

A SINful Approach to Gaussian Graphical Model Selection

MATHIAS DRTON AND MICHAEL D. PERLMAN

ABSTRACT. Multivariate Gaussian graphical models are defined in terms of Markov properties, i.e., conditional independences associated with the underlying graph. Thus, model selection can be performed by testing these conditional independences, which are equivalent to specified zeroes among certain (partial) correlation coefficients. For concentration graphs, covariance graphs, acyclic directed graphs, and chain graphs (both LWF and AMP), we apply Fisher's z -transformation, Šidák's correlation inequality, and Holm's step-down procedure, to simultaneously test the multiple hypotheses obtained from the Markov properties. This leads to a simple method for model selection that controls the overall error rate for incorrect edge inclusion. In practice, we advocate partitioning the simultaneous p -values into three disjoint sets, a significant set S , an indeterminate set I , and a non-significant set N . Then our SIN model selection method selects two graphs, a graph whose edges correspond to the union of S and I , and a more conservative graph whose edges correspond to S only. Prior information about the presence and/or absence of particular edges can be incorporated readily.

1. INTRODUCTION

Graphical models use graphs to represent dependencies between stochastic variables. This graphical approach yields dependence models that are easily visualized and communicated. In graphical model selection one wishes to recover the graph that determines the dependence structure of a set of variables from data.

Two approaches to graphical model selection are commonly taken. One is a score-based search in which one moves in the space of considered graphical models and scores the different models by a criterion that captures how well the model fits the observed data; e.g. see Chickering (2002). One such scoring criterion is, for example, the Bayesian Information Criterion. The second approach is to test the conditional independences that are implied by missing edges. This approach has been considered, for example, by Spirtes et al. (2000). In this paper, we follow the latter approach of testing conditional independences and show how in the case of continuous variables with a multivariate normal distribution this multiple testing problem can be solved in a simultaneous way. This leads to a simple method for

Date: April 9, 2004.

Key words and phrases. Graphical model selection, simultaneous tests, concentration graphs, covariance graphs, ADG, DAG, chain graphs.

model selection that controls the overall error rate for incorrect edge inclusion. Since our method is based on the partitioning of the simultaneous p -values into a significant set S , an indeterminate set I , and a non-significant set N , we call this methodology SIN model selection.

This article is organized as follows. Section 2 gives a brief overview of several important types of Gaussian graphical models that have been considered in the literature. In Section 3, we describe and improve SIN model selection for *undirected graphs* \equiv *concentration graphs*, as introduced in Drton and Perlman (2003, 2004). In Section 4, we show how SIN model selection can be adapted to *bidirected graphs* \equiv *covariance graphs*. In Section 5, we turn to the case of *acyclic directed graphs* (ADG \equiv DAG), for which SIN model selection can be carried out if an *a priori* total ordering (e.g. time-ordering) of the variables is available. In Section 6, we consider *chain graphs* (both LWF and AMP), for which SIN model selection is applicable if the variables can be meaningfully blocked into a totally ordered dependence chain. This means that a total ordering of the blocks is specified *a priori*, but no ordering of the variables within each block is specified. Finally, in Section 7, we show how prior information about the absence or presence of certain edges can be incorporated into SIN model selection, illustrated by the case of undirected graphs. Some proofs are deferred to the Appendix.

2. GAUSSIAN GRAPHICAL MODELS

Let $Y := (Y_1, \dots, Y_p)^t \in \mathbb{R}^p$ be a random vector distributed according to the multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$. It is assumed throughout that the covariance matrix Σ is nonsingular. Let $G := (V, E)$ be a graph with vertex set $V := \{1, \dots, p\}$ and edge set E . If the vertices V of this graph are identified with the variables Y_1, \dots, Y_p , then the edge set E induces conditional independences via so-called Markov properties. In this section, we review several different types of graphs that have been considered in the literature and show how their Markov properties determine statistical models. For introductions to graphical models see the monographs by Edwards (2000), Lauritzen (1996), and Whittaker (1990) that provide general theory as well as methodology for parameter estimation and model testing.

2.1. Undirected graphs. Let $G = (V, E_-)$ be an undirected graph (UG) whose edges are specified by the edge indicators $E_- = (e_{i-j} \mid 1 \leq i \neq j \leq p)$, where $e_{i-j} = e_{j-i} = 1$ or 0 according to whether vertices i and j are adjacent in G or not. The *pairwise undirected Markov property* determined by G states that for all $1 \leq i < j \leq p$,

$$(2.1) \quad e_{i-j} = 0 \quad \implies \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{V \setminus \{i,j\}}.$$

(For the UG in Figure 1(a), (2.1) specifies that $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2, Y_3$ and $Y_2 \perp\!\!\!\perp Y_3 \mid Y_1, Y_4$.) Since $Y \sim \mathcal{N}_p(\mu, \Sigma)$,

$$(2.2) \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{V \setminus \{i,j\}} \quad \iff \quad \rho_{ij \cdot V} = 0,$$

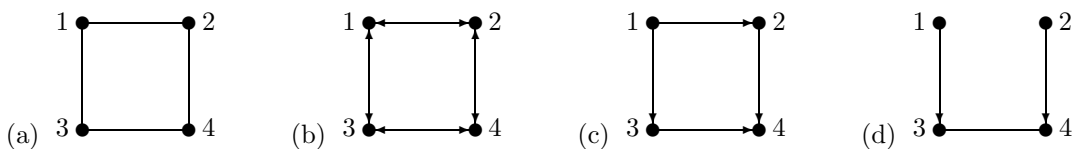


FIGURE 1. (a) An undirected graph, (b) a bidirected graph, (c) an acyclic directed graph, and (d) a chain graph.

where

$$(2.3) \quad \rho_{ij.V} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$$

denotes the ij -th partial correlation, i.e., the correlation between Y_i and Y_j in their conditional distribution given $Y_{V \setminus \{i,j\}}$; compare Lauritzen (1996, p.130). Here, $\Sigma^{-1} := \{\sigma^{ij}\}$ is the *concentration* \equiv *precision* matrix.

Therefore, for an UG $G = (V, E_-)$, the Gaussian graphical model $N_{\text{UG}}(G)$ can be defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ that satisfy the graphical partial correlation restrictions

$$(2.4) \quad e_{i-j} = 0 \quad \implies \quad \rho_{ij.V} = 0$$

for all $1 \leq i < j \leq p$. The model $N_{\text{UG}}(G)$ has been called a *covariance selection model* (Dempster, 1972) and a *concentration graph model* (Cox and Wermuth, 1996). The latter name reflects the fact that $N_{\text{UG}}(G)$ can easily be parameterized using the concentration matrix Σ^{-1} . Statistical theory and inference methodology for Gaussian UG models is well-developed (Edwards, 2000; Lauritzen, 1996; Whittaker, 1990).

2.2. Bidirected graphs. Now let $G = (V, E_{\leftrightarrow})$ be a bidirected graph (BG) with edges represented by the edge indicators $E_{\leftrightarrow} = (e_{i \leftrightarrow j} \mid 1 \leq i \neq j \leq p)$, where $e_{i \leftrightarrow j} = e_{j \leftrightarrow i} = 1$ or 0 according to whether vertices i and j are adjacent in G or not. The *pairwise bidirected Markov property* determined by G states that for all $1 \leq i < j \leq p$,

$$(2.5) \quad e_{i \leftrightarrow j} = 0 \quad \implies \quad Y_i \perp\!\!\!\perp Y_j.$$

(For the BG in Figure 1(b), (2.6) specifies that $Y_1 \perp\!\!\!\perp Y_4$ and $Y_2 \perp\!\!\!\perp Y_3$.) Obviously, since $Y \sim \mathcal{N}_p(\mu, \Sigma)$,

$$(2.6) \quad Y_i \perp\!\!\!\perp Y_j \quad \iff \quad \rho_{ij} = 0,$$

where

$$(2.7) \quad \rho_{ij} := \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

denotes the ij -th correlation, i.e., the correlation between Y_i and Y_j .

Therefore, for a bidirected graph $G = (V, E_{\leftrightarrow})$ the Gaussian graphical model $N_{\text{BG}}(G)$ can be defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ such that the covariance matrix Σ satisfies the graphical correlation restrictions

$$(2.8) \quad e_{i \leftrightarrow j} = 0 \quad \implies \quad \rho_{ij} = 0$$

for all $1 \leq i < j \leq p$. The model $N_{\text{BG}}(G)$ has also been called a *covariance graph model* (Cox and Wermuth, 1996). (Cox and Wermuth (1993, 1996) use dashed lines instead of bidirected edges.) For results regarding the Markov properties of bidirected graphs see Pearl and Wermuth (1994), Kauermann (1996), Banerjee and Richardson (2003), and Richardson (2003). Note also that BG models are a special case of ancestral graph models, which were introduced by Richardson and Spirtes (2002).

2.3. Acyclic directed graphs. A directed graph is called *acyclic* if it contains no directed cycles: $i \rightarrow \cdots \rightarrow i$. Let $G = (V, E_{\rightarrow})$ be such an acyclic directed graph (ADG), with edge indicators $E_{\rightarrow} = (e_{i \rightarrow j} \mid 1 \leq i \neq j \leq p)$, where $e_{i \rightarrow j} = 1$ or 0 according to whether or not there is an edge pointing from vertex i to j . Note that, unlike the UG and BG cases, $e_{i \rightarrow j} = 1$ implies $e_{j \rightarrow i} = 0$.

The directed edges E_{\rightarrow} define a partial ordering \preceq of the vertices $V = \{1, \dots, p\}$ in which $i \preceq j$ if $i = j$ or there is a directed path $i \rightarrow \cdots \rightarrow j$ from i to j in G . Not all pairs of vertices must be ordered with respect to this partial ordering, e.g., the pair of vertices (2, 3) in the graph in Figure 1(c). However, the vertices can be re-labelled so that $i < j$ whenever $i \prec j$; such a labelling is called a *well-numbering* of the vertex set V for G . In the sequel, we assume that, without loss of generality, the vertex set V is well-numbered. This is the case for the graph in Figure 1(c). The *well-numbered directed pairwise Markov property* determined by G states that for all $1 \leq i < j \leq p$,

$$(2.9) \quad e_{i \rightarrow j} = 0 \quad \implies \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{\{1, \dots, j\} \setminus \{i, j\}},$$

compare Edwards (2000, Eqn. (7.2)), also Appendix A. Here, conditional independence given Y_{\emptyset} is understood to be ordinary independence. (In the example of Figure 1(c), (2.9) specifies that $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2, Y_3$ and $Y_2 \perp\!\!\!\perp Y_3 \mid Y_1$.) Since $Y \sim \mathcal{N}_p(\mu, \Sigma)$, for $1 \leq i < j \leq p$ it holds that

$$(2.10) \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{\{1, \dots, j\} \setminus \{i, j\}} \quad \iff \quad \rho_{ij, \{1, \dots, j\}} = 0,$$

where for $\{i, j\} \subseteq K \subseteq V$ we define $\rho_{ij, K}$ to be the partial correlation of Y_i and Y_j given $(Y_k \mid k \in K \setminus \{i, j\})$. Clearly $\rho_{ij, K}$ is a function of the $K \times K$ submatrix of Σ ; compare (2.3).

Let $G = (V, E_{\rightarrow})$ be an ADG with well-numbered vertex set V . The Gaussian graphical model $N_{\text{ADG}}(G)$ can be defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ such that the graphical restrictions

$$(2.11) \quad e_{i \rightarrow j} = 0 \quad \implies \quad \rho_{ij, \{1, \dots, j\}} = 0$$

hold for all $1 \leq i < j \leq p$. The model $N_{\text{ADG}}(G)$ can be parameterized in terms of the Choleski decomposition of Σ^{-1} , which is equivalent to a parameterization in terms of regression parameters and residual variances (Wermuth, 1980; Andersson and Perlman, 1998). The definition of $N_{\text{ADG}}(G)$ does *not* depend on the choice of the underlying well-numbering, which is not unique in general. This follows because for multivariate normal distributions the well-numbered directed pairwise Markov property is equivalent to the well-numbered Markov property and thus equivalent to the global directed Markov property, which clearly does not depend on the choice of the well-numbering (see Theorem 3 in Appendix A, Cowell et al. (1999, Thm. 5.14), and Edwards (2000, §7.1.1)).

2.4. Chain graphs. Consider now a graph $G = (V, E_{\rightarrow}^-)$, which may have both directed and undirected edges. We represent the edges by a pair of edge indicators $E_{\rightarrow}^- = (E_{\rightarrow}, E_{-})$, where $E_{\rightarrow} = (e_{i \rightarrow j} \mid 1 \leq i \neq j \leq p)$ indicates the directed edges and $E_{-} = (e_{i-j} \mid 1 \leq i \neq j \leq p)$ indicates the undirected edges. For any $i, j \in V$, the edge indicators must satisfy

$$(2.12) \quad (e_{i-j} = e_{j-i}) \quad \wedge \quad (e_{i \rightarrow j} = 1 \implies e_{j \rightarrow i} = 0).$$

A graph with directed and/or undirected edges that satisfies (2.12) is called a *chain graph* (CG) iff it has no partially directed cycles, which are paths from one vertex to itself in which there is at least one directed edge and all directed edges point in the same direction. The *chain components* of a CG G are the connected components of the UG obtained by deleting all directed edges in G , that is, the UG over V with edge set E_{-} .

Let $\mathbb{D} = (B_k \mid 1 \leq k \leq q)$, $q \leq p$, be a family of pairwise disjoint blocks of vertices $B_k \neq \emptyset$ such that

1. $\cup(B_k \mid 1 \leq k \leq q) = V$, and
2. each B_k is a union of chain components.

The partitioning \mathbb{D} of the vertex set V is called a *dependence chain* for G if

$$(2.13) \quad k < \ell \implies e_{j \rightarrow i} = 0 \quad \forall i \in B_k, j \in B_{\ell}.$$

In other words, any edges of G between B_k and B_{ℓ} , $k < \ell$, are directed and point from a vertex in B_k to a vertex in B_{ℓ} (see Figure 2). For a dependence chain $\mathbb{D} = (B_k \mid 1 \leq k \leq q)$,

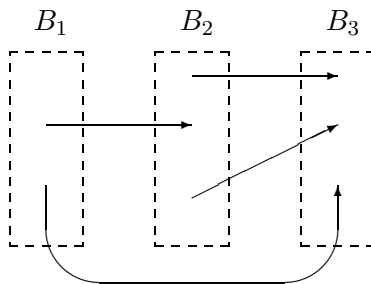


FIGURE 2. Illustration of a dependence chain $\mathbb{D} = (B_1, B_2, B_3)$.

we define the *cumulatives* to be the sets $C_\ell := \cup(B_k \mid k \leq \ell)$, where $0 \leq \ell \leq q$ and $C_0 := \emptyset$. For each $i \in V$, let $1 \leq k(i) \leq q$ be the index such that $i \in B_{k(i)}$. In the sequel, we assume, without loss of generality, that the vertices are labelled such that

$$(2.14) \quad k(i) < k(j) \implies i < j.$$

Since no partially directed cycles occur in a CG, there always exists a dependence chain but, in general, this dependence chain will not be unique. For example, consider the graph with no edges, in which the chain components are singletons and any partitioning of the vertices defines a dependence chain. Less trivially, for the CG depicted in Figure 1(d), a possible dependence chain is $\mathbb{D} = (B_1, B_2)$ with $B_1 = \{1, 2\}$ and $B_2 = \{3, 4\}$. But this is not the only possible choice since both $\tilde{\mathbb{D}} = (\tilde{B}_1, \tilde{B}_2, \tilde{B}_3)$ with $\tilde{B}_1 = \{1\}$, $\tilde{B}_2 = \{2\}$ and $\tilde{B}_3 = \{3, 4\}$, and $\bar{\mathbb{D}} = (\bar{B}_1, \bar{B}_2, \bar{B}_3)$ with $\bar{B}_1 = \{2\}$, $\bar{B}_2 = \{1\}$, and $\bar{B}_3 = \{3, 4\}$ are also dependence chains. Fortunately, the statistical models defined below do not depend on the choice of the dependence chain.

