

A SINFUL APPROACH TO MODEL SELECTION FOR GAUSSIAN CONCENTRATION GRAPHS

MATHIAS DRTON AND MICHAEL D. PERLMAN

ABSTRACT. A multivariate Gaussian graphical Markov model for an undirected graph G , also called a covariance selection model or concentration graph model, is defined in terms of the Markov properties, i.e., conditional independences associated with G , which in turn are equivalent to specified zeroes among the set of pairwise partial correlation coefficients. By means of Fisher's z -transformation and Šidák's correlation inequality, conservative simultaneous confidence intervals for the entire set of partial correlations can be obtained, leading to a simple method for model selection that controls the overall error rate for incorrect edge inclusion. The simultaneous p -values corresponding to the partial correlations are partitioned into three disjoint sets, a significant set S , an indeterminate set I , and a non-significant set N . Our SIN model selection method selects two graphs, a graph \hat{G}_{SI} whose edges correspond to the set $S \cup I$, and a more conservative graph \hat{G}_S whose edges correspond to S only. Prior information about the presence and/or absence of particular edges can be incorporated readily. Similar considerations apply to covariance graph models, which are defined in terms of marginal independence rather than conditional independence.

1. INTRODUCTION

Let $G \equiv (V, E)$ be an *undirected* graph with vertex set $V \equiv \{1, \dots, p\}$ and edge set $E \equiv \{e^{ij}\}$, where $e^{ij} = 1$ or 0 according to whether vertices i and j , $1 \leq i < j \leq p$, are adjacent in G or not. The *Gaussian graphical Markov model* $N(G)$ consists of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ where both the mean vector μ and the covariance matrix Σ (assumed nonsingular) are unknown but where the concentration (or precision) matrix $\Sigma^{-1} \equiv \{\sigma^{ij}\}$ satisfies the following linear restrictions:

$$(1.1) \quad e^{ij} = 0 \quad \implies \quad \sigma^{ij} = 0.$$

The model $N(G)$ has also been called a *covariance selection model* (Dempster [7]) and a *concentration graph model* (Cox and Wermuth [5]); we shall use the latter term in juxtaposition to the term *covariance graph model* to be considered in Section 6. The reader is referred to Edwards [10], Lauritzen [15], or Whittaker [25] for statistical properties

Date: December 11, 2003.

Key words and phrases. Graphical model selection, concentration graphs, covariance selection models, covariance graphs, Šidák's inequality, simultaneous confidence intervals.

of these models, including methods for parameter estimation, model testing, and model selection.

The model $N(G)$ also can be defined in terms of the pairwise conditional independences determined by the Markov properties of G . If $Y \equiv (Y_1, \dots, Y_p)^t \sim \mathcal{N}_p(\mu, \Sigma)$, then

$$(1.2) \quad \sigma^{ij} = 0 \quad \iff \quad Y_i \perp\!\!\!\perp Y_j \mid Y_{V \setminus \{i,j\}} \quad \iff \quad \rho^{ij} = 0,$$

where

$$(1.3) \quad \rho^{ij} \equiv \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$$

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 [15, p.130]). This suggests that model selection, equivalently, determination of the graph G , can be based on the set of sample partial correlations $\{r^{ij}\}$ arising from i.i.d. observations $Y^{(1)}, \dots, Y^{(n+1)} \sim \mathcal{N}_p(\mu, \Sigma)$, where $n \geq p$ is assumed in order to guarantee positive definiteness of the sample covariance matrix

$$(1.4) \quad W \equiv \frac{1}{n} \sum_{m=1}^{n+1} (Y^{(m)} - \bar{Y})(Y^{(m)} - \bar{Y})^t.$$

(The methods presented in Section 2 will require that $n \geq p + 1$.)

The standard approach to model selection, described in Edwards [10, §6.1, also see §3.1] is *backward stepwise selection*. At the first step, the unrestricted (\equiv saturated) model $N(\bar{G})$ is assumed to be true, where \bar{G} is the *complete graph* in which *all* edges are present, and each of the $f_0 \equiv p(p-1)/2$ edges is tested individually at level α . Those edges that are not statistically significant are removed, yielding a reduced graph $G^{(1)}$ with $f_1 \leq f_0$ remaining edges. More precisely, the $p(p-1)/2$ testing problems

$$(1.5) \quad H^{ij} : \rho^{ij} = 0 \quad \text{vs.} \quad K^{ij} : \rho^{ij} \neq 0 \quad (1 \leq i < j \leq p)$$

are tested *individually* at level α , based on the sample partial correlations $\{r^{ij}\}$. If r^{ij} is not significantly different from 0, i.e., if the corresponding individual p-value is larger than α , then H^{ij} is accepted and the edge $i-j$ is removed from \bar{G} .

