \mathcal{A}_U \perp (I - P_{\mathcal{A}_W}) \mathcal{A}_V$;
4. $\mathcal{A}_U \subseteq \ker(P_{\mathcal{A}_{V \cup W}} \ominus P_{\mathcal{A}_W})$;
5. $\mathcal{A}_U \subseteq \ker(P_{\mathcal{A}_{V \cup W}} - P_{\mathcal{A}_W})$;
6. $\mathcal{A}_{U \cup W} \subseteq \ker(P_{\mathcal{A}_{V \cup W}} \ominus P_{\mathcal{A}_W})$;

$$7. \mathcal{A}_{U \cup W} \subseteq \ker(P_{\mathcal{A}_{V \cup W}} - P_{\mathcal{A}_W}).$$

The next theorem shows that ACI satisfies the semi-graphoid axioms, which gives justification of using it to construct graphical models.

Theorem 2

The additive conditional independence relation is a semi-graphoid.

We are now ready to define the additive semi-graphoid model.

Definition 3

A random vector X follows an additive semi-graphoid model with respect to a graph $\mathcal{G} = (\Gamma, E)$ iff

$$X^i \perp\!\!\!\perp_A X^j | X^{-(i,j)} \iff (i,j) \notin E. \quad (4)$$

We write this condition as $X \sim \text{ASG}(\mathcal{G})$.

3 Relation with copula graphical models

While in the last section we have seen it is reasonable to use ACI instead of conditional independence as a criterion for constructing graphs, we now investigate the special cases where ACI reduces conditional independence.

Theorem 3

Suppose

1. X has a gaussian copula distribution with copula functions $f_1, \dots, f_p$;
2. U, V , and W are subvectors of X ;
3. $\mathcal{A}_X^i = \text{span}\{f_i\}$, $i = 1, \dots, p$.

Then $U \perp\!\!\!\perp_A V | W$ if and only if $U \perp\!\!\!\perp V | W$.

Under the specific form of ASG model in Theorem 3, there is no edge between i and j iff the (i, j) th entry of the precision matrix of $(f_1(X^1), \dots, f_p(X^p))^T$ is 0. The latter is population-level criterion used by the GCGM estimator. In this sense, ASG can be regarded as a generalization of the nonparanormal models.

Since our main interests are in the ASG models with $\mathcal{A}_X^i = L_2(P_X^i)$, it is also of interest to ask whether $\perp\!\!\!\perp_A$ coincides with $\perp\!\!\!\perp$ in this case. The next theorem and the subsequent numerical investigation show that they are very nearly equivalent for all practical purposes.

Theorem 4

Suppose

1. X has a gaussian copula distribution with copula functions $f_1, \dots, f_p$;

2. U , V , and W are sub-vectors of X ;

3. $\mathcal{A}_{X^i} = L_2(P_{X^i})$.

Then $U \perp_A V|W$ implies $U \perp V|W$.

The proof, which is omitted, is essentially the same as that of Theorem 3. Interestingly, under $\mathcal{A}_{X^i} = L_2(P_{X^i})$, $\perp$ does not imply $\perp_A$ even under the gaussian copula assumption. However, the following numerical investigation suggests that this implication holds approximately, with vanishingly small error. We are able to carry out this investigation because an upper bound of

$$\rho_{UV|W}^2 = \sup\{\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle^2 : f \in L_2(P_U), g \in L_2(P_V), \|f\| = \|g\| = 1\}$$

can be calculated explicitly under the copula gaussian distribution. To our knowledge this bound has not been recorded in the literature; so it is of interest in its own.

Proposition 1

Suppose X has a gaussian copula distribution with copula functions $f_1, \dots, f_p$. Without loss of generality, suppose $E[f_i(X^i)] = 0$, $\text{var}[f_i(X^i)] = 1$. Let

$$U = f_i(X^i), \quad V = f_j(X^j), \quad W = \{f_k(X^k) : k \neq i, j\},$$

and $R_{VW} = \text{cov}(V, W)$, $R_{WW} = \text{var}(W)$, $R_{WU} = \text{cov}(W, U)$, $\rho_{UV} = \text{cor}(U, V)$. Then

$$\rho_{UV|W} \leq \max\{|\rho_{UV}^\alpha - R_{VW}^{\odot \alpha} (R_{WW}^{\odot \alpha})^{-1} R_{WU}^{\odot \alpha}| : \alpha = 1, 2, \dots\} \equiv \tau_{UV|W}, \quad (5)$$

$A^{\odot \alpha}$ is the α -fold Hadamard product $A \odot \dots \odot A$ and $|\cdot|$ means determinant.

Using this proposition we conduct the following numerical investigation. First we generate a positive definite random matrix C with $C_{12} = C_{21} = 0$ as follows. We generate $\tilde{C}$ from the Wishart distribution $W(I_p, k)$, where k is the degrees of freedom, and then set C_{12} and C_{21} to 0. This does not guarantee C to be positive definite. If it is not, we repeat this process until we get a positive definite matrix, which is then set to be C . Let $R_{XX} = [\text{diag}(C)]^{1/2} C^{-1} [\text{diag}(C)]^{1/2}$. Let R_{WU} , R_{VW} , and R_{WW} be the $(p-2) \times 1$, $(p-2) \times 1$, and $(p-2) \times (p-2)$ matrices

$$\{(R_{XX})_{i1} : i=3, \dots, p\}, \quad \{(R_{XX})_{i2} : i=3, \dots, p\}, \quad \{(R_{XX})_{ij} : i, j=3, \dots, p\},$$

and let $R_{UW} = R_{WU}^\top$, $R_{VW} = R_{WV}^\top$. Note that, if $X \sim N(0, R_{XX})$, then it is guaranteed that $U \perp V|W$, and $\text{var}(X^i) = 1$.

We choose the degrees of freedom k to be relatively small to prevent C/k from converging to I_p , but large enough to avoid singularity. In particular, we take $k = \lceil 5p/4 \rceil, \lceil 6p/4 \rceil, \dots, \lceil 8p/4 \rceil$, which guarantees a wide range of covariance matrices. We take $p = 20, 40, \dots, 100$. We compute $\tau_{U|V|W}$ by taking the maximum of the first 10 terms in (5), which in our simulations is always the global maximum. For each combination of (p, k) , we generate 500 random matrices and present the averages of $\tau_{U|V|W}$ in Table 1. Because $\rho_{U|V|W} = 0$ iff $U \perp\!\!\!\perp_A V|W$, the small numbers in Table 1 indicate that for all practical purposes we can treat $U \perp\!\!\!\perp V|W \Rightarrow U \perp\!\!\!\perp_A V|W$ as valid under the gaussian copula model, even though it is not a mathematical fact.

In the similar way it is related to the gaussian copula graphical models, the ASG is also related to the more general transelliptical graphical models recently introduced by Liu, Han, and Zhang (2012). A random vector X is said to have a transelliptical distribution with shape parameter Σ (a $p \times p$ positive definite matrix) if there exist injections $f_1, \dots, f_p$ on Ω_X to $\mathbb{R}$ such that the distribution of $U = (f_1(X^1), \dots, f_p(X^p))^T$ depends only on $u^T \Sigma u$. Furthermore, X is said to follow a transelliptical graphical model with respect to a graph $\mathcal{G} = (\Gamma, E)$ iff X has a transelliptical distribution and $E = \{(i, j) : (\Sigma^{-1})_{ij} \neq 0\}$. In this case we write $X \sim \text{TEG}(\mathcal{G})$. The following corollary states that ASG reduces to the transelliptical graphical model if the additive families $\mathcal{A}_{X^i}$ are chosen to be $\text{span}(f_i)$ and the underlying distribution of X is assumed to be transelliptical. Its proof is parallel to that of Theorem 3 and is omitted.

Corollary 1

Suppose

1. X has a transelliptical distribution with respect to a positive definite shape parameter Σ and injections $f_1, \dots, f_p$;
2. $\mathcal{A}_{X^i} = \text{span}(f_i)$.

Then $X \sim \text{ASG}(\mathcal{G})$ iff $X \sim \text{TEG}(\mathcal{G})$.

4 Additive conditional covariance operator

In this and the next sections we introduce two linear operators that will give rise to two estimation procedures for ASG. The first is the additive conditional covariance operator (ACCO). Consider the bilinear form

$$\mathcal{A}_U \times \mathcal{A}_V \rightarrow \mathbb{R}, \quad (f, g) \mapsto E[f(U)g(V)].$$

By the Riesz representation theorem, this induces a bounded linear operator $\Sigma_{UV} : \mathcal{A}_V \rightarrow \mathcal{A}_U$ such that $\langle f, \Sigma_{UV} g \rangle = E[f(U)g(V)]$. We call the operator

$$\sum_{U|V|W} = \sum_{UV} - \sum_{UW} \sum_{WV} \quad (6)$$

the additive conditional covariance operator from V to U given W . This operator is different from the conditional variance operator introduced by Fukumizu, Bach, and Jordan (2009),

because of the additive nature of the functional spaces $\mathcal{A}_U$, $\mathcal{A}_V$, and $\mathcal{A}_W$. In particular, the following identity no longer holds:

$$\langle f, \sum_{U \vee V|W} g \rangle = E[\text{cov}(f(U), g(V)|W)].$$

In the following, for a Hilbert space $\mathcal{H}$ a subspace $\mathcal{L}$ of $\mathcal{H}$ and an operator T defined on $\mathcal{H}$ we use $T|_{\mathcal{L}}$ to denote the operator T restricted on $\mathcal{L}$. That is, $T|_{\mathcal{L}}$ is a mapping from $\mathcal{L}$ to $\mathcal{H}$ with $(T|_{\mathcal{L}})f = Tf$ for all $f \in \mathcal{L}$. The following lemma gives a relation between the projection operator and the covariance operator.

Lemma 1

Let $P_{\mathcal{A}_W}: \mathcal{A}_X \rightarrow \mathcal{A}_W$ be the projection on to $\mathcal{A}_W$. Then $P_{\mathcal{A}_W}|_{\mathcal{A}_U} = \Sigma_{WU}$.

The next theorem shows that $\Sigma_{UV|W} = 0$ iff $U \perp_{\mathcal{A}} V|W$, which is the basis for our first ASG estimator. This resembles the relation between conditional covariance operator and conditional independence; that is, if $C_{UV|W}$ is the conditional covariance operator in Fukumizu, Jordan, and Bach (2009), then $C_{UV|W} = 0$ iff $U \perp V|W$.

Theorem 5

$U \perp_{\mathcal{A}} V|W$ iff $\Sigma_{UV|W} = 0$.

5 Additive precision operator

We now introduce the second linear operator – the additive precision operator (APO), and show that a statement resembling the equivalence of (1) and (2) can be made for ASG if we replace the precision matrix by APO. As a result, the estimation of the ASG model can be turned into the sparse estimation of APO.