In the literature two alternative Markov properties for CGs have been considered. The LWF Markov property is due to Lauritzen and Wermuth (1989) and Frydenberg (1990), and the more recent AMP Markov property is studied in Andersson et al. (2001). Yet other Markov properties are discussed in Cox and Wermuth (1996).

2.4.1. The LWF Markov property for chain graphs. Let $G = (V, E_-)$ be a CG and \mathbb{D} an associated dependence chain. The *LWF chain-recursive pairwise Markov property* (Lauritzen, 1996, p.54) for G states that

$$(2.15) \quad (e_{i \rightarrow j} = 0) \wedge (e_{i-j} = 0) \implies Y_i \perp\!\!\!\perp Y_j \mid Y_{C_{k(j)} \setminus \{i, j\}}$$

for all $1 \leq i < j \leq p$; recall (2.14). (For the graph in Figure 1(d), (2.15) specifies that $Y_1 \perp\!\!\!\perp Y_2$, $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2, Y_3$, and $Y_2 \perp\!\!\!\perp Y_3 \mid Y_1, Y_4$.) The Gaussian graphical model $N_{\text{LWF}}(G)$ can be defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ such that

$$(2.16) \quad (e_{i \rightarrow j} = 0) \wedge (e_{i-j} = 0) \implies \rho_{ij.C_{k(j)}} = 0$$

for all $1 \leq i < j \leq p$. If G is in fact an UG then $N_{\text{LWF}}(G) = N_{\text{UG}}(G)$. Similarly, if G is an ADG then $N_{\text{LWF}}(G) = N_{\text{ADG}}(G)$. The model $N_{\text{LWF}}(G)$ can be parameterized as described in Lauritzen (1996, pp.154–155) and in Andersson et al. (2001, p.35).

The restrictions (2.16) that define the model $N_{\text{LWF}}(G)$ depend on the cumulatives C_k and thus on the choice of the dependence chain. However, the model $N_{\text{LWF}}(G)$ itself does *not* depend on the choice of dependence chain (cf. Lauritzen (1996, Thm. 3.34); Lauritzen and Wermuth (1989, §8); Frydenberg (1990, Thm. 3.5)).

2.4.2. The AMP Markov property for chain graphs. The *AMP chain-recursive pairwise Markov property* for a CG $G = (V, E_-)$ states that

$$(2.17) \quad e_{i-j} = 0 \implies Y_i \perp\!\!\!\perp Y_j \mid Y_{C_{k(j)} \setminus \{i, j\}}$$

for all $1 \leq i < j \leq p$ such that $k(i) = k(j)$, and

$$(2.18) \quad e_{i \rightarrow j} = 0 \quad \implies \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{C_{k(j)-1} \setminus \{i\}},$$

for all $1 \leq i < j \leq p$ such that $k(i) < k(j)$. (For the graph in Figure 1(d), (2.17) and (2.18) specify that $Y_1 \perp\!\!\!\perp Y_2$, $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2$, and $Y_2 \perp\!\!\!\perp Y_3 \mid Y_1$.) The Gaussian graphical model $N_{\text{AMP}}(G)$ can be defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ such that

$$(2.19) \quad e_{i-j} = 0 \quad \implies \quad \rho_{ij.C_{k(j)}} = 0$$

for all $1 \leq i < j \leq p$ with $k(i) = k(j)$, and

$$(2.20) \quad e_{i \rightarrow j} = 0 \quad \implies \quad \rho_{ij.C_{k(j)-1} \cup \{j\}} = 0$$

for all $1 \leq i < j \leq p$ with $k(i) < k(j)$. As for the LWF models, if G is an UG then $N_{\text{AMP}}(G) = N_{\text{UG}}(G)$, and if G is an ADG then $N_{\text{AMP}}(G) = N_{\text{ADG}}(G)$. The model $N_{\text{AMP}}(G)$ can be parameterized as described in Andersson et al. (2001, p.35 and §5).

As in the LWF case, the restrictions (2.19) and (2.20) that define the model $N_{\text{AMP}}(G)$ depend on the cumulatives C_k and thus on the choice of the dependence chain. However, the model $N_{\text{AMP}}(G)$ itself does *not* depend on the choice of dependence chain because, as shown in the Appendix A, for Gaussian distributions the AMP chain-recursive pairwise Markov property is equivalent to the AMP global Markov property, which only depends on the underlying CG G .

2.5. Remarks. We have defined five types of graphical models by pairwise Markov properties that associate conditional independences with missing edges. For each of these models, in the Gaussian case, the pairwise Markov property is equivalent to the global Markov property, which allows one to read off the graph G all conditional independences that hold for all distributions in the graphical model based on G . Global and other Markov properties for UGs, ADGs, and LWF CGs are discussed for example in Frydenberg (1990), Lauritzen (1996, §3), and Cowell et al. (1999, §5), whereas more recent results on bidirected graphs and AMP CGs can be found in Richardson (2003) and Andersson et al. (2001), respectively.

In the following sections, for each of these five types of Gaussian graphical models, we describe a simple model selection methodology based on simultaneous testing of the multiple conditional independence hypotheses that define the model. This strategy is applicable provided we need to consider only one conditional independence hypothesis for each possible adjacency in the graph. For UGs and BGs this poses no restriction but an *a priori* total ordering of the variables must be assumed for an ADG, while for a CG, an *a priori* total ordering of blocks of variables must be assumed. In the literature, this ordering or blocking strategy has been advocated in several case studies (Caputo et al., 1999, 2003; Cox and Wermuth, 1996; Didelez et al., 2002; Mohamed et al., 1998; Whittaker, 1990); compare also Wermuth and Lauritzen (1990). The requirement of only one conditional independence hypothesis per possible adjacency does not allow us to deal with the ancestral graphs of

Richardson and Spirtes (2002) in the same way as with the models presented in Sections 2.1-2.4.

3. SIN FOR GAUSSIAN UNDIRECTED GRAPH MODELS

3.1. Methodology for undirected graphs. Let $Y^{(1)}, \dots, Y^{(n)}$ be a sample from the model $N_{\text{UG}}(G)$, where $G = (V, E_-)$ is an unknown UG. The sample information can be summarized by the sufficient statistics

$$(3.1) \quad \bar{Y} := \frac{1}{n} \sum_{m=1}^n Y^{(m)} \in \mathbb{R}^V,$$

the sample mean vector, and

$$(3.2) \quad W := \frac{1}{n-1} \sum_{m=1}^n (Y^{(m)} - \bar{Y})(Y^{(m)} - \bar{Y})^t \in \mathbb{R}^{V \times V},$$

the sample covariance matrix. We assume $n \geq p+1$ in order to guarantee positive definiteness of W . The method presented later in this section requires that $n \geq p+2$.

The definition of the model $N_{\text{UG}}(G)$ in (2.1) and the equivalence (2.2) suggest that we can perform model selection, i.e., recover the graph G , by considering the $p(p-1)/2$ testing problems

$$(3.3) \quad H_{ij.V} : \rho_{ij.V} = 0 \quad \text{vs.} \quad K_{ij.V} : \rho_{ij.V} \neq 0 \quad (1 \leq i < j \leq p).$$

Testing these hypotheses is the first step in stepwise model selection procedures (Edwards, 2000, §6.1, also see §3.1). In *backward stepwise selection*, each hypothesis in (3.3) is tested individually at a fixed significance level α . The largest of the p -values for the hypotheses that are not rejected is determined and the associated edge is removed from the graph. In the next step the remaining edges/hypotheses are tested again in the reduced graph, also at level α . The procedure stops if all remaining hypotheses are rejected at level α . However, stepwise selection procedures “may be regarded as a misuse of significance testing, since the overall error properties are not related in any clear way to the error levels of the individual tests” (Edwards, 2000, p.158). Again on p.172: “Its sampling properties seem to be intractable.”

Instead, in order to provide an alternative model selection procedure that does control the overall error rate (with respect to inclusion of edges), the hypotheses in (3.3) can be tested *simultaneously*, as proposed by Drton and Perlman (2004). Their approach, based on Fisher’s z -transform and Šidák’s (1967) correlation inequality is now described. (Compare also Larntz and Perlman (1988) who use the same idea in a different context.)

Let $r_{ij.V}$ denote the *sample partial correlation* based on W ; recall (2.3). The $(r_{ij.V} \mid 1 \leq i < j \leq p)$ are smooth functions of W^{-1} (Anderson, 2003, Exercise 2.47) and therefore of W . The random matrix W has a Wishart distribution and is asymptotically normal with mean Σ . Thus, it follows from the delta method (Shorack, 2000, §11.6) that the

joint distribution of the $(r_{ij.V} \mid 1 \leq i < j \leq p)$ is also asymptotically normal with mean $(\rho_{ij.V} \mid 1 \leq i < j \leq p)$. The marginal distribution of $r_{ij.V}$ has the same form as the distribution of the ordinary sample correlation r_{ij} , but with the parameter ρ_{ij} replaced by $\rho_{ij.V}$ and the degrees of freedom reduced from $n - 1$ to $(n - 1) - (p - 2) = n - p + 1$ (Anderson, 2003, Thm. 4.3.5).

Let

$$(3.4) \quad z : (-1, 1) \rightarrow \mathbb{R}, \quad r \mapsto \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

be the z -transform. For moderate or large values of n , the z -transform of $r_{ij.V}$, given by

$$(3.5) \quad z_{ij.V} = z(\rho_{ij.V}),$$

substantially increases the accuracy of the normal approximation to the marginal distribution, as follows (Anderson, 2003, §4.2.3):

$$(3.6) \quad \sqrt{n_p} (z_{ij.V} - \zeta_{ij.V}) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $n_p := (n - 3) - (p - 2) = n - p - 1$ and $\zeta_{ij.V} = z(\rho_{ij.V})$. Note that

$$(3.7) \quad \zeta_{ij.V} = 0 \iff \rho_{ij.V} = 0.$$

Furthermore, (3.5) is a *variance-stabilizing* transformation in that the asymptotic variance in (3.6) depends on no unknown parameters. Note that by the delta method the $z_{ij.V}$ are also asymptotically *jointly* normal.

By Šidák's inequality applied to the asymptotic joint normal distribution of the $z_{ij.V}$, under the model $N_{UG}(G)$ the following inequality holds approximately for large n and all $c > 0$:

$$(3.8) \quad \Pr_G[|z_{ij.V} - \zeta_{ij.V}| \leq c, 1 \leq i < j \leq p] \geq \prod_{1 \leq i < j \leq p} \Pr_G[|z_{ij.V} - \zeta_{ij.V}| \leq c].$$

Note that if $b_p(\alpha)$ is determined by

$$(3.9) \quad [2\Phi(b_p(\alpha)) - 1]^{p(p-1)/2} = 1 - \alpha,$$

or equivalently by

$$(3.10) \quad b_p(\alpha) = \Phi^{-1} \left(\frac{1}{2} [(1 - \alpha)^{2/p(p-1)} + 1] \right),$$

with Φ the standard normal cumulative distribution function, then (3.8) implies that

$$(3.11) \quad \Pr_G[|z_{ij.V} - \zeta_{ij.V}| \leq n_p^{-1/2} b_p(\alpha), 1 \leq i < j \leq p] \geq 1 - \alpha.$$

If the hypothesis $H_{ij.V}$ is true and $\zeta_{ij.V} = 0$ (recall (3.7)) then it follows from (3.10) that

$$(3.12) \quad |z_{ij.V}| \leq n_p^{-1/2} b_p(\alpha) \iff \alpha \leq \pi_{ij.V} := 1 - [2\Phi(\sqrt{n_p} |z_{ij.V}|) - 1]^{p(p-1)/2}.$$

The p -values $(\pi_{ij.V} \mid 1 \leq i < j \leq p)$ can be used for model selection as follows. If $\pi_{ij.V} \geq \alpha$ then the data are *compatible* with the hypothesis $H_{ij.V} : \rho_{ij.V} = 0$ at the overall

confidence level $1 - \alpha$, so the edge $i - j$ is *not included* in the selected graph. If $\pi_{ij.V} < \alpha$ then the data are *not compatible* with the hypothesis $H_{ij.V} : \rho_{ij.V} = 0$ at the overall confidence level $1 - \alpha$, so the edge $i - j$ is *included* in the selected graph. Formally, we can define the selected graph $\hat{G}(\alpha) := (V, \hat{E}_-(\alpha))$ by setting

$$(3.13) \quad \hat{E}_-(\alpha) := (\hat{e}_{i-j}(\alpha) \mid 1 \leq i \neq j \leq p),$$

where, for $1 \leq i < j \leq p$,

$$(3.14) \quad \hat{e}_{i-j}(\alpha) = \hat{e}_{j-i}(\alpha) = \begin{cases} 0 & \text{if } \pi_{ij.V} \geq \alpha, \\ 1 & \text{if } \pi_{ij.V} < \alpha. \end{cases}$$

By (3.11) and (3.12), $(\pi_{ij.V} \mid 1 \leq i < j \leq p)$ is a set of *conservative simultaneous p-values* for (3.3) in the sense that

$$(3.15) \quad \Pr_{G_\emptyset}[\pi_{ij.V} \geq \alpha, 1 \leq i < j \leq p] \geq 1 - \alpha,$$

where $G_\emptyset = (V, E_\emptyset)$ denotes the graph with no edges, and $N_{\text{UG}}(G_\emptyset)$ is the model of complete independence. *A fortiori*, for a general UG $G = (V, E_-)$ with $E_- = (e_{i-j} \mid 1 \leq i \neq j \leq p)$,

$$(3.16) \quad \Pr_G[\pi_{ij.V} \geq \alpha \forall ij \in E_0] \geq 1 - \alpha,$$

where $E_0 = \{ij \mid e_{i-j} = 0\}$ indicates the set of edges *absent* in G .

The inequality (3.16) yields results concerning the overall error rate of our proposed model selection procedure. If $G = (V, E_-)$ and $G' = (V, E'_-)$ are two graphs with the same vertex set V , then G' is a *subgraph* of G , denoted by $G' \subseteq G$, if G' is obtained by removing one or more edges from G , that is, if $E'_- \subseteq E_-$. It is readily seen from (2.4) that $N_{\text{UG}}(G')$ is a *submodel* of $N_{\text{UG}}(G)$, i.e., $N_{\text{UG}}(G') \subseteq N_{\text{UG}}(G)$, iff $G' \subseteq G$.