At the second step the model $N(G^{(1)})$ is assumed to be true and one tests each of the f_1 remaining edges individually, again at level α , using the appropriate normal-theory likelihood ratio test. Again those edges that are not statistically significant are removed from $G^{(1)}$, yielding a reduced graph $G^{(2)}$ with $f_2 \leq f_1$ edges. This procedure is repeated until no more edges are removed, resulting in a final graph $G^{(k)}$ and selected model $N(G^{(k)})$, where $k \geq 0$. Note that the actual implementation of the stepwise model selection procedure in Edwards' statistical software package MIM is based on removing only one (the least significant) edge at a time – compare Edwards [10, p.158].

Because the overall error rate for this stepwise method is not controlled, however, its validity as a simultaneous testing procedure is uncertain. Edwards [10, p.158] states that “this [stepwise selection procedure] 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.” Again on p.172: “Its sampling properties seem to be intractable.”

In the present paper we note that a simple and reliable set of simultaneous confidence intervals for the entire set of partial correlations $\{\rho^{ij}\}$ can be obtained by modifying the method of Larntz and Perlman [14] for obtaining simultaneous confidence intervals for the set of ordinary correlations $\{\rho_{ij}\}$. Both sets of simultaneous intervals are obtained by applying Fisher’s variance-stabilizing *z-transformation* to the sample correlations or sample partial correlations, respectively. These *z*-transforms stabilize only the marginal asymptotic distributions of the sample correlations or sample partial correlations rather than their joint asymptotic distribution, but by applying a well-known inequality of Šidák [24], conservative simultaneous $1 - \alpha$ confidence intervals can be obtained that require only the marginal asymptotic distributions. The simultaneous intervals for the partial correlations can then be applied to obtain a conservative simultaneous test for the $p(p-1)/2$ hypotheses in (1.5) at overall level α and thus to select a model in a single step with conservative overall confidence level $1 - \alpha$.

The proposed model selection procedure is described in detail in Section 2, together with results concerning its overall error rate and consistency. In Section 3 this procedure is applied to a series of well-known examples via the SIN approach and the results compared to those obtained using backward stepwise selection via MIM and Bayesian model selection. These examples suggest that our SIN procedure tends to be (appropriately) more conservative with respect to edge inclusion than existing methods – unlike all previous procedures, SIN controls the overall error rate for incorrect edge inclusion. Furthermore, SIN is computationally simpler in that it is accomplished in a single step.

Additional asymptotic properties and the determination of an appropriate value of α are discussed in Section 4, illustrated by a series of simulations from known models. In Section 5 we show that the proposed model selection procedure readily incorporates prior information about the presence and/or absence of particular edges. In Section 6 the original Larntz-Perlman simultaneous confidence intervals for the ordinary correlations are applied to model selection in the class of *covariance graphs*, i.e., graphs where the absence of an edge corresponds to *marginal* independence of the two variables rather than to conditional independence (cf. Cox and Wermuth [5], Kauermann [13], Richardson and Spirtes [19]). Our results are summarized in Section 7.

2. METHODOLOGY

Usually, G shall indicate the true graph, that is, the data $Y^{(1)}, \dots, Y^{(n+1)}$ will be assumed to have been generated from the model $N(G)$. As above, let $\{r^{ij} \mid 1 \leq i < j \leq p\}$ be the sample partial correlations determined from W , a Wishart random matrix with an asymptotically normal joint distribution with mean Σ . The $\{r^{ij}\}$ are smooth functions of W^{-1} (Anderson [1, Exercise 2.47]) and therefore of W . It follows from the delta method (Shorack [23, §11.6]) that the joint distribution of the $\{r^{ij}\}$ is also

asymptotically normal with mean $\{\rho^{ij}\}$. Because the asymptotic covariance matrix also depends on the unknown parameters, direct inversion of this asymptotic distribution to obtain a simultaneous confidence region for the $\{\rho^{ij}\}$ is not feasible. As in Larntz and Perlman [14, p.296-7], however, this difficulty can be overcome by applying Šidák's [24] inequality, as now described.