As a preparation, we first develop the notion of a matrix of operators on a direct-sum functional space. Let $\mathcal{A}_{\oplus X}$ be the Hilbert space of functions consisting of the same set of functions as $\mathcal{A}_X$ but with the inner product

$$\langle f_1 + \cdots + f_p, g_1 + \cdots + g_p \rangle_{\mathcal{A}_{\oplus X}} = \sum_{i=1}^p \langle f_i, g_i \rangle_{\mathcal{A}_{X^i}}.$$

Suppose for each i, j we are given a linear operator $T_{X^j X^i}: \mathcal{A}_{X^i} \rightarrow \mathcal{A}_{X^j}$. Then we can construct a linear operator $T: \mathcal{A}_{\oplus X} \rightarrow \mathcal{A}_{\oplus X}$ as follows

$$T(f_1 + \cdots + f_p) = \sum_{i=1}^p \sum_{j=1}^p T_{X^j X^i} f_i. \quad (7)$$

This construction is legitimate because in the additive case $f_1 + \cdots + f_p$ uniquely determines $(f_1, \dots, f_p)$. In fact, if we fix $(c_2, \dots, c_p)$ in $\Omega_{X^2} \times \cdots \times \Omega_{X^p}$ and vary x^1 in Ω_{X^1} , then using $E[f_1(X^1)] = 0$ we have

$$f_1(x^1) = f_1(x^1) + f_2(c)^2 + \cdots + f_p(c^p) - E[f_1(X^1) + f_2(c^2) + \cdots + f_p(c^p)].$$

Conversely, a linear operator $T: \mathcal{A}_{\oplus X} \rightarrow \mathcal{A}_{\oplus X}$ uniquely determines operators $T_{X^i X^j}: \mathcal{A}_{X^i} \rightarrow \mathcal{A}_{X^j}$ that satisfies (7), as follows. For each $f_i \in \mathcal{A}_{\oplus X}$, Tf_i can be represented as $h_1 + \cdots + h_p$, where $h_j \in \mathcal{A}_{X^j}$. Moreover, $h_1 + \cdots + h_p$ uniquely determines $(h_1, \dots, h_p)$, and hence also h_j . We define $T_{X^i X^j}: \mathcal{A}_{X^i} \rightarrow \mathcal{A}_{X^j}$ by $T_{X^i X^j}(f_i) = h_j$. Thus, there is a one-to-one correspondence between T and $\{T_{X^i X^j}\}$ and they are related by (7). Henceforth we denote this one-to-one correspondence by the equality

$$T = \{T_{X^j X^i}: i, j = 1, \dots, p\}.$$

Moreover, T is bounded iff each $T_{X^i X^j}$ is bounded, as shown in the next proposition.

Proposition 2

The following statements are equivalent

1. $T_{X^i X^j} \in \mathcal{B}(\mathcal{A}_{X^i}, \mathcal{A}_{X^j})$ for $i, j = 1, \dots, p$;
2. $T \in \mathcal{B}(\mathcal{A}_{\oplus X}, \mathcal{A}_{\oplus X})$.

Our construction of $T = \{T_{X^i X^j}\}$ echoes the construction of Bach (2008). However, our construction is relative to a different inner product and will play a new role under ACI. We are now ready to introduce the additive precision operator.

Definition 4

The operator $Y: \mathcal{A}_{\oplus X} \rightarrow \mathcal{A}_{\oplus X}$ defined by the matrix of operators $\{\Sigma_{X^i X^j}: i \in \Gamma, j \in \Gamma\}$ is called the additive covariance operator. If it is invertible then its inverse Θ is called the additive precision operator.

It can be easily checked that, for any $f, g \in \mathcal{A}_{\oplus X}$,

$$\langle f, Yg \rangle_{\mathcal{A}_{\oplus X}} = \text{cov}[f(X), g(X)] = \langle f, g \rangle. \quad (8)$$

This relation also shows why we used $\mathcal{A}_{\oplus X}$ instead of $\mathcal{A}_X$ to define Y in the first place: if we had used $\mathcal{A}_X$, then Y would have satisfied $Yf = f$ for any $f \in \mathcal{A}_X$; that is, Y would have been the identity mapping from $\mathcal{A}_X$ to $\mathcal{A}_X$. In the meantime, we note that, by construction, $Y_{X^i X^j} = \Sigma_{X^i X^j}$ for all $i, j = 1, \dots, p$.

We should note that it is generally unreasonable to assume that the inverse of a covariance operator to be a bounded operator, because we typically assume the sequence of eigenvalues of such operators tend to 0. However, in our construction, the inner product $\langle f_i, f_i \rangle_{\mathcal{A}_{\oplus X}}$ is the marginal variance $\text{var}[f_i(X^i)]$. Thus our covariance operator Y is more like a correlation than a covariance. In particular, Y_{ii} are identity operators by construction. If we assume Y_{ij} are compact for all $i \neq j$, then Y is of the form “identity plus compact perturbation”. Such

operators, if invertible, are bounded from below by cI for some $c > 0$ (Bach, 2008), and therefore have bounded inverses. We summarize this fact in the following proposition.

Proposition 3

Suppose Y is invertible and, for each $i \neq j$, Y_{ij} is compact. Then Y^{-1} is bounded.

To establish the relation between the precision operator Θ and ACI, we need the block inversion formulas for operators. For any subvector U of X , let V denote U^c , and define the matrices of operators

$$\begin{aligned} T_{UU} &= \{T_{X^i X^j} : X_i \in U, X_j \in U\}, & T_{UV} &= \{T_{X^i X^j} : X_i \in U, X_j \in V\}, \\ T_{VU} &= \{T_{X^i X^j} : X_i \in V, X_j \in U\}, & T_{VV} &= \{T_{X^i X^j} : X_i \in V, X_j \in V\}. \end{aligned}$$

From these definitions it follows that $T_{UU} \mathcal{A}_{\oplus U} \subseteq \mathcal{A}_{\oplus U}$, and

$$\langle f, T_{UU} g \rangle_{\mathcal{A}_{\oplus X}} = \langle f, Tg \rangle_{\mathcal{A}_{\oplus X}} \quad \text{for any } f, g \in \mathcal{A}_{\oplus U}.$$

The same can be said of T_{VV} , T_{UV} , and T_{VU} . The following operator identities resemble the matrix identities derived from block-matrix inverse.

Lemma 2

Suppose that $T \in \mathcal{B}(\mathcal{A}_{\oplus X}, \mathcal{A}_{\oplus X})$ is a self adjoint, positive definite operator. Then: the operators

$$T_{UU}, \quad T_{VV}, \quad T_{UU} - T_{UV} T_{VV}^{-1} T_{VU}, \quad T_{VV} - T_{VU} T_{UU}^{-1} T_{UV}$$

are bounded, self adjoint, and positive definite, and the following identities hold

$$\begin{aligned} (T^{-1})_{UU} &= (T_{UU} - T_{UV} T_{VV}^{-1} T_{VU})^{-1}, \\ (T^{-1})_{VV} &= (T_{VV} - T_{VU} T_{UU}^{-1} T_{UV})^{-1}, \\ (T^{-1})_{UV} &= -(T^{-1})_{UU} T_{UV} T_{VV}^{-1} = -T_{UU}^{-1} T_{UV} (T^{-1})_{VV}, \\ (T^{-1})_{VU} &= -(T^{-1})_{VV} T_{VU} T_{UU}^{-1} = -T_{VV}^{-1} T_{VU} (T^{-1})_{UU}, \end{aligned} \quad (9)$$

where $(T^{-1})_{UU}$ and $(T^{-1})_{VV}$ are the blocks of T^{-1} corresponding to $\mathcal{A}_{\oplus U}$ and $\mathcal{A}_{\oplus V}$.

The above lemma shows that if Y has a bounded inverse, then Θ_{UV} is a well defined operator in $\mathcal{B}(\mathcal{A}_{\oplus V}, \mathcal{A}_{\oplus U})$. Moreover, it behaves just like the off-diagonal block of the inverse of a covariance matrix. As we will see in the Appendix, manipulating the blocks of Y and Θ in the same way we do with a gaussian covariance matrix lead to the following equivalence.

Theorem 6

Suppose (U, V, W) is a partition of X and Y is invertible. Then $U \perp\!\!\!\perp_A V|W$ iff $\Theta_{UV} = 0$.

This relation is parallel to the equivalence between (1) and (2) in the gaussian case, which means we can turn estimating ASG into estimating the locations of the zero entries of APO, as shown in the next corollary.

Corollary 2

A random vector X follows an $\text{ASG}(\mathcal{G})$ model if and only if $\Theta_{X^i X^j} = 0$ whenever $(i, j) \notin E$.

Note that the graph $\text{ASG}(\mathcal{G})$ is based entirely on pairwise additive conditional independence (4), which is similar to the *pairwise Markov property* for classical graphical models based on conditional independence. In the classical setting, pairwise Markov property with respect to a graph $\mathcal{G} = (\Gamma, E)$ means

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \iff (i, j) \notin E.$$

Under fairly mild conditions, this property also implies the *global Markov property*, which means if A, B, C are subsets of Γ and C separates A and B in $\mathcal{G}$, then

$$X^A \perp\!\!\!\perp X^B | X^C.$$

See Lauritzen (1996, Chapter 3). It is of theoretical interest to see if the parallel implication holds true in our additive setting. For convenience, let us agree to call property (4) the *pairwise additive Markov property* with respect to $\mathcal{G}$. We are interested in whether $X^A \perp\!\!\!\perp_A X^B | X^C$ holds whenever C separates A and B according to the graph $\mathcal{G}$ determined by (4), a property that can be appropriately defined as the *global additive Markov property* with respect to $\mathcal{G}$.

Using Theorem 6 we can easily derive a partial answer to this question, as follows. Suppose $\{A, B, C\}$ is a partition of Γ and C separates A and B according to $\mathcal{G}$ defined by (4). Then all paths between A and B intersects C . In other words, the block Θ_{AB} and Θ_{BA} of Θ consists of zero operators. However, by Theorem 6, this is equivalent to $X^A \perp\!\!\!\perp_A X^B | X^C$. We will leave it for future research to establish whether this statement is true when $\{A, B, C\}$ is not necessarily a partition.

In the next two sections we introduce two estimators of the ASG model, one based on the ACCO introduced in the last section, and one based on the APO.

6 The ACCO-based estimator

By Theorem 5, it is reasonable to determine the absence of edges by smallness of the sample estimate of $\|\Sigma_{X^i X^j | X^{-(i,j)}}\|$. To develop this estimate, we first derive the finite-sample matrix representation of the operator $\Sigma_{X^i X^j | X^{-(i,j)}}$.

For each $i = 1, \dots, p$, let $\kappa_i : \Omega_{X^i} \times \Omega_{X^i} \rightarrow \mathbb{R}$ be a positive definite kernel. Let

$$\mathcal{A}_{X^i} = \text{span}\{\kappa_i(\cdot, X_1^i) - E_n \kappa_i(X_1^i, X_1^i), \dots, \kappa_i(\cdot, X_n^i) - E_n \kappa_i(X_n^i, X_n^i)\},$$

where $E_n \kappa_i(X^i, X_k^i) = n^{-1} \sum_{\ell=1}^n \kappa_i(X_\ell^i, X_k^i)$. Let I_n be the $n \times n$ identity matrix and 1_n be the n -dimensional vector with all its entries equal to 1. Let Q be the projection matrix $I_n - 1_n 1_n^\top / n$. Let κ_i° denote the vector-valued function

$$x^i \mapsto (\kappa_i(x^i, X_1^i) - E_n \kappa_i(X_1^i, X_1^i), \dots, \kappa_i(x^i, X_n^i) - E_n \kappa_i(X_n^i, X_n^i))^\top.$$

Then, any $f \in \mathcal{A}_{X^i}$ can be expressed as $f = \kappa_i^{\circ\top} [f]$ for some $[f] \in \mathbb{R}^n$, which is called the coordinate of f in the spanning system κ_i° . A coordinate system is relative to a class of functions, and sometimes we need to use symbols such as $[f]_{\mathcal{A}_{X^i}}$ for clarity. But in most of the discussions below this is unnecessary. In this coordinate notation we have

$(f(X_1^i), \dots, f(X_n^i))^\top = Q K_i [f]$, where $K_i = \{\kappa_i(X_k^i, X_\ell^i) : k, \ell = 1, \dots, n\}$. Also, for any $f, g \in \mathcal{A}_{X^i}$,

$$\langle f, g \rangle_{\mathcal{A}_{X^i}} = n^{-1} (Q K_i [f])^\top (Q K_i [g]) = n^{-1} [f]^\top K_i Q K_i [g].$$

Let $K_{-(i,j)}$ denote the $n(p-2) \times n$ matrix obtained by removing the i th and j th block of the $np \times n$ matrix $(K_1, \dots, K_p)^\top$. Then, for any $f \in \mathcal{A}_{X^{-(i,j)}}$ we have

$$(f(X_1^{-(i,j)}), \dots, f(X_n^{-(i,j)}))^\top = Q K_{-(i,j)}^\top [f] \quad (10)$$

for some $[f] \in \mathbb{R}^{n(p-2)}$. Finally, for any linear operator T defined on a finite dimensional Hilbert space $\mathcal{H}$ there is a matrix $[T]$ such that $[Tf] = [T][f]$ for each function $f \in \mathcal{H}$. The matrix $[T]$ is called the matrix representation of T . This notation system is adopted from Horn and Johnson (1985) and was used in Li, Artemiou, and Li (2011), Li, Chun, and Zhao (2012), and Lee, Li, and Chiaromonte (2013).