It follows directly from (3.16) that

$$(3.17) \quad \Pr_G[\hat{G}(\alpha) \subseteq G] \geq 1 - \alpha,$$

up to the accuracy of the normal approximations involved above. Thus, with probability at least $1 - \alpha$ our selection procedure correctly identifies all pairwise conditional independences in the true model. Therefore, the overall error rate for incorrect edge inclusion is controlled.

Furthermore, the selection procedure with fixed simultaneous significance level α is $(1 - \alpha)$ -*consistent* in the following sense. Let G_{faithful} be the graph that is inclusion-minimal amongst all UGs G' for which the data-generating distribution $\mathcal{N}(\mu, \Sigma) \in N_{\text{UG}}(G')$. The distribution $\mathcal{N}(\mu, \Sigma)$ is called *faithful* to G_{faithful} since $\rho_{ij.V} = 0$ iff $e_{i-j} = 0$. Then, asymptotically, (cf. Drton and Perlman, 2004, eqn. (2.18)),

$$(3.18) \quad \liminf_{n \rightarrow \infty} \Pr_G[\hat{G}(\alpha) = G_{\text{faithful}}] \geq 1 - \alpha,$$

which means that the selection procedure identifies G_{faithful} with asymptotic probability at least $(1 - \alpha)$. Moreover, if the sample size n can be chosen large enough, then the asymptotic probability $\Pr_G[\hat{G}(\alpha) \neq G_{\text{faithful}}]$ can be made arbitrarily small by choosing α arbitrarily small (cf. Drton and Perlman, 2004, eqn. (2.19)-(2.21)). In this sense our

procedure is fully consistent. However, the choice of n also depends on the unknown parameter $\min\{|\zeta_{ij.V}| \mid 1 \leq i \neq j \leq p, e_{i-j}^{\text{faithful}} = 1\}$, where $\{e_{i-j}^{\text{faithful}}\}$ are the edge indicators in G_{faithful} .

3.2. Improvement by Holm's step-down procedure. Based on the inequality of Šidák (1967), the p -values $\pi_{ij.V}$ defined in (3.12) allow us to select the graph $\hat{G}(\alpha)$ such that edge inclusion is controlled in the sense of (3.17). However, the step-down procedure of Holm (1979) improves the p -values $\pi_{ij.V}$ while still allowing valid simultaneous testing of the multiple hypotheses in (3.3).

Let

$$(3.19) \quad m_{ij.V} := |\{\pi_{k\ell.V} \mid 1 \leq k < \ell \leq p, \pi_{k\ell.V} < \pi_{ij.V}\}|$$

be the rank of the p -value $\pi_{ij.V}$ (ties amongst the p -values can be ignored since they comprise nullsets). We define the *Holm's adjusted p -values*

$$(3.20) \quad \pi_{ij.V}^H := \max\{1 - (1 - \pi_{k\ell.V})^{t_{k\ell.V}} \mid 1 \leq k < \ell \leq p, m_{k\ell.V} \leq m_{ij.V}\},$$

where

$$(3.21) \quad t_{ij.V} = 1 - \frac{2(m_{ij.V} - 1)}{p(p - 1)} \quad (1 \leq i < j \leq p),$$

compare Dudoit et al. (2003, §2.4). Holm's adjusted p -values $\pi_{ij.V}^H$ are still conservative simultaneous p -values in the sense that if we select the graph $\hat{G}^H(\alpha)$ by replacing $\pi_{ij.V}$ by $\pi_{ij.V}^H$ in (3.14) then

$$(3.22) \quad \Pr_G[\hat{G}^H(\alpha) \subseteq G] \geq 1 - \alpha,$$

compare (3.17). Moreover, $0 < t_{ij.V} \leq 1$ for all $1 \leq i < j \leq p$, which implies that $\pi_{ij.V}^H \leq \pi_{ij.V}$. Therefore, it is always the case that

$$(3.23) \quad \hat{G}(\alpha) \subseteq \hat{G}^H(\alpha),$$

which means that by using the p -values $\pi_{ij.V}^H$ we choose a less conservative graph while still controlling the overall error rate for incorrect edge inclusion. By (3.23), if $\hat{G}(\alpha)$ is replaced by $\hat{G}^H(\alpha)$, the consistency result (3.18) remains valid. (The derivations in Drton and Perlman (2004) are readily amended.)

3.3. The SIN approach. For a given data set, n and p are fixed and the significance level α must be chosen in order to determine the selected model $\hat{G}^H(\alpha)$. In practice, a simple plot of the entire set of simultaneous p -values $\pi_{ij.V}^H$ often reveals a separation into two or three groups, designated by S , I , and N . Group S consists of small, hence clearly Significant, p -values corresponding to edges that definitely should be included. Group N consists of large, hence clearly Non-significant, p -values, corresponding to edges that definitely should be excluded. Group I consists of Indeterminate p -values, corresponding to edges that might

be included under a more liberal significance level. Identifying a set of indeterminate p -values is an explicit way of considering different significance levels; compare e.g. Druzdel and Glymour (1999, p.534) who say about their model selection program that it is “a good practice to run the program at several significance levels”.

This SIN model selection procedure usually leads to two selected graphs, a smaller model \hat{G}_S whose edges correspond to the p -values in S , and a larger model \hat{G}_{SI} whose edges correspond to the p -values in $S \cup I$. As a general rule one might determine the groups S , I , and N by the p -value ranges $(0, 0.05)$, $(0.05, 0.25)$, and $(0.25, 1)$, respectively, but this is both subjective and contextual. For example, for smaller sample sizes one might wish to increase the upper indeterminate value 0.25 to 0.4 or even to 0.5. As already noted, visual examination of the entire set of p -values often suggests an obvious separation into the three groups S , I , and N .

We now apply the SIN model selection method to the well-known mathematics marks data. The selected models \hat{G}_S and \hat{G}_{SI} are compared to the results of the backward stepwise selection method in Edwards’ (2000) software package MIM with the option of *unrestricted* selection, wherein both decomposable and non-decomposable models are considered. Additional examples are given in Drton and Perlman (2003, 2004), where the p -values $\pi_{ij.V}$ were used rather than Holm’s adjusted p -values $\pi_{ij.V}^H$ considered here.

3.4. Example: Mathematics marks. Mardia et al. (1979, pp. 3–4) present the marks of $n = 88$ students in the $p = 5$ examinations in

1. mechanics,
2. vectors,
3. algebra,
4. analysis, and
5. statistics.

The data are also considered in Edwards (2000, Example 3.1.6) and Whittaker (1990, §1.1). Backward stepwise selection in MIM at level $\alpha = 0.05$ yields the “butterfly” graph in Figure 4(a).

Our simultaneous p -values $\pi_{ij.V}^H$ from (3.12) and (3.20) are plotted in Figure 3. The plot suggests that we take $S = \{0.00, 0.01, 0.02\}$ corresponding to included edges 3 – 4, 3 – 5, and 1 – 2, and $N = \{0.93, 1.00^{(3)}\}$ corresponding to excluded edges 2 – 4, 1 – 4, 1 – 5, and 2 – 5. The indeterminate set of p -values is thus $I = \{0.06, 0.11, 0.16\}$ corresponding to possible edges 2 – 3, 4 – 5, and 1 – 3. Hence, SIN selects \hat{G}_{SI} and \hat{G}_S , where $\hat{G}_{SI} = \hat{G}^H(\alpha)$ for $\alpha \in (0.16, 0.93)$ and $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.02, 0.06)$. The graphs \hat{G}_{SI} and \hat{G}_S are shown in Figure 4(a) and (b), respectively.

Note that two p -values $\pi_{ij.V}^H$ and $\pi_{kl.V}^H$ may be equal (an effect strengthened due to rounding). Therefore, we adopt here and in the following examples the shorthand notation $\pi^{(m)}$ for a p -value π in S , I , or N which arises for m edges (at two decimal digits of accuracy).

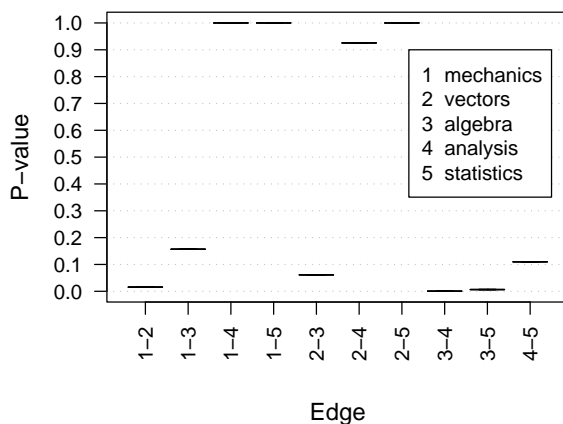


FIGURE 3. Simultaneous p -values $\pi_{ij,V}^H$ for UG model selection.

The SIN model \hat{G}_{SI} coincides with that found by MIM, which suggests that (mechanics, vectors) and (analysis, statistics) are conditionally independent given algebra, a readily interpretable property. The more conservative SIN model \hat{G}_S suggests instead that (mechanics, vectors) is unconditionally independent of (analysis, algebra, statistics) and that analysis and statistics are conditionally independent given algebra, which may also be cognitively interpretable. This illustrates the ease with which SIN reveals alternative models that may convey scientifically meaningful information.

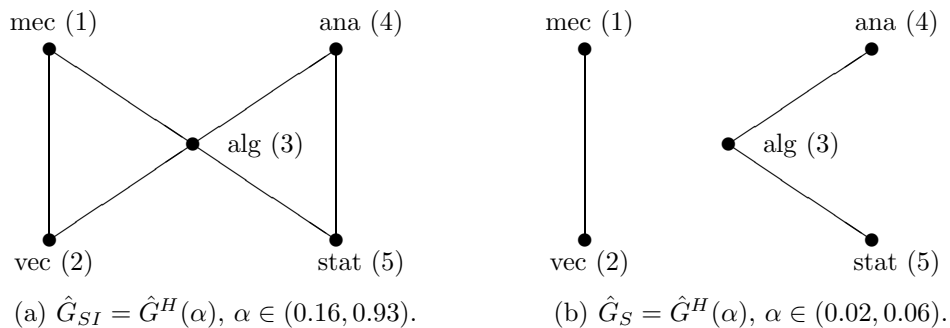


FIGURE 4. UGs for mathematics marks data.

4. SIN FOR GAUSSIAN BIDIRECTED GRAPH MODELS

4.1. Methodology for bidirected graphs. Let $Y^{(1)}, \dots, Y^{(n)}$ be a sample from the model $N_{BG}(G)$, where $G = (V, E_{\leftrightarrow})$ is an unknown BG. Inference for BG models has not been developed as thoroughly as for UG models. In particular, algorithms for maximum likelihood estimation are still under development (Drton and Richardson, 2003) and are not yet implemented in software packages. However, Kauermann (1996) developed a heuristic

inference method based on a “dual likelihood”, which can be carried out in the MIM package as described in Edwards (2000, §7.4). This method treats the inverse sample covariance matrix as if it were the sample covariance matrix and fits a Gaussian concentration graph model to the altered sufficient statistics. Thus, one can perform a backward stepwise model selection by this dualization. In general, the dual likelihood approach and maximum likelihood estimation will not give the same result. In particular, the actual likelihood of a covariance graph model can be multimodal (Drton and Richardson, 2004) whereas the dual likelihood is always unimodal.

A simple way of bypassing these complications is to adapt SIN model selection to BG models. The definition of $N_{\text{BG}}(G)$ in (2.8) suggests that we can recover G from the data by testing the $p(p-1)/2$ hypotheses

$$(4.1) \quad H_{ij} : \rho_{ij} = 0 \quad \text{vs.} \quad K_{ij} : \rho_{ij} \neq 0 \quad (1 \leq i < j \leq p).$$

To test H_{ij} , we use the sample correlations r_{ij} rather than the sample partial correlations $r_{ij.V}$. Denote the z -transforms of r_{ij} and ρ_{ij} by z_{ij} and ζ_{ij} , respectively. As smooth transformations of the sample covariance matrix W , the z -transforms $(z_{ij} \mid 1 \leq i < j \leq p)$ are jointly asymptotically normal with mean $(\zeta_{ij} \mid 1 \leq i < j \leq p)$. Marginally (cf. Anderson, 2003, §4.2.3), a good normal approximation is obtained from the asymptotic result

$$(4.2) \quad \sqrt{n-3} (z_{ij} - \zeta_{ij}) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty.$$

Thus, as in (3.12), conservative simultaneous p -values π_{ij} for (4.1) can be obtained:

$$(4.3) \quad \pi_{ij} := 1 - [2\Phi(\sqrt{n-3}|z_{ij}|) - 1]^{p(p-1)/2};$$

compare also Larntz and Perlman (1988, p.297). The only change from (3.12) is that the sample size adjustment is $n-3$ rather than $n_p = n-p-1$. Applying Holm’s adjustment as in (3.20) and (3.21), we obtain the adjusted p -values π_{ij}^H .

Using these adjusted p -values, our proposed BG model selection proceeds as in Section 3. It selects the graph $\hat{G}^H(\alpha) = (V, \hat{E}_{\leftrightarrow}(\alpha))$ with edges $\hat{E}_{\leftrightarrow}(\alpha) := (\hat{e}_{i \leftrightarrow j}(\alpha) \mid 1 \leq i \neq j \leq p)$ that are given by

$$(4.4) \quad \hat{e}_{i \leftrightarrow j}(\alpha) = \hat{e}_{j \leftrightarrow i}(\alpha) = \begin{cases} 0 & \text{if } \pi_{ij}^H \geq \alpha, \\ 1 & \text{if } \pi_{ij}^H < \alpha. \end{cases}$$

Thus, the edge $i \leftrightarrow j$ is included in $\hat{G}^H(\alpha)$ iff π_{ij}^H is significantly smaller than the overall level α .

It is easy to show that $\hat{G}^H(\alpha)$ again satisfies properties (3.17) and (3.18). For applications, we again advocate the SIN approach, whereby the simultaneous p -values π_{ij}^H are partitioned into groups S , I , and N , leading to two selected models \hat{G}_{SI} and \hat{G}_S , the latter being more conservative with respect to edge inclusion.

4.2. **Example: Glucose control.** Cox and Wermuth (1993, Example 7) report data from an investigation of the determinants of blood glucose control involving $n = 39$ diabetic patients. The data consists of $p = 4$ variables, which are

1. glycosylated hemoglobin (GHb),
2. knowledge about the illness,
3. duration of the illness, and
4. fatalism, a measure of the patient’s attitude to the illness.

By examining the sample covariance matrix, Cox and Wermuth (1993) conclude that the data agree well with the marginal independences $1 \perp\!\!\!\perp 4$, $2 \perp\!\!\!\perp 3$, and $3 \perp\!\!\!\perp 4$, which leads to the graph in Figure 6(a). Backward selection in MIM using Kauermann’s (1996) dual likelihood methodology yields the same graph when the individual significance level 0.05 is chosen.