The marginal distribution of r^{ij} has the same form as the distribution of the ordinary sample correlation r_{ij} , but with the parameter ρ_{ij} replaced by ρ^{ij} and the degrees of freedom reduced from n to $n - (p - 2) = n - p + 2$ (Anderson [1, Thm. 4.3.5]). For moderate or large values of n , the z -transform of r^{ij} , given by

$$(2.1) \quad z^{ij} = \frac{1}{2} \ln \left(\frac{1 + r^{ij}}{1 - r^{ij}} \right),$$

substantially increases the accuracy of the normal approximation to the marginal distribution, as follows (Anderson [1, p.123]):

$$(2.2) \quad \sqrt{n_p} (z^{ij} - \zeta^{ij}) \rightarrow \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $n_p \equiv (n - p + 2) - 2 = n - p$ and ζ^{ij} is the z -transform of ρ^{ij} . Note that

$$(2.3) \quad \zeta^{ij} = 0 \iff \rho^{ij} = 0.$$

Furthermore, (2.1) is a *variance-stabilizing* transformation in that the asymptotic variance in (2.2) depends on no unknown parameters. By Šidák's inequality applied to the asymptotic joint normal distribution of $\{r^{ij}\}$, the following inequality holds approximately for large n :

$$(2.4) \quad \begin{aligned} & \Pr_G[|z^{ij} - \zeta^{ij}| \leq n_p^{-1/2} b_p(\alpha), 1 \leq i < j \leq p] \\ & \geq \prod_{1 \leq i < j \leq p} \Pr_G[|z^{ij} - \zeta^{ij}| \leq n_p^{-1/2} b_p(\alpha)] \\ & = 1 - \alpha, \end{aligned}$$

where $b_p(\alpha)$ is determined by

$$(2.5) \quad [2\Phi(b_p(\alpha)) - 1]^{p(p-1)/2} = 1 - \alpha,$$

or equivalently by

$$(2.6) \quad b_p(\alpha) = \Phi^{-1} \left(\frac{1}{2} [(1 - \alpha)^{2/p(p-1)} + 1] \right),$$

with Φ the standard normal cumulative distribution function. Finally, the inequalities in the first expression in (2.4) can be inverted to yield the following set of conservative simultaneous $1 - \alpha$ confidence intervals for $\{\zeta^{ij}\}$:

$$(2.7) \quad z^{ij} - n_p^{-1/2} b_p(\alpha) \leq \zeta^{ij} \leq z^{ij} + n_p^{-1/2} b_p(\alpha) \quad (1 \leq i < j \leq p),$$

which can be converted to a set of simultaneous intervals for $\{\rho^{ij}\}$ by inverting (2.1).

Because $\rho^{ij} = 0$ iff $\zeta^{ij} = 0$, it follows from (1.1) and (1.2) that the simultaneous intervals (2.7) can be used for model selection as follows: If 0 is contained in the ij -th

interval then the data are compatible with the hypothesis $H^{ij} : \rho^{ij} = 0$ at the overall confidence level $1 - \alpha$, so the edge $i - j$ is *not included* in the selected graph $\hat{G}_\alpha \equiv (V, \hat{E}_\alpha)$. Formally, if $\hat{E}_\alpha \equiv \{\hat{e}_\alpha^{ij}\}$, then

$$(2.8) \quad \hat{e}_\alpha^{ij} = \begin{cases} 0 & \text{if } |z^{ij}| \leq n_p^{-1/2} b_p(\alpha), \\ 1 & \text{if } |z^{ij}| > n_p^{-1/2} b_p(\alpha). \end{cases}$$

Note that the selected graph \hat{G}_α is not restricted to be decomposable, unlike the case e.g. in Giudici and Green [12].