Lemma 3

The following relations hold:

$$\begin{aligned} K_i Q K_i [\sum_{X^i X^j}] &= K_i Q K_j, & K_i Q K_i [\sum_{X^i X^{-(i,j)}}] &= K_i Q K_{-(i,j)}^\top \\ K_{-(i,j)} Q K_{-(i,j)}^\top [\sum_{X^{-(i,j)} X^i}] &= K_{-(i,j)} Q K_i. \end{aligned} \quad (11)$$

Because $K_i Q K_i$ and $K_{-(i,j)} Q K_{-(i,j)}^\top$ are singular, we solve the equations in (11) by Tychonoff regularization:

$$\begin{aligned}
[\sum_{X^i X^j}] &= (K_i Q K_i + \varepsilon_1 I_n)^{-1} K_i Q K_j, \\
[\sum_{X^i X^{-(i,j)}}] &= (K_i Q K_i + \varepsilon_1 I_n)^{-1} K_i Q K_{-(i,j)}^\top, \\
[\sum_{X^{-(i,j)} X^i}] &= (K_{-(i,j)} Q K_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)})^{-1} K_{-(i,j)} Q K_i.
\end{aligned} \quad (12)$$

Note that we use two different tuning constants ε_1 and ε_2 , because the matrices involved have different dimensions. We will discuss how to choose these constants in Section 8. is then The matrix representation of $\Sigma_{X^i X^j | X^{-(i,j)}}$ is then

$$[\sum_{X^j X^i | X^{-(i,j)}}] = [\sum_{X^j X^i}] - [\sum_{X^j X^{-(i,j)}}] [\sum_{X^{-(i,j)} X^i}]. \quad (13)$$

By definition, $\|\Sigma_{X^i X^j | X^{-(i,j)}}\|^2$ is the maximum of

$$\langle \sum_{X^j X^i | X^{-(i,j)}} f, \sum_{X^j X^i | X^{-(i,j)}} f \rangle_{\mathcal{A}_{X^i}} = n^{-1} [f]^\top [\sum_{X^j X^i | X^{-(i,j)}}]^\top K_j Q K_j [\sum_{X^j X^i | X^{-(i,j)}}] [f]$$

subject to $\langle f, f \rangle_{\mathcal{A}_{X^i}} = [f]^\top K_i Q K_i [f] = 1$. To solve this generalized eigenvalue problem, let $v = (K_i Q K_i)^{1/2} [f]$. Then, using inverse Tychonoff regularization, $[f] = (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} v$. So it is equivalent to maximizing

$$v^\top (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} [\sum_{X^j X^i | X^{-(i,j)}}]^\top K_j Q K_j [\sum_{X^j X^i | X^{-(i,j)}}] (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} v$$

subject to $v^\top v = 1$, which implies that $\|\Sigma_{X^i X^j | X^{-(i,j)}}\|^2$ is the largest eigenvalue of

$$\Lambda_{X^i X^j} = (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} [\sum_{X^j X^i | X^{-(i,j)}}]^\top K_j Q K_j [\sum_{X^j X^i | X^{-(i,j)}}] (K_i Q K_i + \varepsilon_1 I_n)^{-1/2}.$$

Since the factor $K_j Q K_j [\Sigma_{X^i X^j | X^{-(i,j)}}]$ in the right-hand side is of the form

$$K_j Q K_j (K_j Q K_j + \varepsilon_1 I_n)^{-1} K_j Q (I_n - K_{-(i,j)}^\top (K_{-(i,j)} Q K_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)})^{-1} K_{-(i,j)}) Q K_i,$$

it is reasonable to replace the first factor $K_j Q K_j$ by $K_j Q K_j + \varepsilon_1 I_n$, so that $\Lambda_{X^i X^j}$ becomes the quadratic form $\Delta_{X^i X^j} \Delta_{X^i X^j}^\top$, where

$$\begin{aligned}
\Delta_{X^i X^j} &= (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} K_i \\
&\quad (Q - Q K_{-(i,j)}^\top (K_{-(i,j)} Q K_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)})^{-1} K_{-(i,j)} Q) K_j (K_j Q K_j + \varepsilon_1 I_n)^{-1/2}.
\end{aligned} \quad (14)$$

Hence, at the sample level, $\|\Sigma_{X^i X^j | X^{-(i,j)}}\|$ is the largest singular value of $X^i X^j$. Because $\Delta_{X^i X^j}^\top = \Delta_{X^j X^i}$, we have the symmetry $\|\Sigma_{X^i X^j | X^{-(i,j)}}\| = \|\Sigma_{X^j X^i | X^{-(i,j)}}\|$.

The above procedure involves inversion of $K_{-(i,j)} Q K_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)}$, which is a large matrix if the graph $\mathcal{G}$ contains many nodes. Nevertheless, as the next proposition shows the actual amount computation is not large — it is essentially that of spectral decomposition of an n by n matrix. Let $\text{diag}(A_1, \dots, A_p)$ be the block-diagonal matrix with diagonal blocks $A_1, \dots, A_p$, and V be a matrix in $\mathbb{R}^{s \times t}$, with $s > t$, which has singular value decomposition

$$V = (L_1, L_0) \text{diag}(D, 0) (R_1, R_0)^\top, \quad (15)$$

where $L_1 \in \mathbb{R}^{s \times u}$, $L_0 \in \mathbb{R}^{s \times (s-u)}$, $R_1 \in \mathbb{R}^{t \times u}$, $R_0 \in \mathbb{R}^{t \times (t-u)}$, and $D \in \mathbb{R}^{u \times u}$ is a diagonal matrix with diagonal elements $\delta_1 \dots \delta_u > 0$.

Proposition 4

If the conditions in the last paragraph are satisfied, then

1. $(VV^\top + \varepsilon I_s)^{-1} = V R_1 D^{-1} ((D^2 + \varepsilon I_u)^{-1} - \varepsilon^{-1} I_u) D^{-1} R_1^\top V^\top + \varepsilon^{-1} I_s$;
2. $V^\top (VV^\top + \varepsilon I_s)^{-1} V = R_1 D (D^2 + \varepsilon I_u)^{-1} D R_1^\top$.

The above proposition allows us to calculate the inversions of $s \times s$ matrix $(VV^\top + \varepsilon I_s)$ essentially by spectral decomposition of $t \times t$ matrix $V^\top V$. We now summarize the estimation procedure as an algorithm. Suppose we observe an i.i.d. sample $X_1, \dots, X_n$ from X , where X_i is a p -dimensional random vector $(X_i^1, \dots, X_i^p)^\top$.

1. Standardize each component of X marginally. That is, for each $i = 1, \dots, p$, we standardize $X_1^i, \dots, X_n^i$, so that $E_n X^i = 0$ and $\text{var}_n X^i = 1$.
2. Choose a positive definite kernel $\kappa: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, such as the gaussian radial kernel. Choose the kernel parameter γ using, for example, the procedure in Li, Chun, and Zhao (2012). Compute the kernel matrices K_i and $K_{-(i,j)}$.
3. Compute $X^i X^j$ in (14), using part 2 of Proposition 4 to compute the matrix $Q K_{-(i,j)}^\top (K_{-(i,j)} Q K_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)})^{-1} K_{-(i,j)} Q$, with $V = K_{-(i,j)} Q$, $s = n(p-2)$, $t = n$, and $u = n-1$. The tuning constants ε_1 and ε_2 are determined using the method described in Section 8.
4. For each $i > j$, compute the largest singular value of $X^i X^j$. If it is smaller than a certain threshold, declare that there is no edge between i and j .

We emphasize again that throughout this procedure, the largest matrix whose spectral decomposition is needed is of dimension $n \times n$, regardless of the dimension p of X .

7 The APO-based estimator

Our second estimator is based on Corollary 2, which states that $X^i \perp\!\!\!\perp_A X^{-(i,j)}$ iff $\Theta_{X^j X^i} = 0$. Thus we can determine the absence of edge between i, j if the sample version of $\|\Theta_{X^j X^i}\|$ is below a threshold. We first derive the matrix representations of the operator Y and its inverse Θ . Let $K_{1:p} = (K_1, \dots, K_p)^\top$.

Theorem 7

The matrix representation of Y satisfies the following equation:

$$\text{diag}(K_1 Q K_1, \dots, K_p Q K_p) [\Upsilon] = K_{1:p} Q K_{1:p}^\top \quad (16)$$

The matrix $[\Theta]$ is obtained by solving (16) for $[Y]^{-1}$ with Tychonoff regularization:

$$[\Theta] = (K_{1:p} Q K_{1:p}^\top + \varepsilon_3 I_{np})^{-1} \text{diag}(K_1 Q K_1, \dots, K_p Q K_p). \quad (17)$$

To find $\|\Theta_{X^j X^i}\|$, we maximize

$$\langle \Theta_{X^j X^i} f_i, \Theta_{X^j X^i} f_i \rangle_{\mathcal{H}_{X^j}} = n^{-1} [f_i]^\top [\Theta_{X^j X^i}]^\top K_j Q K_j [\Theta_{X^j X^i}] [f_i]$$

subject to $\langle f_i, f_i \rangle_{\mathcal{H}_{X^i}} = [f_i]^\top K_i Q K_i [f_i] = 1$. Following the similar development that leads to (14), we see that $\|\Theta_{X^j X^i}\|^2$ is the largest eigenvalue of

$$(K_i Q K_i + \varepsilon_1 I_n)^{-1/2} [\Theta_{X^j X^i}]^\top (K_j Q K_j + \varepsilon_1 I_n) [\Theta_{X^j X^i}] (K_i Q K_i + \varepsilon_1 I_n)^{-1/2},$$

where $[\Theta_{X^j X^i}]$ is the (j, i) th block of (17). Equivalently, $\|\Theta_{X^j X^i}\|$ is the largest singular value of

$$\Gamma_{X^i X^j} = (K_i Q K_i + \varepsilon_1 I_n)^{-1/2} [\Theta_{X^j X^i}]^\top (K_j Q K_j + \varepsilon_1 I_n)^{1/2}.$$

We now summarize the algorithm for the APO-based estimator. The first two steps are the same as in the ACCO-based algorithm; the last two steps are listed below.

3' Compute $[\Theta]$ according to (17), where the inverse $(K_{1:p} Q K_{1:p}^\top + \varepsilon_3 I_{np})^{-1}$ is computed using part 1 of Proposition 4, with $V = K_{1:p} Q$, $\varepsilon = \varepsilon_3$, $s = np$, and $u = n - 1$. The tuning constant ε_3 is determined by the method in Section 8.

4' For each $i, j = 1, \dots, p$, $i > j$, compute the largest singular value of $\Gamma_{X^i X^j}$. If it is smaller than a certain threshold, then there is no edge between i and j .

In concluding this section, we would like to mention that the operator norm used for both the ACCO and APO can be replaced by other norms, such as the Hilbert Schmidt norm. We have experimented both norms in our simulation studies (not presented), and found that their performances are similar.

8 Tuning parameters for regularization

In this section we propose cross-validation and generalized cross-validation procedures to select ε 's in the Tychonoff regularization. We need to invert three types of matrices

$$\text{type I: } K_i Q K_i, \quad \text{type II: } K_{-(i,j)} Q K_{-(i,j)}^\top, \quad \text{type III: } K_{1:p} Q K_{1:p}^\top.$$

As before, we denote the constants in the regularized inverses as ε_1 , ε_2 , and ε_3 .