The simultaneous p -values π_{ij}^H are depicted in Figure 5. Clearly $S = \{0.02, 0.05\}$, cor-

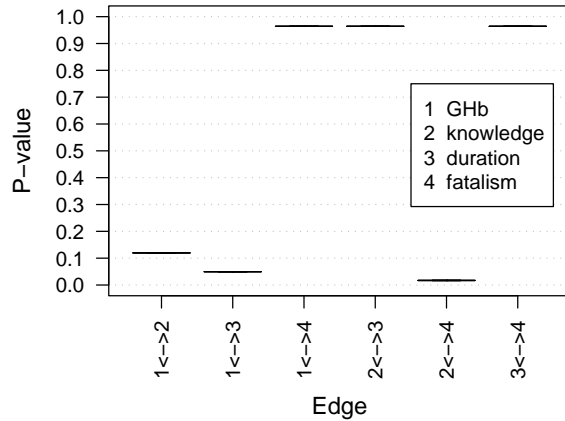


FIGURE 5. Simultaneous p -values π_{ij}^H for BG model selection.

responding to including the edges $2 \leftrightarrow 4$ and $1 \leftrightarrow 3$; $N = \{0.96^{(3)}\}$, corresponding to excluding edges $1 \leftrightarrow 4$, $2 \leftrightarrow 3$, and $3 \leftrightarrow 4$; and $I = \{0.12\}$, corresponding to the possible edge $1 \leftrightarrow 2$. Hence, SIN selects $\hat{G}_{SI} = \hat{G}^H(\alpha)$ for $\alpha \in (0.12, 0.96)$, as illustrated in Figure 6(a), and, more conservatively, $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.05, 0.12)$, as illustrated in Figure 6(b).

5. SIN FOR GAUSSIAN ACYCLIC DIRECTED GRAPH MODELS

5.1. **Methodology for acyclic directed graphs.** Let $Y^{(1)}, \dots, Y^{(n)}$ be a sample from the model $N_{\text{ADG}}(G)$, where $G = (V, E_{\rightarrow})$ is an ADG. The graph G is unknown but we assume to know an *a priori* total ordering $\mathbb{T} = (V, \preccurlyeq)$ of the variable set V that forms a well-ordering for G . Let $\mathcal{D}(\mathbb{T})$ be the set of ADGs, for which \mathbb{T} is a well-ordering. By

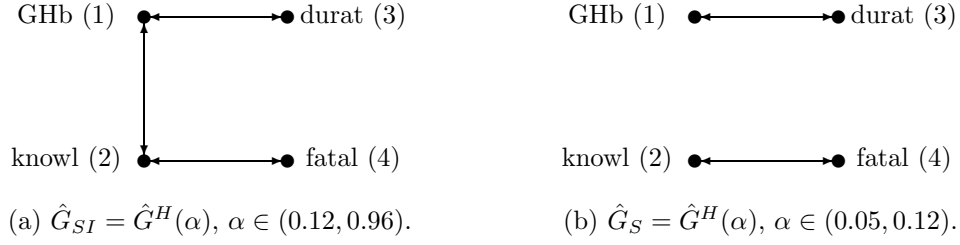


FIGURE 6. BGs for glucose control data.

assumption, the true graph G is in $\mathcal{D}(\mathbb{T})$. Model selection now consists of recovering the graph G from all *a priori* possible graphs $\mathcal{D}(\mathbb{T})$. In an application, the total ordering \mathbb{T} is typically based on *a priori* information about temporal or causal orderings of the variables.

Without loss of generality, we can assume that the variables are labelled as $V = \{1, \dots, p\}$ such that $i \prec j$ iff $i < j$, $i, j \in V$. This yields that if $e_{i \rightarrow j} = 1$ in G then $i < j$. We wish to recover G by simultaneously testing the $p(p-1)/2$ hypotheses

$$(5.1) \quad H_{ij.\{1,\dots,j\}} : \rho_{ij.\{1,\dots,j\}} = 0 \quad \text{vs.} \quad K_{ij.\{1,\dots,j\}} : \rho_{ij.\{1,\dots,j\}} \neq 0 \quad (1 \leq i < j \leq p)$$

that occur in the definition of $N_{\text{ADG}}(G)$; see (2.10) and (2.11).

The hypotheses $H_{ij.\{1,\dots,j\}}$, $i < j$, can be tested using the sample partial correlations $r_{ij.\{1,\dots,j\}}$. Denote the z -transforms of $r_{ij.\{1,\dots,j\}}$ and $\rho_{ij.\{1,\dots,j\}}$ by $z_{ij.\{1,\dots,j\}}$ and $\zeta_{ij.\{1,\dots,j\}}$, respectively. The z -transforms $(z_{ij.\{1,\dots,j\}} \mid 1 \leq i < j \leq p)$ are smooth transformations of the sample covariance matrix W and therefore jointly asymptotically normal with means $(\zeta_{ij.\{1,\dots,j\}} \mid 1 \leq i < j \leq p)$. The marginal distribution of $z_{ij.\{1,\dots,j\}}$ centered at $\zeta_{ij.\{1,\dots,j\}}$ is asymptotically standard normal (Anderson, 2003, §4.2.3 and Thm. 4.3.5). As in the previous sections we can obtain conservative simultaneous p -values $\pi_{ij.\{1,\dots,j\}}$ for (5.1) as follows

$$(5.2) \quad \pi_{ij.\{1,\dots,j\}} := 1 - [2\Phi(\sqrt{n_j} |z_{ij.\{1,\dots,j\}}|) - 1]^{p(p-1)/2},$$

where $n_j = (n-3) - (j-2) = n-j-1$, $j = 2, \dots, p$. Applying Holm's adjustment as in (3.20) and (3.21), we obtain the adjusted p -values $\pi_{ij.\{1,\dots,j\}}^H$.

Using these adjusted p -values, our proposed ADG model selection proceeds as in Sections 3 and 4. It selects the graph $\hat{G}^H(\alpha) = (V, \hat{E}_\rightarrow(\alpha))$ with edges $\hat{E}_\rightarrow(\alpha) := (\hat{e}_{i \rightarrow j}(\alpha) \mid 1 \leq i \neq j \leq p)$ given by

$$(5.3) \quad \hat{e}_{i \rightarrow j}(\alpha) = \begin{cases} 0 & \text{if } i > j, \\ 0 & \text{if } i < j \text{ and } \pi_{ij.\{1,\dots,j\}}^H \geq \alpha, \\ 1 & \text{if } i < j \text{ and } \pi_{ij.\{1,\dots,j\}}^H < \alpha. \end{cases}$$

It is straightforward to show that $\hat{G}^H(\alpha)$ satisfies properties (3.17) and (3.18), provided the true graph G satisfies $G \in \mathcal{D}(\mathbb{T})$.

5.2. **Example: Publishing productivity.** Revisiting a psychological study, Spirtes et al. (2000, Example 5.8.1) consider data on publishing productivity among academics. The $p = 7$ variables are

1. subject’s sex (sex),
2. score of the subject’s ability (ability),
3. measure of the quality of the graduate program attended (GPQ),
4. preliminary measure of productivity (preprod),
5. quality of the first job (QFJ),
6. publication rate (pubs), and
7. citation rate (cites).

The sample comprises $n = 162$ subjects. The ordering of the labels 1 to 7 used in the above listing of the variables corresponds to the common sense time order employed by Spirtes et al. (2000, p.101). In other words, if \prec denotes “being determined before”, then

$$\text{sex (1)} \prec \text{ability (2)} \prec \text{GPQ (3)} \prec \text{preprod (4)} \prec \text{QFJ (5)} \prec \text{pubs (6)} \prec \text{cites (7)}.$$

Let $\mathbb{T} = (V, \preceq)$ be the induced total ordering of $V = \{1, \dots, 7\}$, which agrees with the total ordering of the variable labels. In SIN model selection, we assume that the true graph G satisfies $G \in \mathcal{D}(\mathbb{T})$, that is, $e_{i \rightarrow j} = 1$ in the true graph G only if $i < j$.

The simultaneous p -values $\pi_{ij, \{1, \dots, j\}}^H$ shown in Figure 7 clearly suggest taking $S =$

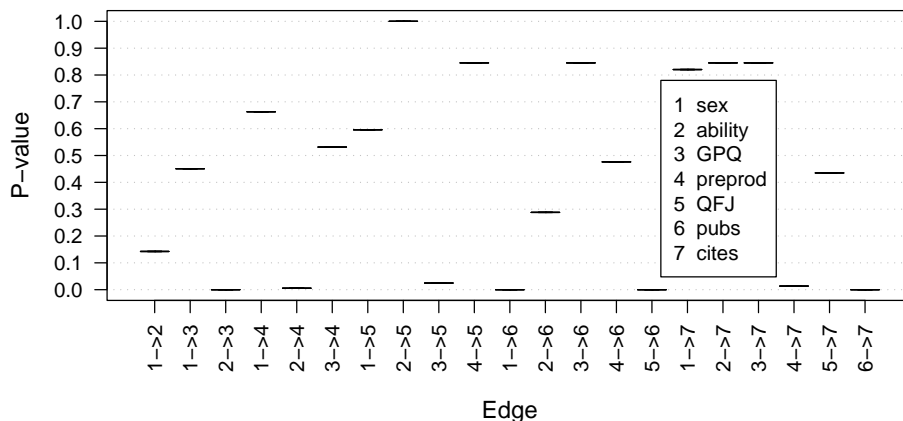


FIGURE 7. Simultaneous p -values $\pi_{ij, \{1, \dots, j\}}^H$ for ADG model selection.

$\{0.00^{(4)}, 0.01^{(2)}, 0.02\}$, corresponding to including the edges $2 \rightarrow 3, 1 \rightarrow 6, 5 \rightarrow 6, 6 \rightarrow 7, 2 \rightarrow 4, 4 \rightarrow 7$ and $3 \rightarrow 5$; and $N = \{0.29, 0.44, 0.45, 0.48, 0.53, 0.60, 0.66, 0.82, 0.85^{(4)}, 1.00\}$, corresponding to excluding the edges $2 \rightarrow 6, 5 \rightarrow 7, 1 \rightarrow 3, 4 \rightarrow 6, 3 \rightarrow 4, 1 \rightarrow 5, 1 \rightarrow 4, 1 \rightarrow 7, 4 \rightarrow 5, 3 \rightarrow 6, 2 \rightarrow 7, 3 \rightarrow 7$ and $2 \rightarrow 5$. There is one indeterminate p -value $I = \{0.14\}$ corresponding to the possible edge $1 \rightarrow 2$. Therefore, SIN selects $\hat{G}_{SI} = \hat{G}^H(\alpha)$

for $\alpha \in (0.14, 0.29)$, illustrated in Figure 8(a), and $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.02, 0.14)$, illustrated in Figure 8(b).

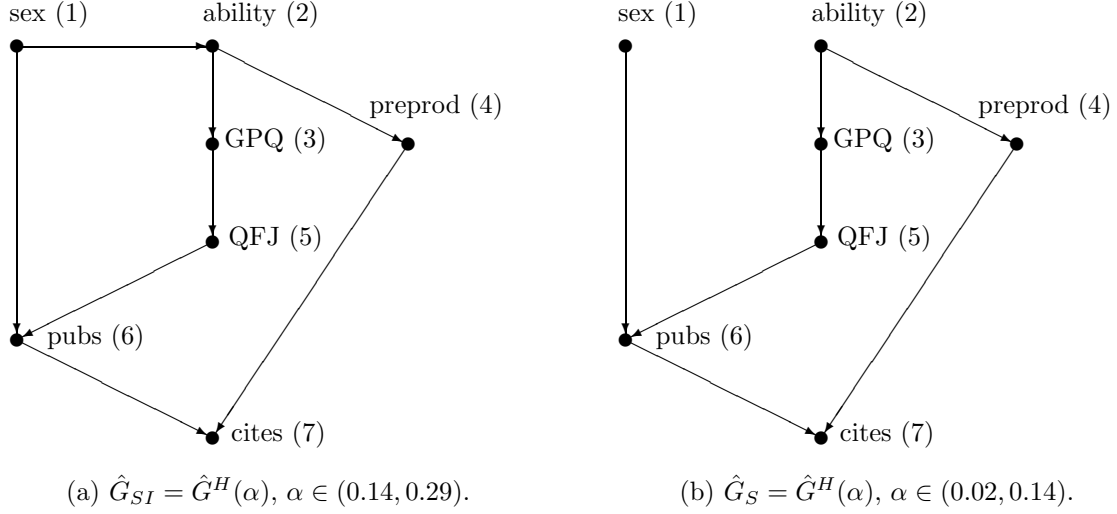


FIGURE 8. ADGs for publishing productivity data.

Backward selection in MIM, at individual level $\alpha = 0.05$, yields a super-graph of \hat{G}_S , which includes the two additional edges $2 \rightarrow 6$ and $5 \rightarrow 7$.

6. SIN FOR GAUSSIAN CHAIN GRAPH MODELS

Let $Y^{(1)}, \dots, Y^{(n)}$ be a sample from the Gaussian graphical model $N_{\text{LWF}}(G)$ or $N_{\text{AMP}}(G)$, where $G = (V, E_{\rightarrow}^-)$ is a CG. The graph G is unknown but we assume to know *a priori* a partitioning $\mathbb{D} = (B_k \mid 1 \leq k \leq q)$, $q \leq p$, of the variable set V that forms a dependence chain for G ; recall (2.13). Let $\mathcal{G}(\mathbb{D})$ be the set of CGs for which \mathbb{D} is a dependence chain. By assumption, the true graph G is in $\mathcal{G}(\mathbb{D})$. Model selection now consists of recovering G from all *a priori* possible graphs $\mathcal{G}(\mathbb{D})$.

For $i \in V$, let $1 \leq k(i) \leq q$ be such that $i \in B_{k(i)}$. Without loss of generality, we can assume that the variables are labelled as $V = \{1, \dots, p\}$ such that $i < j$ whenever $k(i) < k(j)$. Recall the definition of the cumulatives $C_\ell = \cup(B_k \mid k \leq \ell)$.

6.1. Methodology for LWF chain graphs. Assume that the distribution generating the sample is in $N_{\text{LWF}}(G)$ for some $G \in \mathcal{G}(\mathbb{D})$. Then we can recover G by simultaneously testing the $p(p-1)/2$ hypotheses (cf. (2.15), (2.16))

$$(6.1) \quad H_{ij.C_{k(j)}} : \rho_{ij.C_{k(j)}} = 0 \quad \text{vs.} \quad K_{ij.C_{k(j)}} : \rho_{ij.C_{k(j)}} \neq 0 \quad (1 \leq i < j \leq p).$$

In order to test the hypotheses (6.1), we again use the z -transforms $z_{ij.C_{k(j)}}$ of the sample partial correlations $r_{ij.C_{k(j)}}$, $1 \leq i < j \leq p$. As before, the z -transforms $(z_{ij.C_{k(j)}} \mid 1 \leq i <$

$j \leq p$) are jointly asymptotically normal, and marginally they are asymptotically standard normal if centered at $\zeta_{ij.C_{k(j)}} = z(\rho_{ij.C_{k(j)}})$ and suitably standardized.