Since our cross-validation is based on the total error in predicting one set of functions by another set of functions, we first formulate such predictions in generic terms. Let U and V be random vectors and $W = (U^\top, V^\top)^\top$. Let $\mathcal{F} = \{f_0, f_1, \dots, f_r\}$ be a set of functions of U , and $\mathcal{G} = \{g_0, g_1, \dots, g_s\}$ be a set of functions of V , where g_0 and f_0 represent the constant function 1, which are included to account for any constant terms in prediction. Suppose we have an i.i.d. sample $\{W_1, \dots, W_n\}$ of W . Let $A = \{a_1, \dots, a_m\} \subseteq \{1, \dots, n\}$, $B = \{b_1, \dots, b_{n-m}\} = A^c$. Let $\{W_a : a \in A\}$ be the training set and $\{W_b : b \in B\}$ be the testing set. Suppose

$\hat{E}_{V|U}^{(\varepsilon)} : \mathcal{G} \rightarrow \mathcal{F}$ is a training-set-based operator such that, for each $g \in \mathcal{G}$, $\hat{E}_{V|U}^{(\varepsilon)} g$ is the best prediction of $g(V)$ based on functions in $\mathcal{F}$, where ε is the tuning constant to be determined. The total error for predicting any function in $\mathcal{G}$ evaluated at the testing set is

$$\sum_{b \in B} \sum_{\mu=0}^s [g_\mu(V_b) - (\hat{E}_{V|U}^{(\varepsilon)} g_\mu)(U_b)]^2. \quad (18)$$

Now let L_U be the $(r+1) \times m$ matrix whose i th row is the vector $\{f_{i+1}(U_a) : a \in A\}$, and L_V be the $(s+1) \times m$ matrix whose k th row is the vector $\{g_{k+1}(V_a) : a \in A\}$. Let ℓ_U and ℓ_V denote the vector-valued functions $(f_0, \dots, f_r)^\top$ and $(g_0, \dots, g_s)^\top$. Then, following the argument similar to Lee, Li, and Chiaromonte (2013), we can express (18) as

$$\sum_{b \in B} \|\ell_V(V_b) - L_V L_U^\top (L_U L_U^\top + \varepsilon I_{r+1})^{-1} \ell_U(U_b)\|^2,$$

where $\|\cdot\|$ is the Euclidean norm. If we let G_U be the $(r+1) \times (n-m)$ matrix whose b th column is $\ell_U(U_b)$, and G_V be the $(s+1) \times (n-m)$ matrix whose b th column is $\ell_V(V_b)$, then the above can be further simplified as

$$\|G_V - L_V L_U^\top (L_U L_U^\top + \varepsilon I_{r+1})^{-1} G_U\|^2,$$

where $\|\cdot\|$ is the Frobenius matrix norm. Other matrix norms are also reasonable. In our simulation we used the operator norm, which performed well.

Return now to our cross-validation problem. In the k -fold cross validation, data are divided into k subsets. At each step, one subset is used as the testing set and union of the rest is used as the training set. Let $i, j, \dots$ index the p components of x , $a, b, \dots$ index the n subjects, and ν index the k steps in the k -fold cross-validation.

At the ν th step, let $\{X_a : a \in A\}$ be the training set and $\{X_b : b \in B\}$ be the testing set. Reset κ_i° to be the $\mathbb{R}^m$ -valued function $x^i \mapsto (\kappa(x^i, X_{a_1}^i), \dots, \kappa(x^i, X_{a_m}^i))^\top$. Let $\ell_i, \ell_{(i,j)}$, and ℓ be the vector-valued functions defined by

$$\begin{aligned}\ell_i(x^i) &= (1, \kappa_i^\circ(x^i))^\top, & \ell_{(i,j)}(x^i, x^j) &= (1, \kappa_i^\circ(x^i), \kappa_j^\circ(x^j))^\top, \\ \ell(x) &= (1, \kappa_1^\circ(x^1), \dots, \kappa_p^\circ(x^p))^\top.\end{aligned}$$

Let ℓ_{-i} be the $\mathbb{R}^{n(p-1)+1}$ -valued function obtained by removing κ_i° from ℓ , and $\ell_{-(i,j)}$ be the $\mathbb{R}^{n(p-2)+1}$ -valued function obtained by removing κ_i° and κ_j° from ℓ . Let L_i denote the matrix $(\ell_i(X_{a_1}^i), \dots, \ell_i(X_{a_m}^i))$, and define $L_{-i}, L_{(i,j)}$, and $L_{-(i,j)}$ similarly. Let G_i denote the matrix $(\ell_i(X_{b_1}^i), \dots, \ell_i(X_{b_{n-m}}^i))$, and define $G_{-i}, G_{(i,j)}$, and $G_{-(i,j)}$ similarly.

To determine ε_1 , we use the functions in ℓ_i to predict the functions in ℓ_{-i} at X_b for any $b \in B$. That is,

$$\begin{aligned}\text{CV}_1^{(\nu)}(\varepsilon_1) &= \sum_{i=1}^p \sum_{b \in B} \|\ell_{-i}(X_b^{-i}) - L_{-i} L_i^\top (L_i L_i^\top + \varepsilon_1 I_{m+1})^{-1} \ell_i(X_b^i)\|^2 \\ &= \sum_{i=1}^p \|G_{-i} - L_{-i} L_i^\top (L_i L_i^\top + \varepsilon_1 I_{m+1})^{-1} G_i\|^2.\end{aligned}\quad (19)$$

Repeat this process for $\nu = 1, \dots, k$ and form $\text{CV}_1(\varepsilon_1) = \sum_{\nu=1}^k \text{CV}_1^{(\nu)}(\varepsilon_1)$. We choose ε_1 by minimizing $\text{CV}_1(\varepsilon_1)$ over a grid. To determine ε_2 , we use the functions in $\ell_{-(i,j)}$ to predict the functions in $\ell_{(i,j)}$ at each $X_b, b \in B$, which leads to

$$\text{CV}_2^{(\nu)}(\varepsilon_2) = 2 \sum_{i>j} \|G_{(i,j)} - L_{(i,j)} L_{-(i,j)}^\top (L_{-(i,j)} L_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)+1})^{-1} G_{-(i,j)}\|^2. \quad (20)$$

Let $\text{CV}_2(\varepsilon_2) = \sum_{\nu=1}^k \text{CV}_2^{(\nu)}(\varepsilon_2)$. We choose ε_2 by minimizing $\text{CV}_2(\varepsilon_2)$ over a grid. Here, we use part 1 of Proposition 4 to calculate $(L_{-(i,j)} L_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)+1})^{-1}$.

Alternatively, we can use the generalized cross validation criteria parallel to (19) and (20) to speed up computation (Wahba, 1990). Rather than splitting the data into training and testing samples, we compute the residual sum of squares from the entire data, and regularize it by the effective degrees of freedom as defined by the traces of relevant projection matrices. Specifically, the GCV-criterion parallel to (19) is

$$\text{GCV}_1(\varepsilon_1) = \sum_{i=1}^p \frac{\|L_{-i} - L_{-i} L_i^\top (L_i L_i^\top + \varepsilon_1 I_{n+1})^{-1} L_i\|^2}{\{1 - \text{tr}(L_i^\top (L_i L_i^\top + \varepsilon_1 I_{n+1})^{-1} L_i) / n\}^2}.$$

Here, L_i and L_{-i} are computed from the full data, and G_i in (19) is now replaced by L_i . Similarly, the GCV-criterion parallel to (20) is

$$\text{GCV}_2(\varepsilon_2) = 2 \sum_{i>j} \frac{\|L_{(i,j)} - L_{(i,j)} L_{-(i,j)}^\top (L_{-(i,j)} L_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)+1})^{-1} L_{-(i,j)}\|^2}{\{1 - \text{tr}(L_{-(i,j)}^\top (L_{-(i,j)} L_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)+1})^{-1} L_{-(i,j)}) / n\}^2}.$$

Here, again, the matrix $L_{-(i,j)}^\top (L_{-(i,j)} L_{-(i,j)}^\top + \varepsilon_2 I_{n(p-2)+1})^{-1} L_{-(i,j)}$ is to be calculated by part 2 of Proposition 4 with $V = L_{-(i,j)}$, $s = n(p-2) + 1$ and $u = 1$. Finally, we recommend taking ε_3 to be the same as ε_2 . This is based on the intuition that $K_{-(i,j)} Q K_{-(i,j)}^\top$ and $K_{1:p} Q K_{1:p}^\top$ are of the similar forms and dimensions. This choice works well in our simulation studies.

From our experience, the GCV-criteria give the similar tuning constants as the CV-criteria but increase the computation speed by several folds. For this reason, we used the GCV criteria in our subsequent simulations and application.

9 Simulation studies

In this section we compare our ASG estimators with GGM and GCGM estimators, as proposed by Yuan and Lin (2007) and Liu et al (2013). Since our estimators are based on thresholding the norms of the operators $\Sigma_{X^i X^j | X^{-(i,j)}}$ and $\Theta_{X^i X^j}$, it would be fair to compare them with the thresholded versions of GGM and GCGM. However, because L_1 -regularization is widely used for these estimators, we will instead compare them with the L_1 -regularized GGM and GCGM.

Comparison 1

We first compare the ASG estimators with GGM and GCGM when the copula assumption is violated. Consider the following models:

$$\text{Model I : } X_1 = \varepsilon_1, X_2 = X_1^2 + \varepsilon_2, X_3 = \varepsilon_3,$$

$$\text{Model II : } X_1 = \varepsilon_1, X_2 = \varepsilon_2, X_3 = X_1^2 + \varepsilon_3, X_4 = X_1^2 + X_2^2 + \varepsilon_4, X_5 = \varepsilon_5,$$

where $\varepsilon_1, \dots, \varepsilon_5$ are i.i.d. $N(0, 1)$. The edge set for Model I is $\{(1, 2)\}$; the edge set for Model II is $\{(1, 3), (1, 4), (2, 4)\}$. The sample size is $n = 100$ and we repeated the simulation 100 times and the averaged ROC curves for correct estimation of the edge sets are presented in the upper panels of Figure 2. As expected, our ASG estimators perform much better than the GGM and GCGM when the copula assumption does not hold.

For this and all the other comparisons, we have used the radial basis function (RBF) kernel with $\gamma = 1/(2p)$, where γ is as it appears in the parametric form used in Li, Chun, and Zhao (2012). In Comparison 6 we give some results for alternative kernels and parameters.

Comparison 2

We now consider two models that satisfy the gaussian or copula gaussian assumption, which are therefore favorable to GGM and GCGM, to see how the ASG estimators compare with them:

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

$$\text{Model IV : } (f_1(X^1), \dots, f_p(X^p))^T \sim N(0, \Theta^{-1}),$$

where Θ is the 10×10 precision matrix with diagonal entries 1, 1, 1, 1.333, 3.010, 3.203, 1.543, 1.270, 1.554, and nonzero off-diagonal entries $\theta_{3,5} = 1.418$, $\theta_{4,6} = -1.484$, $\theta_{4,10} = -0.744$, $\theta_{5,9} = 0.519$, $\theta_{5,10} = -0.577$, $\theta_{7,10} = 1.104$, and $\theta_{8,9} = -0.737$. For Model IV, $X^1, \dots, X^p$ each has a marginal gamma distribution with rate parameter 1 and shape parameter generated from $\text{unif}(1, 5)$. The sample size is 50. The averaged ROC curves (over 100 replications) for the 4 estimators are shown in the lower panels in Figure 2. From these curves we can see the ASG estimators are comparable to GGM and GCGM when the gaussian or copula gaussian assumption hold.

The tuning parameters are selected by GCV_1 and GCV_2 from the coarse grids ($5 \times 10^{-4}, \dots, 5 \times 10^3$). The optimal tuning parameters do vary across repetitions, but the ballpark figures are relatively stable. For example, for model IV, the typical values are 5×10^{-4} for ε_1 and 5×10^{-1} for ε_2 . To evaluate the GCV criteria, in Figure 3 we compare the ROC curves using the GCV-tuned $\hat{\varepsilon}_1, \hat{\varepsilon}_2$ with those for $\hat{\varepsilon}_1 \times 10^{-1}$ and $\hat{\varepsilon}_2 \times 10^{-1}$ and those for $\hat{\varepsilon}_1 \times 10$ and $\hat{\varepsilon}_2 \times 10$. We see that smaller tuning parameters performs better at low specificity, larger tuning parameters performs better at high specificity, but the GCV tuning criteria balances these two cases.