By analogy to the previous sections, we are led to the conservative simultaneous p -values

$$(6.2) \quad \pi_{ij.C_{k(j)}} := 1 - [2\Phi(\sqrt{n_{c_{k(j)}}} |z_{ij.C_{k(j)}}|) - 1]^{p(p-1)/2},$$

where $c_k = |C_k|$ and $n_{c_{k(j)}} = (n-3) - (c_{k(j)} - 2) = n - c_{k(j)} - 1$; compare (5.2). Applying Holm's adjustment as in (3.20) and (3.21), we obtain the adjusted p -values $\pi_{ij.C_{k(j)}}^H$. These p -values are used to select the (LWF) CG $\hat{G}^H(\alpha) = (V, \hat{E}_-(\alpha))$ whose edges $\hat{E}_-(\alpha) = (\hat{E}_+(\alpha), \hat{E}_-(\alpha))$ are determined as follows. The directed edges $\hat{E}_-(\alpha)$ are given by

$$(6.3) \quad \hat{e}_{i \rightarrow j}(\alpha) = \begin{cases} 0 & \text{if } k(i) \geq k(j), \\ 0 & \text{if } k(i) < k(j) \text{ and } \pi_{ij.C_{k(j)}}^H \geq \alpha, \\ 1 & \text{if } k(i) < k(j) \text{ and } \pi_{ij.C_{k(j)}}^H < \alpha, \end{cases}$$

while the undirected edges $\hat{E}_-(\alpha)$ are given by

$$(6.4) \quad \hat{e}_{i-j}(\alpha) = \hat{e}_{j-i}(\alpha) = \begin{cases} 0 & \text{if } k(i) \neq k(j), \\ 0 & \text{if } k(i) = k(j) \text{ and } \pi_{ij.C_{k(j)}}^H \geq \alpha, \\ 1 & \text{if } k(i) = k(j) \text{ and } \pi_{ij.C_{k(j)}}^H < \alpha. \end{cases}$$

If the true graph G is in $\mathcal{G}(\mathbb{D})$, then $\hat{G}^H(\alpha)$ again satisfies properties (3.17) and (3.18).

6.2. Methodology for AMP chain graphs. Assume now that the data-generating distribution is in $N_{\text{AMP}}(G)$ for some $G \in \mathcal{G}(\mathbb{D})$. To simplify notation, for $1 \leq i < j \leq p$ let

$$(6.5) \quad \tilde{C}(i, j) = \begin{cases} C_{k(i)} = C_{k(j)} & : \text{ if } k(i) = k(j), \\ C_{k(j)-1} \cup \{j\} & : \text{ if } k(i) < k(j). \end{cases}$$

We can recover G by simultaneously testing the $p(p-1)/2$ hypotheses (recall (2.17)-(2.20))

$$(6.6) \quad H_{ij.\tilde{C}(i,j)} : \rho_{ij.\tilde{C}(i,j)} = 0 \quad \text{vs.} \quad K_{ij.\tilde{C}(i,j)} : \rho_{ij.\tilde{C}(i,j)} \neq 0 \quad (1 \leq i < j \leq p).$$

As before, in order to test the hypotheses (6.1), we use the z -transforms $z_{ij.\tilde{C}(i,j)}$ of the sample partial correlations $r_{ij.\tilde{C}(i,j)}$, $1 \leq i < j \leq p$.

From the asymptotics, we obtain the conservative simultaneous p -values

$$(6.7) \quad \pi_{ij.\tilde{C}(i,j)} := 1 - [2\Phi(\sqrt{n_{\tilde{c}(i,j)}} |z_{ij.\tilde{C}(i,j)}}|) - 1]^{p(p-1)/2},$$

where $n_{\tilde{c}(i,j)} = (n-3) - (\tilde{c}(i,j) - 2) = n - \tilde{c}(i,j) - 1$ and $\tilde{c}(i,j) = |\tilde{C}(i,j)|$ (cf. (5.2) and (6.2)). Holm's adjustment (cf. 3.20) and (3.21)) yields the adjusted p -values $\pi_{ij.\tilde{C}(i,j)}^H$. With these p -values, we select the (AMP) CG $\hat{G}^H(\alpha) = (V, \hat{E}_-(\alpha))$ whose edges $\hat{E}_-(\alpha) =$

$(\hat{E}_{\rightarrow}(\alpha), \hat{E}_{-}(\alpha))$ are determined as follows. The directed edges $\hat{E}_{\rightarrow}(\alpha)$ are given by

$$(6.8) \quad \hat{e}_{i \rightarrow j}(\alpha) = \begin{cases} 0 & \text{if } k(i) \geq k(j), \\ 0 & \text{if } k(i) < k(j) \text{ and } \pi_{ij, \tilde{C}(i,j)}^H \geq \alpha, \\ 1 & \text{if } k(i) < k(j) \text{ and } \pi_{ij, \tilde{C}(i,j)}^H < \alpha, \end{cases}$$

while the undirected edges $\hat{E}_{-}(\alpha)$ are given by

$$(6.9) \quad \hat{e}_{i-j}(\alpha) = \hat{e}_{j-i}(\alpha) = \begin{cases} 0 & \text{if } k(i) \neq k(j), \\ 0 & \text{if } k(i) = k(j) \text{ and } \pi_{ij, \tilde{C}(i,j)}^H \geq \alpha, \\ 1 & \text{if } k(i) = k(j) \text{ and } \pi_{ij, \tilde{C}(i,j)}^H < \alpha. \end{cases}$$

Again, $\hat{G}^H(\alpha)$ satisfies properties (3.17) and (3.18) if the true graph G is in $\mathcal{G}(\mathbb{D})$.

SIN model selection is particularly attractive for AMP CGs for the simple practical reason that statistical inference methodology for AMP CGs is still under development (Drton, 2004) and not yet available in software packages. In addition, the likelihood function for AMP CG models may be multimodal as shown in Drton and Richardson (2004), who analyze the model arising from the conditioning $(Y_3, Y_4 \mid Y_1, Y_2)$ in the AMP CG model based on the graph in Figure 1(d).

6.3. Example: University graduation rates. Druzdzel and Glymour (1999) analyze data from a study carried out for the purpose of college ranking. After preprocessing the data available for the year 1993, they focus on $p = 8$ variables and $n = 159$ universities (Druzdzel and Glymour, 1999, Table 3). The eight variables are

- a1. average spending per student (spend),
- a2. student-teacher ratio (strat),
- a3. faculty salary (salar),
- b1. rejection rate (rejr),
- b2. percentage of admitted students who accept university's offer (pacc),
- c1. average test scores of incoming students (tstsc),
- c2. class standing of incoming freshmen (top10), and
- d1. average percentage of graduation (apgra).

Let $A = \{a1, a2, a3\}$, $B = \{b1, b2\}$, $C = \{c1, c2\}$, and $D = \{d1\}$, then $\mathbb{D} = (A, B, C, D)$ is a partition of the variable set. As argued in Druzdzel and Glymour (1999, p.534), the temporal order $A < B < C < D$ seems appropriate. Thus, we restrict ourselves to the set $\mathcal{G}(\mathbb{D})$ of all CGs for which $\mathbb{D} = (B_1, B_2, B_3, B_4) = (A, B, C, D)$ is a dependence chain. In the following, we will illustrate SIN model selection for AMP and LWF CGs $G \in \mathcal{G}(\mathbb{D})$.

6.3.1. Selecting an LWF chain graph by SIN. The simultaneous p -values $\pi_{ij, C_{k(j)}}^H$ are depicted in Figure 9. We choose $S = \{0.00^{(9)}, 0.01, 0.02, 0.04^{(2)}\}$, corresponding to including the edges $a1 - a2$, $a1 - a3$, $a2 - a3$, $a3 \rightarrow b1$, $a3 \rightarrow b2$, $a1 \rightarrow c1$, $a2 \rightarrow c1$, $c1 - c2$, $c2 \rightarrow d1$, $b1 - b2$, $a1 \rightarrow b1$, $a2 \rightarrow c2$ and $b2 \rightarrow d1$. We consider as clearly non-significant the

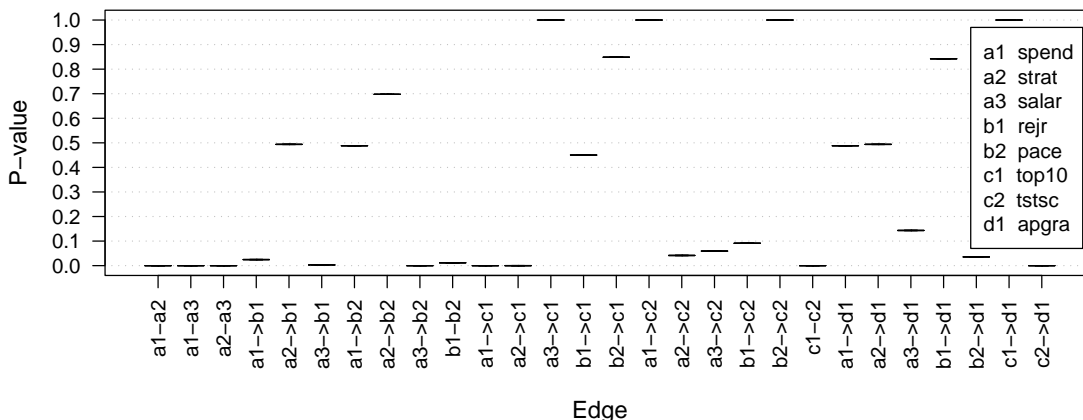


FIGURE 9. Simultaneous p -values $\pi_{ij.C_k(j)}^H$ for LWF CG model selection.

p -values in $N = \{0.45, 0.49^{(4)}, 0.70, 0.84, 0.85, 1.00^{(4)}\}$, which correspond to the excluded edges $b1 \rightarrow c1$, $a2 \rightarrow b1$, $a1 \rightarrow b2$, $a1 \rightarrow d1$, $a2 \rightarrow d1$, $a2 \rightarrow b2$, $b2 \rightarrow c1$, $b1 \rightarrow d1$, $a3 \rightarrow c1$, $a1 \rightarrow c2$, $b2 \rightarrow c2$, and $c1 \rightarrow d1$. Finally, we consider as indeterminate the p -values $I = \{0.06, 0.09, 0.14\}$ with associated possible edges $a3 \rightarrow c2$, $b1 \rightarrow c2$, and $a3 \rightarrow d1$. Thus, SIN selects $\hat{G}_{SI} = \hat{G}^H(\alpha)$ for $\alpha \in (0.14, 0.45)$, illustrated in Figure 10(a), and $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.04, 0.06)$, illustrated in Figure 10(b).

For comparison, backward selection in MIM, at individual level 0.05, yields a CG with skeleton similar to but less sparse than \hat{G}_{SI} . The graph selected by MIM has four additional edges, which are $a1 \rightarrow c2$, $a3 \rightarrow c1$, $b1 \rightarrow c1$, and $c1 \rightarrow d1$, but the edge $c2 \rightarrow d1$ that is present in \hat{G}_{SI} is absent in the MIM graph. Backward selection includes the edges $a1 \rightarrow c2$, $a3 \rightarrow c1$, and $c1 \rightarrow d1$, even though their associated simultaneous p -values equal 1.00, and omits the edge $c2 \rightarrow d1$ with simultaneous p -value 0.00.

6.3.2. *Selecting an AMP chain graph by SIN.* Figure 11 illustrates the simultaneous p -values $\pi_{ij.\tilde{C}(i,j)}^H$. We choose $S = \{0.00^{(11)}, 0.01, 0.03^{(3)}\}$, corresponding to including the edges $a1 - a2$, $a1 - a3$, $a2 - a3$, $a3 \rightarrow b2$, $b1 - b2$, $a1 \rightarrow c1$, $a2 \rightarrow c1$, $b1 \rightarrow c1$, $b1 \rightarrow c2$, $c1 - c2$, $c2 \rightarrow d1$, $a3 \rightarrow c2$, $a1 \rightarrow b1$, $a3 \rightarrow b1$ and $b2 \rightarrow d1$. We consider as clearly non-significant the p -values in $N = \{0.49^{(4)}, 0.61, 0.66, 0.79^{(3)}, 1.00^{(2)}\}$, which correspond to the excluded edges $a2 \rightarrow b1$, $a3 \rightarrow c1$, $a1 \rightarrow d1$, $a2 \rightarrow d1$, $a1 \rightarrow b2$, $a2 \rightarrow b2$, $b2 \rightarrow c1$, $a2 \rightarrow c2$, $b1 \rightarrow d1$, $b2 \rightarrow c2$ and $c1 \rightarrow d1$. Finally, we consider as indeterminate the p -values $I = \{0.09, 0.13\}$ with associated possible edges $a1 \rightarrow c2$ and $a3 \rightarrow d1$. Thus, SIN selects $\hat{G}_{SI} = \hat{G}^H(\alpha)$ for $\alpha \in (0.13, 0.49)$, illustrated in Figure 12(a), and $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.03, 0.09)$, illustrated in Figure 12(b).

6.3.3. *Differences in the selected LWF and AMP chain graphs.* Even though $N_{AMP}(G) \neq N_{LWF}(G)$ in general, the LWF and AMP CGs selected in this example are very similar.

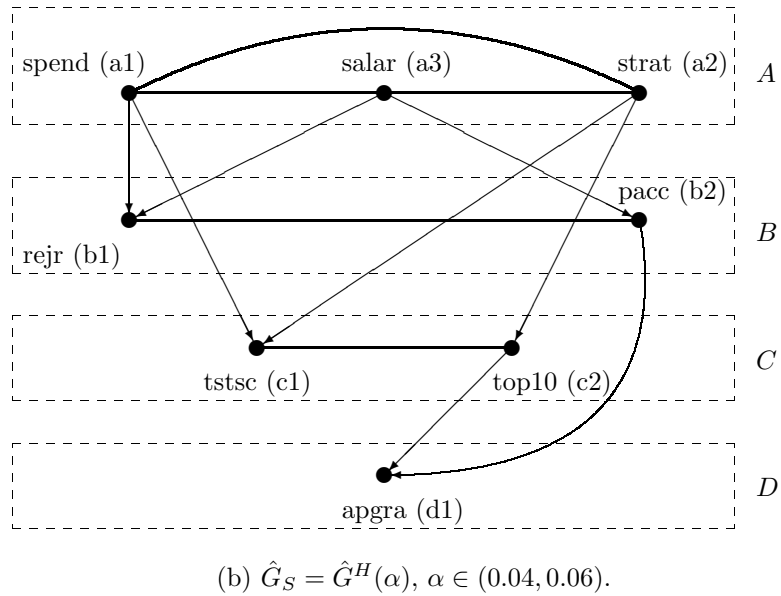
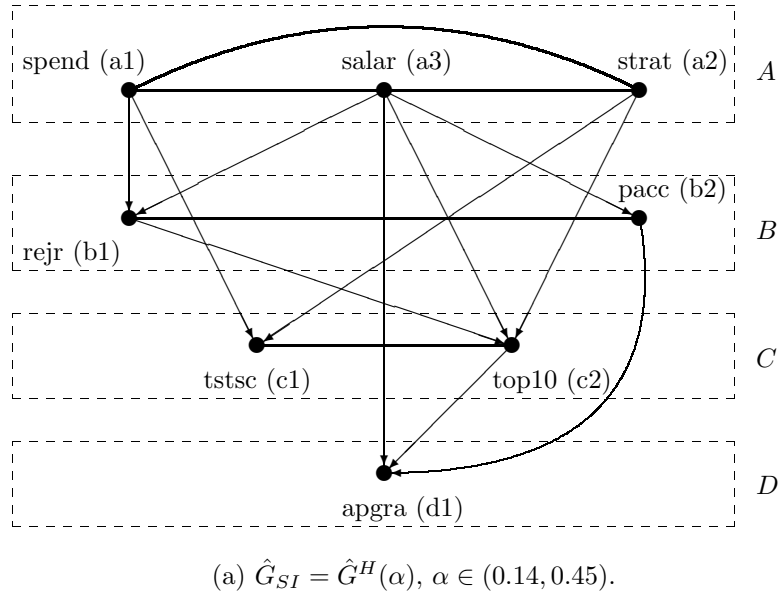


FIGURE 10. LWF CGs for university graduation rates data.

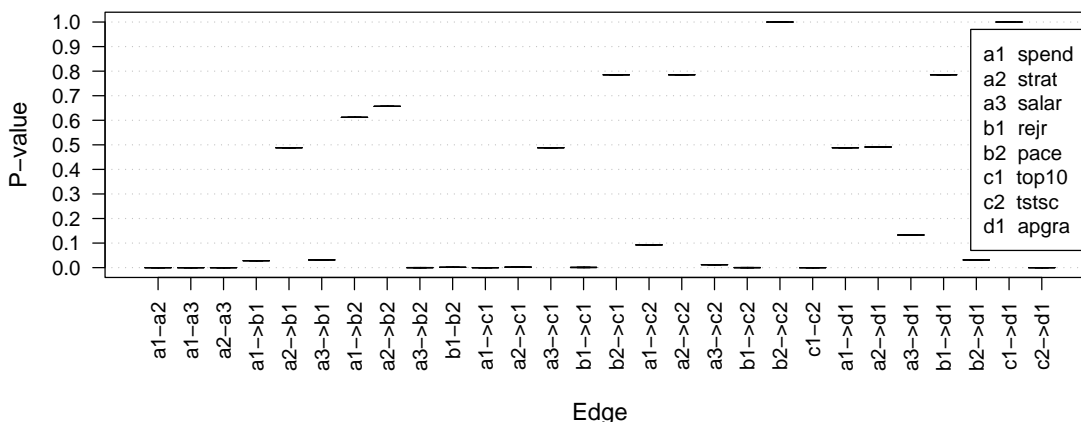


FIGURE 11. Simultaneous p -values $\pi_{ij, \tilde{C}(i,j)}^H$ for AMP CG model selection.

First note that if $i, j \in A = \{a1, a2, a3\}$, the first block in the dependence chain, then the hypotheses tested to determine the absence or presence of the undirected edge $i - j$ are the same for both the LWF and the AMP selection procedure. Thus, differences in the undirected subgraph over A can only arise due to Holm's p -value adjustment. This is not true for the other blocks. Here, however, the undirected subgraphs over A , as well as the undirected subgraphs over B and C are the same for all the LWF and AMP CGs selected by SIN (cf. Figures 10 and 12). Moreover, the hypotheses tested to determine the absence or presence of the directed edge $i \rightarrow d1$ are the same for both the LWF and the AMP procedure because the block $D = \{d1\}$ is a singleton, but again differences in the selected graph may occur due to Holm's p -value adjustment. Here, the directed edges in LWF and AMP CGs \hat{G}_{SI} and \hat{G}_S have the same directed edges $i \rightarrow d1$, respectively. For the remaining edges the hypotheses tested are different depending upon whether an LWF or an AMP CG is selected.

The LWF and AMP CGs \hat{G}_{SI} differ by three edges: the edge $a2 \rightarrow c2$ only occurs in the LWF CG, whereas the edges $a1 \rightarrow c2$ and $b1 \rightarrow c1$ only occur in the AMP CG. The more conservative LWF and AMP CGs \hat{G}_S differ by four edges: the edge $a2 \rightarrow c2$ only occurs in the LWF graph, whereas the edges $b1 \rightarrow c1$, $b1 \rightarrow c2$, and $a3 \rightarrow c2$ only occur in the AMP CG.

7. INCORPORATING PRIOR INFORMATION ABOUT THE PRESENCE OR ABSENCE OF EDGES

Suppose it is known that certain edges $E^{d,0}$ of the true graph $G \equiv (V, E)$ are definitely absent and that certain other edges $E^{d,1}$ are definitely present. Thus, $E = E^{d,0} \dot{\cup} E^{d,1} \dot{\cup} E^u$, where E^u denotes the remaining set of uncertain edges. Model selection now reduces to the problem of determining the absence or presence of the edges in E^u . Let $G^{up} := (V, E^{d,0} \dot{\cup} E^{d,1} \dot{\cup} E^{u,1})$ denote the *upper graph*, where $E^{u,1}$ replaces all uncertain edges by

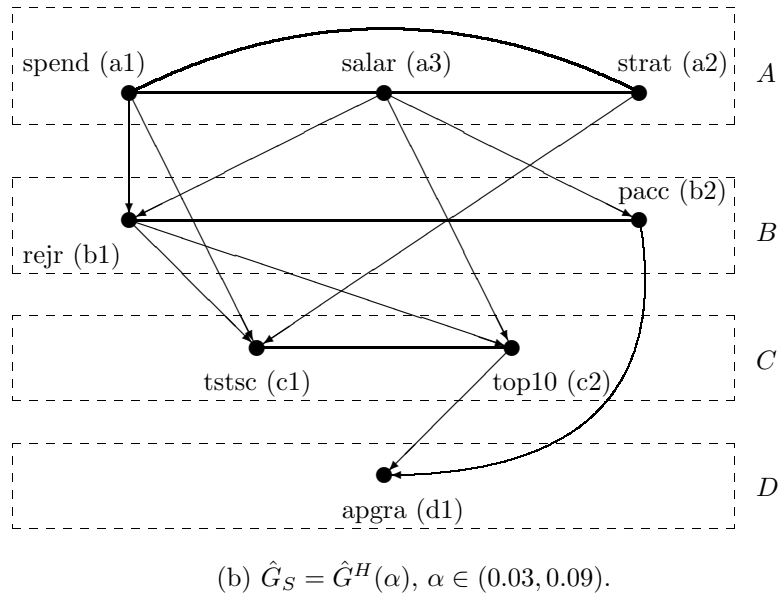
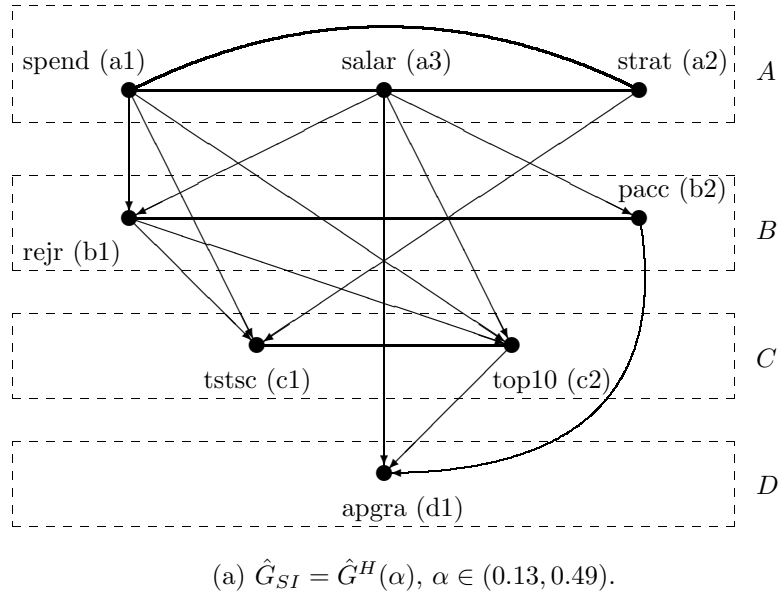


FIGURE 12. AMP CGs for university graduation rates data.

present edges, and let $G^{\text{low}} := (V, E^{d,0} \cup E^{d,1} \cup E^{u,0})$ denote the *lower graph*, where $E^{u,0}$ replaces all uncertain edges by absent edges. Thus the true graph G satisfies $G^{\text{low}} \subseteq G \subseteq G^{\text{up}}$, where G^{low} and G^{up} are known. If all edges are uncertain then the upper and lower graph are the complete and the empty graph, respectively. Consomni and Leucari (2001) call the upper graph the “full graph”, although solely in the context of directed graphs.

The methodology and SIN approach presented in Sections 3-6 extend readily to the present case by reducing the $p(p-1)/2$ simultaneous testing problems to the $q := |E^u|$ testing problems corresponding to the uncertain edges only. Since $q \leq p(p-1)/2$, we have fewer testing problems to consider and gain power in simultaneous testing. Furthermore, since $G \subseteq G^{\text{up}}$, the conditional independences holding in G^{up} also hold in G , which may allow for additional power gain in SIN model selection. In this section, we detail this approach for UGs, then comment briefly on its application to the other graph types considered in Sections 4-6.

For the case of an UG $G = (V, E_-)$, the definitely absent edges $E^{d,0}$, the definitely present edges $E^{d,1}$, and the uncertain edges E^u are all undirected, so we denote them by $E_-^{d,0}$, $E_-^{d,1}$ and E_-^u , respectively. Let e_{i-j}^{up} be the edge indicators for the upper UG G^{up} and define

$$\text{nb}^{\text{up}}\{i, j\} := \{k \in V \setminus \{i, j\} \mid e_{i-k}^{\text{up}} = 1 \vee e_{j-k}^{\text{up}} = 1\},$$

the neighbors of $\{i, j\}$ in the specified graph G^{up} . Furthermore, let

$$(7.1) \quad \text{Nb}^{\text{up}}\{i, j\} = \text{nb}^{\text{up}}\{i, j\} \cup \{i, j\}.$$

The global Markov property for UGs (Lauritzen, 1996, p.32) implies that for a data-generating distribution in the model $N_{\text{UG}}(G) \subseteq N_{\text{UG}}(G^{\text{up}})$,

$$(7.2) \quad \rho_{ij.V} = \rho_{ij.\text{Nb}^{\text{up}}\{i,j\}}$$

for all uncertain edges $ij \in E_-^u$. Therefore, the hypothesis $H_{ij.V}$ in (3.3) can be tested by means of the corresponding sample partial correlation $r_{ij.\text{Nb}^{\text{up}}\{i,j\}}$, or equivalently by its sample z -transform $z_{ij.\text{Nb}^{\text{up}}\{i,j\}}$. The approximation (3.6) is now replaced by

$$(7.3) \quad \sqrt{n_{ij}^{\text{up}}} (z_{ij.\text{Nb}^{\text{up}}\{i,j\}} - \zeta_{ij.\text{Nb}^{\text{up}}\{i,j\}}) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $n_{ij}^{\text{up}} = (n-3) - |\text{nb}^{\text{up}}\{i, j\}|$ and $\zeta_{ij.\text{Nb}^{\text{up}}\{i,j\}}$ is the z -transform of $\rho_{ij.\text{Nb}^{\text{up}}\{i,j\}}$. Šidák's inequality now yields that

$$(7.4) \quad \Pr_G[|z_{ij.\text{Nb}^{\text{up}}\{i,j\}} - \zeta_{ij.\text{Nb}^{\text{up}}\{i,j\}}| \leq (n_{ij}^{\text{up}})^{-1/2} \bar{b}_q(\alpha) \forall ij \in E_-^u] \geq 1 - \alpha$$

approximately for large n , where $\bar{b}_q(\alpha)$ is determined by

$$(7.5) \quad [2\Phi(\bar{b}_q(\alpha)) - 1]^q = 1 - \alpha,$$

or equivalently by

$$(7.6) \quad \bar{b}_q(\alpha) = \Phi^{-1} \left(\frac{1}{2} [(1 - \alpha)^{1/q} + 1] \right).$$

The corresponding set of conservative simultaneous p -values for the q testing problems $(H_{ij.V} \mid ij \in E_-^u)$ is given by

$$(7.7) \quad \bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}} := \bar{\pi}(z_{ij.\text{Nb}^{\text{up}}\{i,j\}}) = 1 - [2\Phi(\sqrt{n_{ij}^{\text{up}}} |z_{ij.\text{Nb}^{\text{up}}\{i,j\}}|) - 1]^q, \quad ij \in E_-^u.$$

Because $n_{ij}^{\text{up}} \geq n_p$ and $\bar{b}_q(\alpha) \leq b_p(\alpha)$, the corresponding tests are more powerful than the ones considered in Section 3. As in Section 3.2, we can improve the p -values $\bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}}$ by Holm's procedure to obtain the adjusted p -values $\bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}}^H$.

The selected graph $\hat{G}^H(\alpha) := (V, E_-^{d,0} \cup E_-^{d,1} \cup \hat{E}_-^u(\alpha))$, where $\hat{E}_-^u(\alpha) := \{\hat{e}_{i-j}(\alpha) \mid ij \in E_-^u\}$, is now given by

$$(7.8) \quad \hat{e}_{i-j}(\alpha) = \begin{cases} 0 & \text{if } \bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}}^H \geq \alpha, \\ 1 & \text{if } \bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}}^H < \alpha. \end{cases}$$

The results (3.17) and (3.18) concerning the overall error rate of the selection procedure $\hat{G}^H(\alpha)$ remain valid here. The SIN approach to model selection is applied here by simply replacing the entire set of p -values $(\pi_{ij.V}^H \mid 1 \leq i < j \leq p)$ by the reduced set $(\bar{\pi}_{ij.\text{Nb}^{\text{up}}\{i,j\}}^H \mid ij \in E_-^u)$, obtaining two selected UGs \hat{G}_S and \hat{G}_{SI} as in Section 3.4.

Note that in this approach of testing only uncertain edges, we use the Markov property of the upper UG but we still employ sample correlations as opposed to maximum likelihood estimates of correlations, for example. The reason for using sample partial correlations, which may not be asymptotically efficient, is the fact that the z -transform is variance-stabilizing for sample correlations but not necessarily for other estimates of correlations (Roverato, 1996).

7.1. Example: Prior knowledge for mathematics marks data. We revisit the mathematics marks data set from Section 3.4 and assume that we know how a student's performance in statistics relates to the performance in the other mathematical subjects. More precisely, we assume that in the true underlying UG G , the edges "statistics–algebra" and "statistics–analysis" are present, whereas the edges "statistics–mechanics" and "statistics–vectors" are absent. Labelling the variables by 1 to 5 as in Section 3.4, this means that $e_{1-5} = e_{2-5} = 0$ form the edge set $E^{d,0}$ and $e_{3-5} = e_{4-5} = 1$ form the edge set $E^{d,1}$. There remain $q = 6$ uncertain edges in the set E^u , which comprises all the edges amongst the variables 1 to 4, i.e. amongst all the variables except "statistics".