Comparison 3

We now investigate the performance of our additive methods under a nonadditive model and compare with a fully nonparametric method that does not rely on the additive assumption. The fully nonparametric method is based on the conditional covariance operator (CCO) introduced by Fukumizu, Bach, and Jordan (2009). Consider the model

$$\text{Model V : } X_i = \varepsilon_i, i = 1, 2, 5; X_3 = X_1 + \varepsilon_3; X_4 = (X_1 + X_2)^2 + \varepsilon_4,$$

where $\varepsilon_1, \dots, \varepsilon_5$ be i.i.d. $N(0, 1)$. The nonadditivity is introduced through the term $(X_1 + X_2)^2$.

Let $\mu : \mathbb{R}^{p-2} \times \mathbb{R}^{p-2} \rightarrow \mathbb{R}$ be a positive definite kernel and $M_{X^{-(i,j)}}$ be the kernel matrix $\{ \mu(X_r^{-(i,j)}, X_s^{-(i,j)}), s = 1, \dots, n \}$. Let

$$G_{X^i} = QK_{X^i}Q, \quad G_{X^j} = QK_{X^j}Q, \quad G_{X^{-(i,j)}} = QM_{X^{-(i,j)}}Q, \\ Q_{X^{-(i,j)}}(\varepsilon_2) = I_n - G_{X^{-(i,j)}}(G_{X^{-(i,j)}} + \varepsilon_2 I_n)^{-1}.$$

Let $V_{X^i X^j | X^{-(i,j)}}$ be the CCO and $\hat{V}_{X^i X^j | X^{-(i,j)}}$ be its sample estimate. Along the lines of Fukumizu, Bach, and Jordan (2009) we derive the operator norm $\hat{V}_{X^i X^j | X^{-(i,j)}}$ as the largest singular value of

$$(G_{X^i}^{1/2} + \varepsilon_1 I_n)^{1/2} Q_{X^{-(i,j)}}(\varepsilon_2) (G_{X^j} + \varepsilon_1 I_n)^{1/2}. \quad (21)$$

We declare no edge between node i and j if this value is smaller than a threshold. A derivation of this is sketched in the Appendix. The tuning parameters are determined the similar GCV criteria as described in Section 8 except that the univariate kernel $K_{X^{-(i,j)}}$ is replaced by the multivariate kernel $M_{X^{-(i,j)}}$. We ran this simulation for $n = 50$ using ACCO, APO, GCGM, and this fully nonparametric estimator (denoted as fuku in the plot). The averaged ROC curves are shown in Figure 4. We can see that our methods work better Fukumizu et al's method even in this nonadditive case.

Comparison 4

In this study we compare the 4 estimators by applying them to the data sets from the DREAM4 Challenges at <http://wiki.c2b2.columbia.edu/dream/index.php?title=D4c2> in which subnetworks are extracted from a real biological networks, and gene expressions are simulated from stochastic differential equations, reflecting their biological interactions. Under this simulation scenario, two genes are not guaranteed to be linearly associated. In the DREAM 4 samples, $p = 10$, $n = 31$ for all five networks. In the analysis, we combined all data sets from various experiments (wildtype, knockout, knockdown and multifactorial) except for the time-course experiments, and apply each method to the combined data sets across all five network configurations in the DREAM4 Challenges. Figure 5 shows the averaged ROC curves for the 5 subnetworks. We note that APO and ACCO perform the best for the first subnetwork, and APO performs the best in the fifth subnetwork. In the other three subnetworks there is no clear separation among the four estimators. These comparisons suggest a possible benefit of our ASG estimators for finding non-linear biological interactions.

Comparison 5

In this study we increase the sample size n and dimension p to investigate the accuracy and computational feasibility of our methods. We use Model IV in Comparison 2 with $(p, n) = (200, 400)$ and $(p, n) = (400, 400)$. We compare APO with the fully nonparametric method and the averaged ROC curves are given in Figure 6. For larger p and n the ACCO is more computationally expensive than APO because one has to compute $\|\hat{\Sigma}_{X^i X^j | X^{-(i,j)}}\|$ for all $i, j = 1, \dots, p$, $i \neq j$. However, this problem is mitigated by APO, where the large $(np \times np)$ matrix $[\Theta]$ only needs to be calculated once.

For $p = 200$ and $n = 400$, we selected (by GCV) $\hat{\varepsilon}_1 = 5 \times 10^{-5}$, $\hat{\varepsilon}_2 = 0.5$. Each run with fixed ε_1 and ε_2 took around 900 seconds (15 mins) in an Intel(R) Core(TM) i7-3770 CPU at 3.40GHz machine. For $p = 400$ and $n = 400$, we selected by GCV $\hat{\varepsilon}_1 = 5e^{-5}$, $\hat{\varepsilon}_2 = 0.5$. Each run with fixed ε_1 and ε_2 took around 1900 seconds in the same computer. The GGM were not used for this simulation, since it took a very long time to get the estimates (more than a day for a single tuning parameter). We also performed 10 replications. It appears that our additive operator performs better than Fukumizu et al's operator in high-dimensional cases.

Comparison 6

We now investigate the behavior of different kernels using Models III and IV. We have experimented with 3 types of kernels: the radial basis function kernel, the polynomial (POLY) kernel, and the rational quadratic (RATIONAL) kernel. The specific parametric forms we used for these kernels are as appear in Li, Chun, and Zhao (2012). For each kernel, we used two tuning parameters: for RBF, we use $\gamma = 1/9, 1/(9p)$; for POLY, we use $r = 2, 3$; for RATIONAL, $c = 200, 400$. The ROC curves for these two models are given in Figure 7. We see that, with the exception of polynomial kernel with degree 3, these performance of these kernels are somewhat similar, with RBF, $\gamma = 1/(9p)$ most favorable. The cubic polynomial kernel does not perform very well because it tends to amplify outlying observations, making their results unstable.

10 Pathway analysis

We applied the four estimators to a real data set from Dimas et al (2009), in which gene expressions were measured from 85 individuals participating in the GenCord project. The same data set was also used in Chun, Zhang, and Zhao (2012). The participants were of Western European origin, and cell lines were obtained from their umbilical cords. In this analysis, we focused on one cell type, the B cells (lymphoblastoid cell lines). As in Chun et al (2012), we mapped the probes to 17,945 autosomal RefSeq genes, and use the resulting gene level data in our analysis. We then extracted the pathway information from the KEGG database (<http://www.genome.jp/kegg/>), and used it for partitioning genes and for evaluating the estimators by taking the interactions in the database as true interactions.

Due to the limited sample sizes, we only used the pathways whose sizes (i.e. number of genes) are greater than 2 and less than 36. There are a total of 300 such pathways. After ignoring directionality of regulations in KEGG pathways, we computed the false positives and true positives of the estimates and computed the area under curves (AUC) of the ROCs. We then evaluated the significance of ROC curves based on the p -values of the Wilcoxon rank sum test. Figure 8 shows the AUCs of all networks by ACCO and APO. The left panel shows AUCs based on ACCO, with significant values highlighted by green circles. The right panel shows those based on APO, with significant values highlighted by blue circles. In each case significance means the p -value is less than 10^{-4} . We see that ACCO yields a larger number of significant ROCs than GGM. To reconfirm this pattern we computed the number of significant ROC curves by ACCO and GGM for three different significance levels, as presented in Table 2, which again shows that the numbers of significant ROC curves by ACCO are consistently higher at all three significance levels.

To provide further insights into the comparison, we listed in Table 3, Panel I, all the pathways with significant ROCs by ACCO (p -value $< 10^{-5}$) whose AUCs by ACCO are larger than those by GGM. There were a total of 6 such pathways. In Table 3, Panel II, we listed all the pathways with significant ROCs by GGM whose AUCs by GGM are larger than those by ACCO. There were a total of 2 such pathways.

Some pathways were selected by more than one methods. Specifically,

$$\begin{aligned} \#(\text{ACCO} \cap \text{APO})=4, & \quad \#(\text{ACCO} \cap \text{GGM})=1, & \quad \#(\text{ACCO} \cap \text{GCGM})=1, \\ \#(\text{APO} \cap \text{GGM})=1, & \quad \#(\text{APO} \cap \text{GCGM})=1, & \quad \#(\text{GGM} \cap \text{GCGM})=1, \end{aligned}$$

where $\#(A \cap B)$ means the number of pathways found by method A and method B . It appears that ACCO and APO yield most similar results. Furthermore, single pathway was found by all methods.

From the analysis of these 300 pathways we see that the ASG estimators perform better than GGM and GCGM in a majority of cases. A plausible interpretation is that intrinsically nonlinear interactions (i.e. those that cannot be described by the gaussian copula assumption) are prevalent among genetic networks, and using a more flexible interaction assumption such as the ASG has real benefits in modeling of gene regulations.

11 Conclusions

In this paper we introduced two new estimators for graphical models based on the notion of additive conditional independence. They substantially broaden the applicability of the current graphical model estimators based on the gaussian or copula gaussian assumption, but at the same time avoid using high-dimensional kernels that can hamper estimation accuracy. As we have demonstrated by simulation and data analysis, the new methods work particularly well when the interactions exhibit intrinsic nonlinearity, which cannot be captured by the gaussian or copula gaussian models, and are comparable to existing methods when the copula gaussian assumption holds.

Our generalization of precision matrix to additive precision operator allows us to adopt essentially the same estimation strategy of the gaussian graphical model at the operator level. Using this generalization we succeeded in preserving the simplicity of the gaussian dependence structure without suffering from its austere distributional restriction.

This basic idea opens up many other possibilities that cannot be explored within the scope of this paper. For example, although both the ACCO and APO estimators are based on truncating small operators (as measured by the operator norm), it is conceivable that more sophisticated sparse techniques, such as LASSO, SCAD, and adaptive LASSO, can be adapted to this setting as well. How to implement these techniques in the operator setting, and how much they would improve the truncation-based estimators presented here, would be interesting questions for future research. Moreover, the asymptotic properties of the proposed estimators, such as consistency and asymptotic distributions, also need to be developed in further studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Bing Li's research was supported in part by NSF grant DMS-1106815.

Hyonho Chun's research was supported in part by NSF grant DMS-1107025.

Hongyu Zhao's research was supported in part by NIH grants R01-GM59507, P01-CA154295, and NSF grant DMS-1106738.

We greatly appreciate the thorough and thought provoking reviews by three referees and an Associate Editor, which contain many useful suggestions and lead to substantial improvements of this work.