The upper UG G^{up} for this prior knowledge is shown in Figure 13. Let $1 \leq i < j \leq 4$ and recall $V = \{1, \dots, 5\}$. From Figure 13 we see immediately that $\text{nb}^{\text{up}}\{i, j\} = V \setminus \{i, j\}$ if $(i, j) \neq (1, 2)$. If $(i, j) = (1, 2)$, then $\text{nb}^{\text{up}}\{1, 2\} = \{3, 4\}$. Thus, for testing the absence of uncertain edges we will use the z -transforms

$$(7.9) \quad z_{ij.\text{Nb}^{\text{up}}\{i,j\}} = \begin{cases} z_{ij.V} & \text{if } (i, j) \neq (1, 2), \\ z_{12.34} & \text{if } (i, j) = (1, 2). \end{cases}$$

The resulting conservative simultaneous p -values are illustrated in Figure 14. We par-

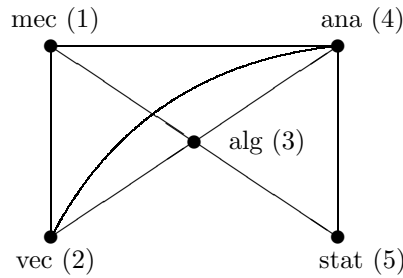


FIGURE 13. The upper UG G^{up} for the mathematics marks data assuming prior knowledge about edges involving “statistics”.

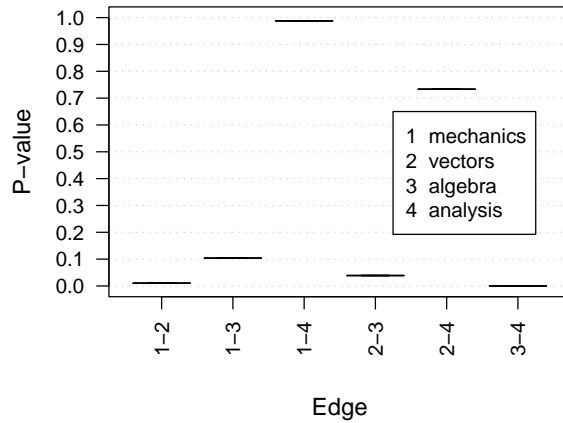


FIGURE 14. Simultaneous p -values $\bar{\pi}_{ij, \text{Nb}^{\text{up}}\{i,j\}}^H$ for UG model selection with prior information about edges.

tition these p -values into the set $S = \{0.00, 0.01, 0.04\}$ corresponding to included edges $3 - 4$, $1 - 2$ and $2 - 3$, the set $I = \{0.10\}$ corresponding to possible edge $1 - 3$, and the set $N = \{0.73, 0.99\}$ corresponding to excluded edges $2 - 4$ and $1 - 4$. Comparing these p -values to the ones given in Section 3.4, one can see that, as expected, the prior knowledge leads to smaller p -values for the uncertain edges. Hence, SIN with prior knowledge selects $\hat{G}_{SI} = \hat{G}^H(\alpha)$ for $\alpha \in (0.10, 0.73)$ and $\hat{G}_S = \hat{G}^H(\alpha)$ for $\alpha \in (0.04, 0.10)$. The graph \hat{G}_{SI} is identical to the “butterfly” graph shown in Figure 4(a) and the graph \hat{G}_S equals the graph in Figure 4(b) but with the additional edge $2 - 3$, which connects “vectors” and “algebra”.

7.2. Incorporating prior information about the presence or absence of edges in other types of graphs. In SIN selection for graphical models other than undirected ones, *a priori* knowledge about the presence and absence of edges again can be exploited by testing only hypotheses associated with uncertain edges. Furthermore, for all but the BGs

the sparsity of the upper graph may further improve the selection procedure by reducing the sizes of the conditioning sets, as in (7.2). This explained further as follows.

In the BG selection, the sample correlations, or rather their z -transforms z_{ij} , are used to test for the absence of an edge. Here, when testing only uncertain edges the z_{ij} remain unchanged. For ADGs, however, the following improvement is possible. Let $i < j$ be two vertices, $G_{\{1, \dots, j\}}^{\text{up}}$ the subgraph of the upper ADG G^{up} that is induced by the vertex set $\{1, \dots, j\}$, and $(G_{\{1, \dots, j\}}^{\text{up}})^m$ the moralized, thus undirected, subgraph. Then by the global Markov property for ADGs (Lauritzen, 1996, p.47), the partial correlation in (5.2) equals

$$(7.10) \quad z_{ij.\{1, \dots, j\}} = z_{ij.\text{Nb}^{\text{up}, m}\{i, j\}},$$

where $\text{nb}^{\text{up}, m}\{i, j\}$ is the set of neighbors of i and j in $(G_{\{1, \dots, j\}}^{\text{up}})^m$ and $\text{Nb}^{\text{up}, m}\{i, j\} = \text{nb}^{\text{up}, m}\{i, j\} \cup \{i, j\}$. In the LWF CG case, let $(G_{C_{k(j)}}^{\text{up}})^m$ be the moralized induced subgraph of the upper CG G^{up} . Then by the global Markov property for LWF CGs (Lauritzen, 1996, p.55), the partial correlation in (6.2) equals

$$(7.11) \quad z_{ij.C_{k(j)}} = z_{ij.\text{Nb}^{\text{up}, m}\{i, j\}},$$

where $\text{nb}^{\text{up}, m}\{i, j\}$ are the neighbors of i and j in $(G_{C_{k(j)}}^{\text{up}})^m$. Finally, in the AMP CG case, we consider two cases. If $k(i) = k(j)$, then $z_{ij.\tilde{C}(i, j)} = z_{ij.C_{k(j)}}$ in (6.7). Let $(G_{C_{k(j)}}^{\text{up}})^a$ be the augmented induced subgraph of the upper CG G^{up} (Andersson et al., 2001, §2). Then by the global Markov property for AMP chain graphs (Andersson et al., 2001, Def. 6),

$$(7.12) \quad z_{ij.C_{k(j)}} = z_{ij.\text{Nb}^{\text{up}, a}\{i, j\}},$$

where $\text{nb}^{\text{up}, a}\{i, j\}$ are the neighbors of i and j in the UG $(G_{C_{k(j)}}^{\text{up}})^a$ and $\text{Nb}^{\text{up}, a}\{i, j\} = \text{nb}^{\text{up}, a}\{i, j\} \cup \{i, j\}$. If $k(i) < k(j)$, then

$$(7.13) \quad z_{ij.\tilde{C}(i, j)} = z_{ij.C_{k(j)-1}} = z_{ij.\text{Nb}^{\text{up}, a}\{i, j\}}$$

but $\text{nb}^{\text{up}, a}\{i, j\}$ become the neighbors of i and j in $(G^{\text{up}}[C_{k(j)-1} \cup \{j\}])^a$, where the operation $G^{\text{up}} \mapsto G^{\text{up}}[A]$, $A \subseteq V$, is defined as in Andersson et al. (2001, §2).

8. CONCLUDING REMARKS

We have described SIN model selection, a new method for model selection in Gaussian graphical models based on simultaneous testing of the conditional independences that define the models, then partitioning the resulting p -values into a ‘‘Significant’’, an ‘‘Indeterminate’’ and a ‘‘Nonsignificant’’ set. The use of the variance-stabilizing z -transform and Šidák’s inequality in the simultaneous tests was first proposed by Drton and Perlman (2004) in the context of UGs. Drton and Perlman (2004) also present simulation results to justify the SIN approach. In this paper we improved the simultaneous tests by Holm’s p -value adjustment and extended the methodology to other types of graphs. The methodology easily extends to incorporate prior information about edges. The R package (Ihaka and Gentleman, 1996) ‘SIN’, which can be downloaded from the R server, implements SIN model selection.

In the cases of Gaussian BGs and AMP CGs, our method is attractive from a practical point of view since likelihood inference for these models is not yet fully developed. In addition, our method avoids problems with the possibly multimodal likelihood of BG and AMP CG models. Here, stepwise selection procedures might have to fit inaccurate models in some intermediate steps, which in turn might cause problems with a multimodal likelihood (Drton and Richardson, 2004).

Our method based on conservative simultaneous tests tends to select sparser (more parsimonious) models than standard procedures such as backward stepwise selection. It would be interesting to compare the simultaneous SIN procedure to other model selection procedures in a systematic simulation study. Furthermore, simultaneous testing methodology other than Holm's p -value adjustment could lead to further improvement of SIN model selection while still controlling the overall rate of incorrect edge inclusion. A recent review of simultaneous testing methodology in the framework of statistical genetics can be found in Dudoit et al. (2003).

ACKNOWLEDGEMENTS

We warmly thank Steen Andersson, Sanjay Chaudhuri, Thomas Richardson, Galen Shorack, Jon Wellner, and Graham Wood for helpful comments.

REFERENCES

- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ.
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Statist.* **24** 81–102.
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28** 33–85.
- ANDERSSON, S. A. and PERLMAN, M. D. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.* **66** 133–187.
- BANERJEE, M. and RICHARDSON, T. S. (2003). On a dualization of graphical Gaussian models: A correction note. *Scand. J. Statist.* **30** 817–820.
- CAPUTO, A., FORAITA, R., KLASSEN, S. and PIGEOT, I. (2003). Undernutrition in Benin - An analysis based on graphical models. *Social Science & Medicine* **56** 1677–1691.
- CAPUTO, A., HEINICKE, A. and PIGEOT, I. (1999). A graphical chain model derived from a model selection strategy for the sociologists graduates study. *Biom. J.* **41** 217–234.
- CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498.
- CONSONNI, G. and LEUCARI, V. (2001). Model determination for directed acyclic graphs. *The Statistician* **50** 243–256.

- COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science, Springer-Verlag, New York.
- COX, D. R. and WERMUTH, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8** 204–218.
- COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.
- DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.
- DIDELEZ, V., PIGEOT, I., DEAN, K. and WISTER, A. (2002). A comparative analysis of graphical interaction and logistic regression modelling: self-care and coping with a chronic illness in later life. *Biom. J.* **44** 410–432.
- DRTON, M. (2004). *Maximum Likelihood Estimation in Gaussian AMP Chain Graph Models and Gaussian Ancestral Graph Models*. Ph.D. thesis, Dept. of Statistics, University of Washington, Seattle.
- DRTON, M. and PERLMAN, M. D. (2003). A SInful approach to model selection for Gaussian concentration graphs. Tech. Rep. 429, Dept. of Statistics, University of Washington, Seattle.
- DRTON, M. and PERLMAN, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **0** 00–00. To appear.
- DRTON, M. and RICHARDSON, T. S. (2003). A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In *Uncertainty in Artificial Intelligence: Proceedings of the 19th Conference* (U. Kjærulff and C. Meek, eds.). Morgan Kaufmann, San Francisco, CA, 184–191.
- DRTON, M. and RICHARDSON, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91** 00–00. To appear.
- DRUZDZEL, M. J. and GLYMOUR, C. (1999). Causal inferences from databases: Why universities lose students. In *Computation, Causation, and Discovery* (C. Glymour and G. F. Cooper, eds.), chap. 19. AAAI Press, Menlo Park, CA, 521–539.
- DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103.
- EDWARDS, D. M. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer-Verlag, New York.
- FRYDENBERG, M. (1990). The chain graph Markov property. *Scand. J. Statist.* **17** 333–353.
- HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.
- IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5** 299–314.
- KAUERMANN, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23** 105–116.
- LARNTZ, K. and PERLMAN, M. D. (1988). A simple test for the equality of correlation matrices. In *Statistical Decision Theory and Related Topics, IV, Vol. 2 (West Lafayette, Ind., 1986)*. Springer, New York, 289–298.
- LAURITZEN, S. L. (1996). *Graphical Models*. Oxford University Press, New York.
- LAURITZEN, S. L. and WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

- MOHAMED, W. N., DIAMOND, I. and SMITH, P. F. (1998). The determinants of infant mortality in malaysia: a graphical chain modelling approach. *J. R. Statist. Soc. A* **161** 349–366.
- PEARL, J. and WERMUTH, N. (1994). When can association graphs admit a causal interpretation? In *Selecting Models from Data: Artificial Intelligence and Statistics IV* (P. Cheeseman et al., ed.), vol. 89 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 205–214.
- RICHARDSON, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.* **30** 145–157.
- RICHARDSON, T. S. and SPIRITES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030.
- ROVERATO, A. (1996). Partial correlation coefficient comparison in graphical Gaussian models. In *COMPSTAT. Proceedings in Computational Statistics, 12th Symposium*. Physica-Verlag, Heidelberg, 429–434.
- SHORACK, G. R. (2000). *Probability for Statisticians*. Springer-Verlag, New York.
- ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633.
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA.
- WERMUTH, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972.
- WERMUTH, N. and LAURITZEN, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. B* **52** 21–50, 51–72.
- WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

APPENDIX A. EQUIVALENCE OF THE CHAIN-RECURSIVE PAIRWISE AND GLOBAL MARKOV PROPERTIES FOR AMP CHAIN GRAPHS

The following properties of conditional independence will be used in the subsequent proofs. Let W , X , Y and Z be random variables, and let h be a measurable function. Then,

- (CI1) $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$,
- (CI2) $X \perp\!\!\!\perp Y \mid Z \implies h(X) \perp\!\!\!\perp Y \mid Z$,
- (CI3) $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z, h(X)$,
- (CI4) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid Y, Z \iff X \perp\!\!\!\perp (W, Y) \mid Z$,
- (CI5) $X \perp\!\!\!\perp Y \mid W, Z$ and $X \perp\!\!\!\perp W \mid Y, Z \implies X \perp\!\!\!\perp (W, Y) \mid Z$,
- (CI5*) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid Z \implies X \perp\!\!\!\perp (W, Y) \mid Z$.

The properties (CI1)-(CI4) hold for any underlying distribution but (CI5) and (CI5*) are not generally true. A sufficient condition for (CI5) to hold is that W , X , Y and Z have a positive joint density with respect to a product measure (Andersson et al., 1997, Rem. 3.3). A sufficient condition for (CI5*) to be true is that W , X , Y and Z are jointly normal.

First we show the equivalence of the well-numbered directed and well-numbered directed pairwise Markov properties for ADGs. Let $G = (V, E_{\rightarrow})$ be an ADG with well-numbered

vertex set $V = \{v_1, \dots, p\}$. Here, we show that for any distribution satisfying (CI5), the well-numbered directed Markov property is equivalent to the well-numbered directed pairwise Markov property defined in (2.9). The former property has also been called the ordered directed Markov property (Cowell et al., 1999, p.73) and is equivalent to the global directed Markov property even if (CI5) does not hold (Cowell et al., 1999, Thm. 5.14).

Definition 1. *The well-numbered directed Markov property states that for all $1 \leq i \leq p$,*

$$(A.1) \quad i \perp\!\!\!\perp (\{1, \dots, i-1\} \setminus \text{pa}(i)) \mid \text{pa}(i),$$

where $\text{pa}(i) = \{j \in V \mid j \rightarrow i\}$.