References

- Bellman, RE. Dynamic Programming. Princeton University Press; 1957.
- Bach FR. Consistency of the Group Lasso and Multiple Kernel Learning. *Journal of Machine Learning Research*. 2008; 9:1179–1225.
- Bickel PJ, Levina E. Covariance Regularization by Thresholding. *Annals of Statistics*. 2008; 36:2577–2604.
- Chun, H.; Zhang, X.; Zhao, H. Gene regulation network inference with joint sparse Gaussian graphical models. 2012. Unpublished manuscript
- Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009; 325:1246–1250. [PubMed: 19644074]
- Dawid AP. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*. 1979; 41:1–31.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*. 2001; 96:1348–1360.
- Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008; 9:432–441. [PubMed: 18079126]
- Fukumizu K, Bach FR, Jordan MI. Kernel dimension reduction in regression. *Annals of Statistics*. 2009; 4:1871–1905.
- Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. *Biometrika*. 2009; 98:1–15. [PubMed: 23049124]
- Kelley, JL. General Topology. Springer; New York: 1955.
- Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*. 2009; 37:4254–4278. [PubMed: 21132082]
- Lancaster. Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*. 1957; 44:289–292.
- Lauritzen, SL. Graphical Models. Oxford: Clarendon Press; 1996.
- Lauritzen SL, Dawid AP, Larsen BN, Leimer HG. Independence properties of directed Markov fields. *Networks*. 1990; 20:491–505.
- Lee K-Y, Li B, Chiaromonte F. A general theory for nonlinear sufficient dimension reduction: formulation and estimation. *Annals of Statistics*. 2013 In print.
- Li B, Artemiou A, Li L. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Annals of Statistics*. 2011; 36:3182–3210.
- Li B, Chun H, Zhao H. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*. 2012; 107:152–167. [PubMed: 24574574]

- Liu, H.; Han, F.; Zhang, C. Advances in Neural Information Processing System. Vol. 25. Curran Associates, Inc; 2013. Transelliptical Graphical Models; p. 800-808.
- Liu H, Han F, Yuan M, Lafferty J, Wasserman L. High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*. 2013 In print.
- Liu H, Lafferty J, Wasserman L. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*. 2009; 10:2295–2328.
- Meinshausen N, Bühlmann P. High-dimensional graphs with the lasso. *Annals of Statistics*. 2006; 34:1436–1462.
- Pearl J, Geiger D, Verma T. Conditional independence and its representations. *Kybernetika*. 1989; 25:33–44.
- Pearl, J.; Paz, A. Graphoids: a graph-based logic for reasoning about relevancy relations. *Proceedings of the European Conference on Artificial Intelligence*; Brighton, UK. 1986.
- Pearl, J.; Verma, T. The logic of representing dependencies by directed acyclic graphs. *Proceedings of the American Association of Artificial Intelligence Conference*; Seattle, Washington. 1987.
- Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *Journal of American Statistical Association*. 2009; 104:735–746.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Wahba, G. Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics; 1990.
- Xue L, Zou H. Regularized Rank-based Estimation of High-dimensional Nonparanormal Graphical Models. *Annals of Statistics*. 2013 In print.
- Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007; 94:19–35.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

Appendix: proofs

We first state without proof several simple geometric facts that will be useful for proving Theorems 1 and 2. Let $P_{\mathcal{L}}$ be the projection on to a subspace $\mathcal{L}$ of $L_2(P_X)$, U and W be subvectors of X . Then

$$\mathcal{A}_{U \cup W} = \mathcal{A}_U + \mathcal{A}_W. \quad (22)$$

$$P_{\mathcal{A}_{U \cup W}} = P_{\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W} + P_{\mathcal{A}_W}. \quad (23)$$

$$\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W = (I - P_{\mathcal{A}_W})\mathcal{A}_U \quad (24)$$

Proof of Theorem 1

$1 \Leftrightarrow 2$ follows from relation (22); $2 \Leftrightarrow 3$ follows from relation (24); $6 \Leftrightarrow 7$ follows from (23); $4 \Leftrightarrow 5$ follows from the relation (23).

$2 \Rightarrow 4$. By statement 2, $P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}(\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W) = \{0\}$. By relation (24) and the fact that $\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W \perp \mathcal{A}_W$, we have

$$P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}(\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W) = P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}(I - P_{\mathcal{A}_W})\mathcal{A}_U = P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}\mathcal{A}_U.$$

4 $\Rightarrow$ 6. This is because

$$P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}\mathcal{A}_{U \cup W} = P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}(\mathcal{A}_U + \mathcal{A}_W) = P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}\mathcal{A}_W = \{0\},$$

where the first equality follows from (22); the second from assertion 4; the third from $\mathcal{A}_W \perp \mathcal{A}_{V \cup W} \ominus \mathcal{A}_W$.

6 $\Rightarrow$ 2. Because $\ker(P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}) = [\text{ran}(P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W})]^\perp = (\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W)^\perp$, we have

$$\mathcal{A}_{U \cup W} \perp \mathcal{A}_{V \cup W} \ominus \mathcal{A}_W,$$

which implies $\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W \perp \mathcal{A}_{V \cup W} \ominus \mathcal{A}_W$ because $\mathcal{A}_{U \cup W} \ominus \mathcal{A}_W \subseteq \mathcal{A}_{U \cup W}$.

Proof of Theorem 2

Axiom 1 is obvious. To prove 2, 3, 4, let U, V, W, R be disjoint sub-vectors of X and $\mathcal{A}_U, \mathcal{A}_V, \mathcal{A}_W, \mathcal{A}_R$ be their respective additive families.

2. If $U \perp_A V \cup R|W$, then

$$(\mathcal{A}_U + \mathcal{A}_W) \ominus \mathcal{A}_W \perp (\mathcal{A}_{V \cup R} + \mathcal{A}_W) \ominus \mathcal{A}_W.$$

Axiom 2 follows because $(\mathcal{A}_V + \mathcal{A}_W) \ominus \mathcal{A}_W \subseteq (\mathcal{A}_{V \cup R} + \mathcal{A}_W) \ominus \mathcal{A}_W$.

3. By Theorem 1, part 3, $U \perp_A V \cup R|W$ is equivalent to

$$(I - P_{\mathcal{A}_W})\mathcal{A}_U \perp (I - P_{\mathcal{A}_W})\mathcal{A}_{V \cup R} \quad (25)$$

By Axiom 2, we have $U \perp_A R|W$, which, by Theorem 1, part 5, implies $P_{\mathcal{A}_W}f = P_{\mathcal{A}_{W \cup R}}f$ for all $f \in \mathcal{A}_U$. Hence the left hand side of (25) can be rewritten as $(I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_U$. Meanwhile, the right hand side of (25) contains $(I - P_{\mathcal{A}_W})\mathcal{A}_V$. Hence

$$(I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_U \perp (I - P_{\mathcal{A}_W})\mathcal{A}_V. \quad (26)$$

Because $\mathcal{A}_W \subseteq \mathcal{A}_{W \cup R}$, we have $\text{ran}(I - P_{\mathcal{A}_{W \cup R}}) \perp \text{ran}(P_{\mathcal{A}_{W \cup R}})$. Hence

$$(I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_U \perp P_{\mathcal{A}_{W \cup R} \ominus \mathcal{A}_W}\mathcal{A}_V. \quad (27)$$

Because

$$(I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_V = (I - P_{\mathcal{A}_W} - P_{\mathcal{A}_{W \cup R} \ominus \mathcal{A}_W})\mathcal{A}_V \subseteq (I - P_{\mathcal{A}_W})\mathcal{A}_V + P_{\mathcal{A}_{W \cup R} \ominus \mathcal{A}_W}\mathcal{A}_V,$$

relation (26) and the second expression in (27) implies the desired relation

$$(I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_U \perp (I - P_{\mathcal{A}_{W \cup R}})\mathcal{A}_V.$$

4. By Theorem 1, part 3, $U \perp_A V|W$ and $U \perp_A R|W \cup V$ are equivalent to

$$(I - P_{\mathcal{A}_W})\mathcal{A}_U \perp (I - P_{\mathcal{A}_W})\mathcal{A}_V, \quad (I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_U \perp (I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_R. \quad (28)$$

We need to show $(I - P_{\mathcal{A}_W})\mathcal{A}_U \perp (I - P_{\mathcal{A}_W})(\mathcal{A}_V + \mathcal{A}_R)$, which, by the first relation in (28), is implied by

$$(I - P_{\mathcal{A}_W})\mathcal{A}_U \perp (I - P_{\mathcal{A}_W})\mathcal{A}_R. \quad (29)$$

Because $U \perp_A V|W$, the left side of (29) is $(I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_U$. By (23), the right side of (29) is $(I - P_{\mathcal{A}_{V \cup W}} + P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W})\mathcal{A}_R$. So, under (28), relation (29) is equivalent to

$$(I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_U \perp (I - P_{\mathcal{A}_{V \cup W}} + P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W})\mathcal{A}_R.$$

Since $(I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_U \perp (I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_R$, it suffices to show that

$$(I - P_{\mathcal{A}_{V \cup W}})\mathcal{A}_U \perp P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}\mathcal{A}_R, \quad (30)$$

which is true because

$$\text{ran}(P_{\mathcal{A}_{V \cup W} \ominus \mathcal{A}_W}) = \mathcal{A}_{V \cup W} \ominus \mathcal{A}_W = (I - P_{\mathcal{A}_W})(\mathcal{A}_{V \cup W}) = \text{ran}((I - P_{\mathcal{A}_W})P_{\mathcal{A}_{V \cup W}})$$

and $(I - P_{\mathcal{A}_{V \cup W}})(I - P_{\mathcal{A}_W})P_{\mathcal{A}_{V \cup W}} = 0$.

Proof of Theorem 3

The statement $U \perp_A V|W$ means

$$\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle = 0 \quad \forall f \in \mathcal{A}_U, g \in \mathcal{A}_V. \quad (31)$$

Let A , B , and C be the sets of indices for U , V , and W and $h(W)$ be the vector $\{f_k(X^k) : k \in C\}$. By condition 3, f and g can be written as $\sum_{i \in A} \alpha_i f_i$ and $\sum_{j \in B} \beta_j f_j$ for some $\alpha = (\alpha_1, \dots, \alpha_r)^\top$ and $\beta = (\beta_1, \dots, \beta_s)^\top$. Since $(f(U), h(W))$ is jointly gaussian, the conditional expectation $E[f(U)|h(W)]$ belongs to $\mathcal{A}_W$. For the same reason $E[g(V)|h(W)]$ belongs to $\mathcal{A}_W$. It follows that $P_{\mathcal{A}_W} f = E[f(U)|h(W)]$, $P_{\mathcal{A}_W} g = E[g(V)|h(W)]$, and

$$\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle = E[\text{cov}(f(U), g(V)|h(W))] = \text{cov}[f(U), g(V)|h(W)],$$

where the second equality holds because the random vector $(f(U), g(V), h(W))$ is multivariate gaussian. Consequently, (31) is equivalent to $f(U) \perp\!\!\!\perp g(V)|h(W)$ for all $f \in \mathcal{A}_U$ and $g \in \mathcal{A}_V$, which is equivalent to

$$\{f_i(X^i) : i \in A\} \perp\!\!\!\perp \{f_j(X^j) : j \in B\} | h(W).$$

Because $f_1, \dots, f_p$ are injective, this is equivalent to $U \perp\!\!\!\perp V|W$.

Proof of Proposition 1

Let $f \in L_2(P_U)$ and $g \in L_2(P_V)$ be any functions with $\|f\| = \|g\| = 1$. Because $P_{\mathcal{A}_W}$ is idempotent and self adjoint, we have

$$\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle = \langle f, g \rangle - \langle f, P_{\mathcal{A}_W} g \rangle.$$

For each $\gamma = 1, \dots, p-2$, let $\{\psi_{\alpha\gamma}^w : \alpha = 1, 2, \dots\}$ be the hermite polynomials in the variable w^γ . Similarly, let $\{\psi_{\alpha}^u : \alpha = 1, 2, \dots\}$ and $\{\psi_{\beta}^v : \beta = 1, 2, \dots\}$ be the hermite polynomials in the variables u and v . By Lancaster (1957), these polynomials form orthonormal bases in $L_2(P_W)$, $L_2(P_U)$, and $L_2(P_V)$. It follows that the set $\{\psi_{\alpha\gamma}^w : \alpha = 1, 2, \dots, \gamma = 1, \dots, p-2\}$ is an orthonormal basis for $\mathcal{A}_W$. Moreover, by Lancaster (1957), whenever $\alpha \neq \beta$,

$$\{\psi_{\alpha}^u\} \cup \{\psi_{\beta}^v\} \cup \{\psi_{\alpha\gamma}^w : \gamma = 1, \dots, p-2\} \perp \{\psi_{\beta}^u\} \cup \{\psi_{\alpha}^v\} \cup \{\psi_{\beta\gamma}^w : \gamma = 1, \dots, p-2\}.$$

So, if we let $\mathcal{A}_W^\alpha = \text{span}\{\psi_{\alpha\gamma}^w : \gamma = 1, \dots, p-2\}$, then $P_{\mathcal{A}_W} g = \sum_{\alpha=1}^{\infty} P_{\mathcal{A}_W^\alpha} g$. Each term in this sum is

$$P_{\mathcal{A}_W^\alpha} g = P_{\mathcal{A}_W^\alpha} \left(\sum_{\beta=1}^{\infty} \langle g, \psi_{\beta}^v \rangle \psi_{\beta}^v \right) = \sum_{\beta=1}^{\infty} \langle g, \psi_{\beta}^v \rangle P_{\mathcal{A}_W^\alpha} \psi_{\beta}^v = \langle g, \psi_{\alpha}^v \rangle P_{\mathcal{A}_W^\alpha} \psi_{\alpha}^v$$