We restate the well-numbered directed pairwise Markov property from (2.9).

Definition 2. *The well-numbered directed pairwise Markov property states that for all $1 \leq i \leq p$ and $1 \leq j < i$,*

$$(A.2) \quad j \not\rightarrow i \implies i \perp\!\!\!\perp j \mid (\{1, \dots, i-1\} \setminus \{j\}).$$

Theorem 3. *If (CI5) holds, then the well-numbered directed pairwise Markov property and the well-numbered directed Markov property are equivalent.*

Proof. (\implies): Let $\{1, \dots, i-1\} \setminus \text{pa}(i) = \{w_1, \dots, w_t\}$. Recall that $\text{pa}(i) \subseteq \{1, \dots, i-1\}$. By (A.2),

$$\begin{aligned} i \perp\!\!\!\perp w_1 \mid (\text{pa}(i) \cup \{w_2, \dots, w_t\}) \\ i \perp\!\!\!\perp w_2 \mid (\text{pa}(i) \cup \{w_1, w_3, \dots, w_t\}) \\ \vdots \\ i \perp\!\!\!\perp w_t \mid (\text{pa}(i) \cup \{w_1, \dots, w_{t-1}\}) \end{aligned}$$

By repeated application of (CI5), we find first that

$$i \perp\!\!\!\perp \{w_1, w_2\} \mid (\text{pa}(i) \cup \{w_3, \dots, w_t\}),$$

and ultimately that

$$i \perp\!\!\!\perp \{w_1, \dots, w_t\} \mid \text{pa}(i),$$

which is (A.1).

(\impliedby): Follows from the converse of (CI4) since $j \in \{1, \dots, i-1\} \setminus \text{pa}(i)$. \square

We now proceed to the main result of this Appendix. Let $G = (V, E_{\rightarrow})$ be a CG. We show that for jointly normal variables the AMP chain-recursive pairwise Markov property is equivalent to the block-recursive Markov property for AMP CGs, which is equivalent to the AMP global Markov property (Andersson et al., 2001, Thm. 2).

Let $\mathbb{D} = (B_k \mid 1 \leq k \leq q)$ be a dependence chain for G , and let $C_\ell = \cup(B_k \mid k \leq \ell)$ be the cumulatives for the dependence chain. Furthermore, let τ_1, \dots, τ_r , $r \geq q$, be the

chain components of G , ordered such that for $v \in \tau_i$ and $w \in \tau_j$ with $v \rightarrow w$ it holds that $i < j$. In the following we adopt the graphical terminology and notation in Andersson et al. (2001). In particular, the ADG $\mathcal{D} = \mathcal{D}(G)$ is the ADG of chain components.

We now restate the AMP chain-recursive pairwise Markov property (cf. §2.4.2).

Definition 4. *The AMP chain-recursive pairwise Markov property states that*

- (i) for all $1 \leq k \leq q$ and $v, w \in B_k$:

$$v \not\perp w \implies v \perp\!\!\!\perp w \mid C_k \setminus \{v, w\};$$

- (ii) for all $1 \leq k < \ell \leq q$ and $v \in B_k, w \in B_\ell$:

$$v \not\perp w \implies v \perp\!\!\!\perp w \mid C_{\ell-1} \setminus \{v\}.$$

We will prove that this pairwise Markov property is equivalent to the AMP block-recursive Markov property (Andersson et al., 2001, Def. 5).

Definition 5. *The AMP block-recursive Markov property states that*

- (i) for all $1 \leq i \leq r$:

$$\tau_i \perp\!\!\!\perp (\text{nd}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_{\mathcal{D}}(\tau_i)) \mid \text{pa}_{\mathcal{D}}(\tau_i);$$

- (ii) for all $1 \leq i \leq r$ the conditional distribution $(\tau_i \mid \text{pa}_{\mathcal{D}}(\tau_i))$ is globally Markov with respect to the UG G_{τ_i} ;

- (iii) for all $1 \leq i \leq r$ and all $\sigma \subseteq \tau_i$:

$$\sigma \perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)) \mid \text{pa}_G(\sigma).$$

Lemma 6. *Let τ_i be a chain component of G . Let $k(\tau_i) \in \{1, \dots, q\}$ be such that $\tau_i \subseteq B_{k(\tau_i)}$. If (CI5) holds, then the AMP chain-recursive pairwise Markov property implies that*

$$(A.3) \quad \tau_i \perp\!\!\!\perp B_{k(\tau_i)} \setminus \tau_i \mid C_{k(\tau_i)-1}.$$

Proof. By Definition 4(i), the conditional distribution $(B_{k(\tau_i)} \mid C_{k(\tau_i)-1})$ is pairwise Markov with respect to the UG $G_{B_{k(\tau_i)}}$. Since (CI5) holds, the pairwise Markov property is equivalent to the global Markov property (Lauritzen, 1996, Thm. 3.7). By the definition of a chain component, no vertex in τ_i is adjacent to a vertex in $B_{k(\tau_i)} \setminus \tau_i$, and the global Markov property for $G_{B_{k(\tau_i)}}$ implies (A.3). \square

Lemma 7. *If (CI5) holds, then the AMP block-recursive Markov property is equivalent to the following requirements:*

- (i) for all $1 \leq i < j \leq r$ such that no vertex in τ_i is a parent of a vertex in τ_j :

$$\tau_i \perp\!\!\!\perp \tau_j \mid (\cup (\tau_1, \dots, \tau_{j-1}) \setminus \tau_i),$$

- (ii) for all $1 \leq i \leq r$ the conditional distribution $(\tau_i \mid \text{pa}_{\mathcal{D}}(\tau_i))$ is pairwise Markov with respect to G_{τ_i} ,

- (iii) as in Definition 5.

Proof. (i) follows from the equivalence of the local and the well-numbered directed pairwise Markov property because τ_1, \dots, τ_r are a well-numbering for the ADG \mathcal{D} . The equivalence of the Markov properties is a consequence of Theorem 3 and Theorem 5.14 in Cowell et al. (1999). (ii) follows because under (CI5), the undirected pairwise Markov property is equivalent to the undirected global Markov property (Lauritzen, 1996, Thm. 3.7). \square

Theorem 8. *If (CI5) and (CI5*) hold, then the AMP chain-recursive pairwise Markov property and the AMP block-recursive Markov property are equivalent.*

Proof of Sufficiency. We need to show that the AMP chain-recursive pairwise Markov property implies the statements (i)-(iii) in Lemma 7.

(i). Let τ_i and τ_j be two chain components such that $\tau_i \in \{\tau_1, \dots, \tau_{j-1}\} \setminus \text{pa}_{\mathcal{D}}(\tau_i)$. Then the AMP chain-recursive pairwise Markov property and Lemma 6 imply that for all $v \in \tau_j$ and $w \in \tau_i$,

$$(A.4) \quad v \perp\!\!\!\perp w \mid (\cup(\tau_1, \dots, \tau_{j-1}) \setminus \{w\}).$$

Let $\tau_j = \{\gamma_1, \dots, \gamma_J\}$, $\tau_i = \{\delta_1, \dots, \delta_I\}$ and $\pi = \cup(\tau_1, \dots, \tau_{j-1})$. By (A.4),

$$\begin{aligned} \gamma_1 &\perp\!\!\!\perp \delta_1 \mid (\pi \setminus \{\delta_1\}) \\ \gamma_1 &\perp\!\!\!\perp \delta_2 \mid (\pi \setminus \{\delta_2\}) \\ &\vdots \\ \gamma_1 &\perp\!\!\!\perp \delta_J \mid (\pi \setminus \{\delta_J\}). \end{aligned}$$

Applying (CI5), we obtain that

$$\gamma_1 \perp\!\!\!\perp \{\delta_1, \delta_2\} \mid (\pi \setminus \{\delta_1, \delta_2\}).$$

Repeated application of (CI5) yield

$$\gamma_1 \perp\!\!\!\perp \tau_i \mid (\pi \setminus \tau_i).$$

Similarly, we obtain

$$\gamma_2 \perp\!\!\!\perp \tau_i \mid (\pi \setminus \tau_i), \dots, \gamma_I \perp\!\!\!\perp \tau_i \mid (\pi \setminus \tau_i),$$

which, by repeated application of (CI5*), yields

$$\tau_i \perp\!\!\!\perp \tau_j \mid (\pi \setminus \tau_i),$$

as required for statement Lemma 7(i).

(ii). Let $1 \leq i \leq r$. Since we have established Lemma 7(i) it follows that Definition 5(i) holds and thus

$$(\tau_i \mid \tau_1, \dots, \tau_{i-1}) = (\tau_i \mid \text{pa}_{\mathcal{D}}(\tau_i)).$$

Therefore Lemma 7(ii) is equivalent to $(\tau_i \mid \tau_1, \dots, \tau_{i-1})$ being pairwise Markov with respect to the UG G_{τ_i} . This is the case if for all non-adjacent $v, w \in \tau_i$ the conditional independence

$$v \perp\!\!\!\perp w \mid (\cup (\tau_1, \dots, \tau_i) \setminus \{v, w\})$$

holds. This, however, follows directly from Definition 4(i) and Lemma 6.

(iii). Let τ_i be a chain component and $\sigma \subseteq \tau_i$ a subset. Let $v \in \sigma$ and $w \in \text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(v)$. In particular, this implies that $w \not\sim v$. Furthermore, let $\tau_i \subseteq B_{k(\tau_i)}$. Then by Definition 4(ii), it follows that

$$(A.5) \quad v \perp\!\!\!\perp w \mid (C_{k(\tau_i)-1} \setminus \{w\}).$$

Moreover, by the already established property (i) in Definition 5, it holds that

$$(A.6) \quad \begin{aligned} & \tau_i \perp\!\!\!\perp (C_{k(\tau_i)-1} \setminus \text{pa}_{\mathcal{D}}(\tau_i)) \mid \text{pa}_{\mathcal{D}}(\tau_i) \\ \stackrel{(CI2)}{\implies} & v \perp\!\!\!\perp (C_{k(\tau_i)-1} \setminus \text{pa}_{\mathcal{D}}(\tau_i)) \mid \text{pa}_{\mathcal{D}}(\tau_i). \end{aligned}$$

Now rewrite (A.5) as

$$(A.7) \quad v \perp\!\!\!\perp w \mid ([\text{pa}_{\mathcal{D}}(\tau_i) \setminus \{w\}] \cup [C_{k(\tau_i)-1} \setminus \text{pa}_{\mathcal{D}}(\tau_i)])$$

and (A.6) as

$$v \perp\!\!\!\perp [C_{k(\tau_i)-1} \setminus \text{pa}_{\mathcal{D}}(\tau_i)] \mid ([\text{pa}_{\mathcal{D}}(\tau_i) \setminus \{w\}] \cup \{w\}).$$

Then it follows from (CI5) and (CI2) that

$$(A.8) \quad v \perp\!\!\!\perp w \mid (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \{w\}).$$

Obviously, for all $v \in \sigma$, $\text{pa}_G(v) \subseteq \text{pa}_G(\sigma)$ and therefore $\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma) \subseteq \text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(v)$. Hence, (A.8) holds in particular for all $v \in \sigma$ and $w \in \text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)$.

Next, let $\sigma = \{v_1, \dots, v_s\}$ and $\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma) = \{w_1, \dots, w_t\}$. Then, by (A.8),

$$\begin{aligned} v_1 & \perp\!\!\!\perp w_1 \mid (\text{pa}_G(\sigma) \cup \{w_2, \dots, w_t\}) \\ v_1 & \perp\!\!\!\perp w_2 \mid (\text{pa}_G(\sigma) \cup \{w_1, w_3, \dots, w_t\}) \\ & \vdots \\ v_1 & \perp\!\!\!\perp w_t \mid (\text{pa}_G(\sigma) \cup \{w_1, \dots, w_{t-1}\}) \end{aligned}$$

By repeated application of (CI5), we find first that

$$v_1 \perp\!\!\!\perp \{w_1, w_2\} \mid (\text{pa}_G(\sigma) \cup \{w_3, \dots, w_t\}),$$

and ultimately that

$$v_1 \perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)) \mid \text{pa}_G(\sigma).$$

Similarly,

$$\begin{aligned} v_2 &\perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)) \mid \text{pa}_G(\sigma) \\ &\vdots \\ v_s &\perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)) \mid \text{pa}_G(\sigma). \end{aligned}$$

Applying (CI5*) repeatedly, we are able to establish that,

$$\sigma \perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau_i) \setminus \text{pa}_G(\sigma)) \mid \text{pa}_G(\sigma),$$

which is Definition 5(iii). \square

Proof of Necessity. In order to show that the AMP block-recursive Markov property implies the AMP chain-recursive pairwise Markov property, we use that the AMP block-recursive Markov property is equivalent to the AMP global Markov property (Andersson et al., 2001, Def. 6, Thm. 2). We consider two cases.

First, let $v, w \in B_k$. Then the AMP global Markov property implies statement (i) in Definition 4 if $C_k \setminus \{v, w\}$ separates v and w in the augmented graph $G[C_k]^a$. Clearly, $\text{An}(C_k) = C_k$ and $\text{Co}(C_k) = C_k$, hence

$$G[C_k] := G_{\text{An}(C_k)} \cup G_{\text{Co}(\text{An}(C_k))}^\wedge = G_{C_k}.$$

The vertices v and w are not adjacent in G_{C_k} and any adjacency that v or w form with another vertex by a directed edge has the arrowhead at v or w , respectively. Thus, v and w are not adjacent in $(G_{C_k})^a = G[C_k]^a$ and the separation implying statement (i) in Definition 4 holds.

Second, let $v \in B_k$ and $w \in B_\ell$, where $k < \ell$. The AMP global Markov property implies statement (ii) in Definition 4 if $C_{\ell-1} \setminus \{v\}$ separates v and w in the augmented graph $G[C_{\ell-1} \cup \{w\}]^a$. Now, $\text{An}(C_{\ell-1} \cup \{w\}) = C_{\ell-1} \setminus \{w\}$ and $\text{Co}(C_{\ell-1} \cup \{w\}) = C_{\ell-1} \cup \tau$, where τ is the chain component $w \in \tau$. It follows that

$$G[C_{\ell-1} \cup \{w\}] = G_{C_{\ell-1} \cup \{w\}} \cup G_\tau.$$

Clearly, v and w are not adjacent in $G[C_{\ell-1} \cup \{w\}]$. Moreover, in the graph $G[C_{\ell-1} \cup \{w\}]$, all directed edges pointing into τ are directed edges pointing into w . Therefore, the subgraph $v \rightarrow \tilde{w} - w$ cannot occur in $G[C_{\ell-1} \cup \{w\}]$, and v and w remain nonadjacent in the augmented graph $G[C_{\ell-1} \cup \{w\}]^a$. This in turn implies that $C_{\ell-1} \setminus \{v\}$ separates v and w in $G[C_{\ell-1} \cup \{w\}]^a$, which concludes the proof. \square

DEPARTMENT OF STATISTICS, UNIVERSITY OF WASHINGTON, SEATTLE, WASHINGTON, U.S.A.

E-mail address: drton@stat.washington.edu, michael@ms.washington.edu