By Lancaster (1957) again,

$$\langle \psi_\alpha^v, \psi_{\alpha\gamma}^w \rangle = \rho_{VW\gamma}^\alpha, \quad \langle \psi_{\alpha\gamma}^w, \psi_{\alpha\delta}^w \rangle = \rho_{W\gamma W\delta}^\alpha,$$

which yield $P_{\mathcal{A}_W^\alpha} \psi_\alpha^v = R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} \psi_{\alpha*}^w$, where $\psi_{\alpha*}^w = (\psi_{\alpha 1}^w, \dots, \psi_{\alpha, p-2}^w)^\top$. Hence

$$P_{\mathcal{A}_W} g = \sum_{\alpha=1}^{\infty} P_{\mathcal{A}_W^\alpha} g = \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} \psi_{\alpha*}^w.$$

Using this we compute

$$\begin{aligned} \langle f, P_{\mathcal{A}_W} g \rangle &= \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle \langle f, R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} \psi_{\alpha*}^w \rangle \\ &= \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle \sum_{\beta=1}^{\infty} \langle f, \psi_\beta^u \rangle \langle \psi_\beta^u, R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} \psi_{\alpha*}^w \rangle \\ &= \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle \langle f, \psi_\alpha^u \rangle \langle \psi_\alpha^u, R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} \psi_{\alpha*}^w \rangle \\ &= \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle \langle f, \psi_\alpha^u \rangle R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} R_{WU}^{\odot\alpha}. \end{aligned}$$

By the similar calculation we found $\langle f, g \rangle = \sum_{\alpha=1}^{\infty} \langle f, \psi_\alpha^u \rangle \langle g, \psi_\alpha^v \rangle \rho_{UV}^\alpha$. Hence

$$\langle f, g \rangle - \langle f, P_{\mathcal{A}_W} g \rangle = \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle \langle f, \psi_\alpha^u \rangle (\rho_{UV}^\alpha - R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} R_{WU}^{\odot\alpha})$$

Finally, by the Cauchy-Schwarz inequality,

$$\begin{aligned} (\langle f, g \rangle - \langle f, P_{\mathcal{A}_W} g \rangle)^2 &\leq \sum_{\alpha=1}^{\infty} \langle g, \psi_\alpha^v \rangle^2 \sum_{\alpha=1}^{\infty} \langle f, \psi_\alpha^u \rangle^2 (\rho_{UV}^\alpha - R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} R_{WU}^{\odot\alpha})^2 \\ &= \sum_{\alpha=1}^{\infty} \langle f, \psi_\alpha^u \rangle^2 (\rho_{UV}^\alpha - R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} R_{WU}^{\odot\alpha})^2 \\ &\leq \max\{(\rho_{UV}^\alpha - R_{VW}^{\odot\alpha} (R_{WW}^{\odot\alpha})^{-1} R_{WU}^{\odot\alpha})^2 : \alpha=1, 2, \dots\}, \end{aligned}$$

as desired.

Proof of Lemma 1

For any $f \in \mathcal{A}_U$ and $g \in \mathcal{A}_W$, we have

$$\langle (P_{\mathcal{A}_W} |_{\mathcal{A}_U} - \sum_{WU}) f, g \rangle = \langle P_{\mathcal{A}_W} f, g \rangle_{\mathcal{A}_W} - \langle \sum_{WU} f, g \rangle. \quad (32)$$

The first term on the right-hand side is

$$\langle P_{\mathcal{A}_W} f, g \rangle = \langle f, P_{\mathcal{A}_W} g \rangle = \langle f, g \rangle = E[f(U)g(W)].$$

Hence $\langle (P_{\mathcal{A}_W} | \mathcal{A}_U - \Sigma_{WU})f, g \rangle_{\mathcal{A}_W} = 0$ for all $f \in \mathcal{A}_W$ and $g \in \mathcal{A}_U$.

Proof of Theorem 5

If $U \perp_A V|W$, then, for any $f \in \mathcal{A}_U$, $g \in \mathcal{A}_V$,

$$0 = \langle (I - P_{\mathcal{A}_W})f, (I - P_{\mathcal{A}_W})g \rangle = \langle f, g \rangle - \langle f, P_{\mathcal{A}_W}g \rangle, \quad (33)$$

where the second equality follows from the definition of a projection. Using Lemma 1, we calculate the two terms on the right as

$$\langle f, g \rangle = \langle f, \sum_{UV} g \rangle, \quad \langle f, P_{\mathcal{A}_W}g \rangle = \langle f, \sum_{WV} g \rangle = \langle f, \sum_{UW} \sum_{WV} g \rangle.$$

Hence (33) is equivalent to $\langle f, (\Sigma_{UV} - \Sigma_{UW} \Sigma_{WV})g \rangle = 0$, which implies $\Sigma_{UV} - \Sigma_{UW} \Sigma_{WV} = 0$. The arguments can obviously be reversed.

Proof of Proposition 2

1 $\Rightarrow$ 2. For any $f \in \mathcal{A}_{\oplus X}$, $\|f\|_{\mathcal{A}_{\oplus X}} = 1$, we have, by the triangular inequality,

$$\begin{aligned} \|Tf\|_{\mathcal{A}_{\oplus X}} &\leq \sum_{i=1}^p \sum_{j=1}^p \|T_{X^j X^i} f_i\| \\ &\leq \sum_{i=1}^p \sum_{j=1}^p \|T_{X^j X^i}\| \|f_i\| \leq p^2 \max_{i,j=1,\dots,p} \|T_{X^j X^i}\| < \infty. \end{aligned}$$

2 $\Rightarrow$ 1. For any $f_k \in \mathcal{A}_{X^k}$ and $\|f_k\|_{\mathcal{A}_k} = 1$, we have

$$\|Tf_k\|_{\mathcal{A}_{\oplus X}}^2 = \langle Tf_k, Tf_k \rangle_{\mathcal{A}_{\oplus X}} = \sum_{i=1}^p \langle T_{X^i X^k} f_k, T_{X^i X^k} f_k \rangle \geq \|T_{X^{\ell} X^k} f_k\|^2,$$

which implies

$$\begin{aligned} \sup\{\|T_{X^{\ell} X^k} f_k\| : \|f_k\| \leq 1\} &\leq \sup\{\|Tf_k\|_{\mathcal{A}_{\oplus X}} : \|f_k\| \leq 1\} \\ &= \sup\{\|Tf_k\|_{\mathcal{A}_{\oplus X}} : \|f_k\|_{\mathcal{A}_{\oplus X}} \leq 1\} = \|T\| < \infty. \end{aligned}$$

Proof of Lemma 2

First, we show that T_{UU} (and hence also T_{VV}) is bounded, self adjoint, and positive definite.

Let $h \in \mathcal{A}_{\oplus U}$. Then $T_{UU}h \in \mathcal{A}_{\oplus U}$. Hence

$$\langle T_{UU}h, T_{UU}h \rangle_{\mathcal{A}_{\oplus X}} = \langle T_{UU}h, Th \rangle_{\mathcal{A}_{\oplus X}} \leq \|T_{UU}h\|_{\mathcal{A}_{\oplus X}} \|Th\|_{\mathcal{A}_{\oplus X}}$$

which implies that $\|T_{UU}h\|_{\mathcal{A}_{\oplus X}}$ is either 0 or no greater than $\|T\| \|h\|$. Hence T_{UU} is bounded. That T_{UU} is self adjoint follows from $\langle g, T_{UU}h \rangle_{\mathcal{A}_{\oplus X}} = \langle g, Th \rangle_{\mathcal{A}_{\oplus X}}$ for all $g, h \in \mathcal{A}_{\oplus U}$. Because $T > 0$, if $h \in \mathcal{A}_{\oplus U}$ and $h \neq 0$, then $\langle h, Th \rangle_{\mathcal{A}_{\oplus X}} > 0$. Because $\langle h, Th \rangle_{\mathcal{A}_{\oplus X}} = \langle h, T_{UU}h \rangle_{\mathcal{A}_{\oplus X}}$ we have $\langle h, T_{UU}h \rangle_{\mathcal{A}_{\oplus X}} > 0$. Thus $T_{UU} > 0$.

That $T_{UU} - T_{UW}T_{WW}^{-1}T_{WU}$ and $T_{VV} - T_{VW}T_{WW}^{-1}T_{WV}$ are bounded and self adjoint can be proved similarly. To show that they are positive definite and the subsequent identities, let I_{UU} be the identity operator from $\mathcal{A}_{\oplus U}$ to $\mathcal{A}_{\oplus U}$, and O_{UV} be the zero operator from $\mathcal{A}_{\oplus V}$ to $\mathcal{A}_{\oplus U}$. Define I_{VV} and O_{VU} similarly. Then

$$\begin{pmatrix} (T^{-1})_{UU} & (T^{-1})_{UV} \\ (T^{-1})_{VU} & (T^{-1})_{VV} \end{pmatrix} \begin{pmatrix} T_{UU} & T_{UV} \\ T_{VU} & T_{VV} \end{pmatrix} = \begin{pmatrix} I_{UU} & O_{UV} \\ O_{VU} & I_{VV} \end{pmatrix}.$$

Hence

$$(T^{-1})_{UU}T_{UU} + (T^{-1})_{UV}T_{VU} = I_{UU} \quad (34)$$

$$(T^{-1})_{UU}T_{UV} + (T^{-1})_{UV}T_{VV} = O_{UV}. \quad (35)$$

By part 1, T_{VV} is invertible. Hence, by (35),

$$(T^{-1})_{UV} = -(T^{-1})_{UU}T_{UV}T_{VV}^{-1}. \quad (36)$$

Substitute this into (34) to obtain

$$(T^{-1})_{UU}(T_{UU} - T_{UV}T_{VV}^{-1}T_{VU}) = I_{UU} \quad (37)$$

which implies that $T_{UU} - T_{UV}T_{VV}^{-1}T_{VU}$ is invertible and the first two identities in (9) hold. Similarly, we can prove that $T_{VV} - T_{VU}T_{UU}^{-1}T_{UV}$ is invertible and the last two identities in (9) hold.

Proof of Theorem 6

Let A, B, C be the index sets of $U, V, W, Z = U \cup V$, and

$$\begin{aligned} \Upsilon_{ZZ} &= \{\sum_{X^i X^j} : i, j \in A \cup B\}, & \Upsilon_{ZW} &= \{\sum_{X^i X^j} : i \in A \cup B, j \in C\}, \\ \Upsilon_{WZ} &= \{\sum_{X^i X^j} : i \in C, j \in A \cup B\}, & \Upsilon_{WW} &= \{\sum_{X^i X^j} : i \in C, j \in C\}. \end{aligned}$$

These operators are different from the operators such as Σ_{UV} in Section 4 because they are defined by different inner products. For example, Σ_{UU} is the identity mapping but Υ_{UU} is not. By Lemma 2, $\Theta_{ZZ} = (\Upsilon_{ZZ} - \Upsilon_{ZW} \Upsilon_{WW}^{-1} \Upsilon_{WZ})^{-1}$. Because

$$\Upsilon_{ZZ} - \Upsilon_{ZW} \Upsilon_{WW}^{-1} \Upsilon_{WZ} = \begin{pmatrix} \Upsilon_{UU} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WU} & \Upsilon_{UV} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV} \\ \Upsilon_{VU} - \Upsilon_{VW} \Upsilon_{WW}^{-1} \Upsilon_{WU} & \Upsilon_{VV} - \Upsilon_{VW} \Upsilon_{WW}^{-1} \Upsilon_{WV} \end{pmatrix},$$

we have, by Lemma 2 again, $(\Theta_{ZZ}^{-1})_{UV} = -(\Theta_{ZZ}^{-1})_{UU} \Theta_{UV} \Theta_{VV}^{-1}$. This can be rewritten as

$$\Upsilon_{UV} = \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV} = -(\Upsilon_{UU} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WU}) \Theta_{UV} \Theta_{VV}^{-1}. \quad (38)$$

Therefore, $\Theta_{UV} = 0$ iff $\Upsilon_{UV} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV} = 0$.

Now let $f \in \mathcal{A}_U$, $g \in \mathcal{A}_V$. Then, by the definition of a projection,

$$\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle = \langle f, g \rangle - \langle f, P_{\mathcal{A}_W} g \rangle.$$

The first term on the right is $\langle f, \Upsilon_{UV} g \rangle_{\mathcal{A} \oplus X}$. Because, for any $h \in \mathcal{A}_W$,

$$\langle h, \Upsilon_{WW}^{-1} \Upsilon_{WV} g \rangle = \langle h, \Upsilon_{WV} g \rangle_{\mathcal{A} \oplus X} = \langle h, g \rangle,$$

we have $\Upsilon_{WW}^{-1} \Upsilon_{WV} = P_{\mathcal{A}_W} |_{\mathcal{A}_V}$. Hence

$$\langle f, P_{\mathcal{A}_W} g \rangle = \langle f, \Upsilon_{WW}^{-1} \Upsilon_{WV} g \rangle = \langle f, \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV} g \rangle_{\mathcal{A} \oplus X}.$$

That is, we have shown that

$$\langle f - P_{\mathcal{A}_W} f, g - P_{\mathcal{A}_W} g \rangle = \langle f, (\Upsilon_{UV} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV}) g \rangle_{\mathcal{A} \oplus X},$$

which means $\Upsilon_{UV} - \Upsilon_{UW} \Upsilon_{WW}^{-1} \Upsilon_{WV} = 0$ iff $U \perp_{\mathcal{A}} V|W$.

Proof of Lemma 3

The proof of the three equalities are similar. So we only prove the last one. By (10), for any $f \in \mathcal{A}_X^{-(i,j)}$, $g \in \mathcal{A}_X^i$, we have

$$\langle f, \sum_{X^{-(i,j)} X^i} g \rangle_{\mathcal{A}_{X^{-(i,j)}}} = n^{-1} [f]^\top K_{-(i,j)} Q K_{-(i,j)}^\top [\sum_{X^{-(i,j)} X^i} g].$$

Alternatively,

$$\langle f, \sum_{X^{-(i,j)} X^i} g \rangle_{\mathcal{A}_{X^{-(i,j)}}} = \text{cov}_n[f(X), g(X)] = n^{-1} [f]^\top K_{-(i,j)} Q K_i [g].$$

Hence the right-hand sides of the above two equalities are the same for all $[f]$ and $[g]$, which implies the third relation in (11).

Proof of Proposition 4

1. Write (15) as $V = L\Psi R^\top$ where $L = (L_1, L_0)$, $R = (R_1, R_0)$, and $\Psi = \text{diag}(D, 0)$. Because L and R are orthogonal matrices,

$$(VV^\top + \varepsilon I_s)^{-1} = (L\Psi\Psi^\top L^\top + \varepsilon I_s)^{-1} = L(\Psi\Psi^\top + \varepsilon I_s)^{-1} L^\top. \quad (39)$$

By construction,

$$(\Psi\Psi^\top + \varepsilon I_s)^{-1} = \{\text{diag}(D^2 + \varepsilon I_u, \varepsilon I_{s-u})\}^{-1} = \text{diag}\{(D^2 + \varepsilon I_u)^{-1}, \varepsilon^{-1} I_{s-u}\}. \quad (40)$$

Hence the right-hand side of (39) is $L_1(D^2 + \varepsilon I_u)^{-1} L_1^\top + \varepsilon^{-1} L_0 L_0^\top$, which, because L is an orthogonal matrix, can be rewritten as

$$L_1(D^2 + \varepsilon I_u)^{-1} L_1^\top + \varepsilon^{-1} (I_s - L_1 L_1^\top) = L_1((D^2 + \varepsilon I_u)^{-1} - \varepsilon^{-1} I_u) L_1^\top + \varepsilon^{-1} I_s.$$

By (15) we have $V = L_1 D R_1^\top$, and hence $L_1 = V R_1 D^{-1}$. Substitute this into the right-hand side to prove part 1.

2. Multiply both sides of (39) by $V^\top$ from the left and by V from the right, to obtain

$$V^\top (VV^\top + \varepsilon I_s)^{-1} V = R \Psi^\top (\Psi \Psi^\top + \varepsilon I_s)^{-1} \Psi R.$$

Substitute $R \Psi^\top = R_1 D$ and (40) into the right-hand side above to prove part 2.

Proof of Theorem 7

Let $f, g \in \mathcal{A}_{\oplus X}$. Then $f = f_1 + \dots + f_p$ and $g = h_1 + \dots + h_p$ where $f_k, h_k \in \mathcal{A}_X^k$, $k = 1, \dots, p$. Hence

$$\langle f, \Upsilon g \rangle_{\mathcal{A}_{\oplus X}} = n^{-1} \sum_{i=1}^p \langle f_i, h_i \rangle_{\mathcal{A}_{X^i}} = n^{-1} \sum_{i=1}^p [f_i]^\top K_i Q K_i [h_i] \quad (41)$$

Because

$$[f]_{\mathcal{A}_{\oplus X}} = ([f_1]_{\mathcal{A}_{X^1}}^\top, \dots, [f_p]_{\mathcal{A}_{X^p}}^\top)^\top, \quad [\Upsilon g]_{\mathcal{A}_{\oplus X}} = ([h_1]_{\mathcal{A}_{X^1}}^\top, \dots, [h_p]_{\mathcal{A}_{X^p}}^\top)^\top,$$

The right-hand side of (41) can be rewritten in matrix form as

$$n^{-1} [f]^\top \text{diag}(K_1 Q K_1, \dots, K_p Q K_p) [\Upsilon] [g]. \quad (42)$$

Alternatively, the inner product in (41) can be written as

$$\begin{aligned} \langle f, \Upsilon g \rangle_{\mathcal{A}_{\oplus X}} &= n^{-1} \sum_{i=1}^p \sum_{j=1}^p \langle f_i, \sum_{X^i X^j} g_j \rangle_{\mathcal{A}_{X^i}} \\ &= n^{-1} \sum_{i=1}^p \sum_{j=1}^p [f_i]^\top K_i Q K_j [g_j] = n^{-1} [f]^\top K_{1:p} Q K_{1:p}^\top [g]. \end{aligned} \quad (43)$$

Since (42) is equal to (43) for all $[f]$ and $[g]$, equality (16) holds.

Operator norm of CCO

By Fukumizu et al (2009),

$$[\hat{V}_{X^i X^j | X^{-(i,j)}}] = n^{-1} Q_{X^{-(i,j)}} (\varepsilon_2) G_{X^i}. \quad (44)$$

The operator norm of $\hat{V}_{X^i X^j | X^{-(i,j)}}$ is the maximum of $\langle f_i, \hat{V}_{X^i X^j | X^{-(i,j)}} f_j \rangle^2$ subject to $\langle f_i, f_i \rangle = \langle f_j, f_j \rangle = 1$. By linear algebra, this is equivalent to

$$\text{maximizing} \quad \langle f_i, \hat{V}_{X^i X^j | X^{-(i,j)}} \hat{V}_{X^j X^i | X^{-(i,j)}} f_i \rangle \quad \text{subject to} \quad \langle f_i, f_i \rangle = 1. \quad (45)$$

By (44), the first expression in (45) is $n^{-2} [f_i]^\top G_{X^i} Q_{X^{-(i,j)}} (\varepsilon_2) G_{X^i} Q_{X^{-(i,j)}} (\varepsilon_2) G_{X^i} [f_i]$. The second expression in (45) is $[f_i]^\top G_{X^i} [f_i] = 1$. This is a generalized eigenvalue problem. To convert this to an ordinary eigenvalue problem, let $\phi_i = G_{X^i}^{1/2} [f_i]$. Solve this equation for $[f_i]$ with Tychonoff regularization, to obtain $[f_i] = (G_{X^i} + \varepsilon_1 I_n)^{-1/2} \phi_i$. The optimization problem (45) reduces to maximizing (ignoring n^{-2}):

$$\phi_i^\top (G_{X^i} + \varepsilon_1 I_n)^{-1/2} G_{X^i} Q_{X^{-(i,j)}} (\varepsilon_2) G_{X^j} Q_{X^{-(i,j)}} (\varepsilon_2) G_{X^i} (G_{X^i} + \varepsilon_1 I_n)^{-1/2} \phi_i$$

subject to $\phi_i^\top \phi_i = 1$. As we did with ACCO and APO, we use the symmetrized version

$$\phi_i^\top (G_{X^i} + \varepsilon_1 I_n)^{1/2} Q_{X^{-(i,j)}}(\varepsilon_2) (G_{X^j} + \varepsilon_1 I_n) Q_{X^{-(i,j)}}(\varepsilon_2) (G_{X^i} + \varepsilon_1 I_n)^{1/2} \phi_i,$$

which is of the form $AA^\top$ with A being the matrix in (21).

Figure 1.

Violation of copula assumption. The vertical axis is $\Phi^{-1} \circ F_V(V)$. Both U and $\Phi^{-1} \circ F_V(V)$ are $N(0, 1)$ but their joint distribution is strongly non-gaussian.

Figure 2.

Comparison of ACCO, APO, GGM, and GCGM. Upper panels: ROC curves for Models I and II, where the gaussian copula assumption is violated. Lower panels: ROC curves for Models III and IV, where the gaussian copula assumption holds.

Figure 3.
Performance of GCV under Model IV.

Figure 4.
Comparison of ASG and the fully nonparametric estimator on Model V

Figure 5. Comparison of the ROC curves by ACCO, APO, GGM, GCGM estimators on 5 subnetworks in DREAM4 Challenges.

Figure 6.

Comparison of ASG and fully nonparametric method for $p = 200, n = 400$ (left panel) and $p = 400, n = 400$ (right panel) using model IV

Figure 7.
Comparing different kernels using Model III (left panel) and Model IV (right panel).

Figure 8. Comparison of number of significant ROC curves by the ACCO and by GGM estimators for DREAM4 networks with p -values $< 10^{-4}$.

Table 1

Average values of $\tau_{UV|W}$ under $U \perp\!\!\!\perp V|W$

p	k			
	$\lceil 5p/4 \rceil$	$\lceil 6p/4 \rceil$	$\lceil 7p/4 \rceil$	$\lceil 8p/4 \rceil$
20	0.00310	0.00270	0.00114	0.00087
40	0.00472	0.00240	0.00095	0.00051
60	0.00485	0.00185	0.00069	0.00024
80	0.00435	0.00183	0.00053	0.00018
100	0.00483	0.00125	0.00036	0.00018

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Numbers of significant pathways with varying significance levels

<i>p</i> -values	ACCO	APO	GCGM	GGM
$< 10^{-3}$	17	21	14	10
$< 10^{-4}$	10	9	5	5
$< 10^{-5}$	8	7	2	2

Table 3

Significant pathways whose AUCs by ACCO are larger than those by GGM (panel I) and whose AUCs by ACCO are smaller than those by GGM (panel II)

Panel	Pathway ^a	p	AUC		
			ACCO	APO	GGM
I	1	35	0.63	0.59	0.52
	2	24	0.66	0.66	0.56
	3	25	0.61	0.58	0.54
	4	18	0.67	0.60	0.62
	5	18	0.69	0.63	0.62
	6	16	0.69	0.59	0.55
II	7	14	0.60	0.65	0.62
	8	17	0.73	0.71	0.64

^aLabels in this column indicate the following pathways. 1: Activation of Csk by cAMP-dependent Protein Kinase Inhibits Signaling through the T Cell Receptor; 2: Cell Cycle: G2 M Checkpoint; 3: Cyclins and Cell Cycle Regulation; 4: Lck and Fyn tyrosine kinases in initiation of TCR Activation; 5: Th1 Th2 Differentiation; 6: hsa00710; 7: Proteasome Complex pathway; 8: Human Peptide GPCRs.