Our model selection procedure can be described equivalently in terms of the set of *simultaneous* p-values $\{\pi^{ij} \equiv \pi(z^{ij}) \mid 1 \leq i < j \leq p\}$ determined by the intervals (2.7) for the $p(p-1)/2$ testing problems (1.5). It follows from (2.6) and (2.7) that

$$(2.9) \quad |z^{ij}| \leq n_p^{-1/2} b_p(\alpha) \iff \alpha \leq 1 - [2\Phi(\sqrt{n_p} |z^{ij}|) - 1]^{p(p-1)/2} \equiv \pi(z^{ij}),$$

so our selected graph $\hat{G}_\alpha \equiv (V, \hat{E}_\alpha)$ is given by

$$(2.10) \quad \hat{e}_\alpha^{ij} = \begin{cases} 0 & \text{if } \pi^{ij} \geq \alpha, \\ 1 & \text{if } \pi^{ij} < \alpha. \end{cases}$$

Thus, the edge $i - j$ is included in \hat{G}_α iff π^{ij} is significantly small at overall level α .

Note that by (2.4) and (2.9), $\{\pi^{ij}\}$ is a set of *conservative simultaneous* p-values for (1.5) in the sense that

$$(2.11) \quad \Pr_{G_\emptyset}[\pi^{ij} \geq \alpha, 1 \leq i < j \leq p] \geq 1 - \alpha,$$

where $G_\emptyset \equiv (V, E_\emptyset)$ denotes the graph with no edges. ($N(G_\emptyset)$ is the model of complete independence.) Again by (2.4) and (2.9), for a general graph $G \equiv (V, E \equiv \{e^{ij}\})$,

$$(2.12) \quad \Pr_G[\pi^{ij} \geq \alpha \forall ij \in E_0] \geq 1 - \alpha,$$

where $E_0 = \{ij \mid e^{ij} = 0\}$ indicates the set of edges *absent* in G .

The inequality (2.12) yields results concerning the overall error rate of our proposed model selection procedure. For two graphs $G \equiv (V, E)$ and $G' \equiv (V, E')$ with the same vertex set V , we say that G' is a *subgraph* of G (denoted by $G' \subseteq G$) if $E' \subseteq E$, that is, if G' is obtained by removing one or more edges from G . It is readily seen from (1.1) and (1.2) that $N(G')$ is a *submodel* of $N(G)$ (i.e., $N(G') \subseteq N(G)$) iff $G' \subseteq G$.

First, it follows directly from (2.12) that

$$(2.13) \quad \Pr_G[\hat{G}_\alpha \subseteq G] \geq 1 - \alpha$$

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

Second, by (2.8), (1.1), (1.2), and (2.2),

$$\begin{aligned}
\Pr_G[\hat{G}_\alpha \supseteq G] &= \Pr_G[\hat{e}_\alpha^{ij} = 1 \forall ij \in E_1] \\
&= \Pr_G[|z^{ij}| > n_p^{-1/2} b_p(\alpha) \forall ij \in E_1] \\
&= 1 - \Pr_G[|z^{ij}| \leq n_p^{-1/2} b_p(\alpha) \text{ for some } ij \in E_1] \\
&\geq 1 - \sum_{ij \in E_1} \Pr_G[|z^{ij}| \leq n_p^{-1/2} b_p(\alpha)] \\
&= 1 - \sum_{ij \in E_1} \Pr_G[n_p (z^{ij})^2 \leq b_p^2(\alpha)] \\
(2.14) \qquad &= 1 - \sum_{ij \in E_1} \Pr[\chi_1^2(n_p (\zeta^{ij})^2) \leq b_p^2(\alpha)]
\end{aligned}$$

$$(2.15) \qquad \geq 1 - |E_1| \cdot \Pr[\chi_1^2(n_p \Lambda^2) \leq b_p^2(\alpha)]$$

(to the accuracy of the normal approximations), where $E_1 = \{ij \mid e^{ij} = 1\}$ indicates the set of edges *present* in G ,

$$(2.16) \qquad \Lambda \equiv \Lambda(G, \Sigma) = \min\{|\zeta^{ij}| \mid ij \in E_1\},$$

and $\chi_1^2(\delta)$ denotes a noncentral chi-square variate with one degree of freedom and noncentrality parameter δ . Assume now that the data-generating distribution is *faithful* to the graph, i.e., $\sigma^{ij} = 0$ iff $e^{ij} = 0$. Then for $G \supsetneq G_\emptyset$ it follows that $|E_1| > 0$ and $\Lambda > 0$, so the bound (2.15) goes to 1 as $n \rightarrow \infty$, hence

$$(2.17) \qquad \lim_{n \rightarrow \infty} \Pr_G[\hat{G}_\alpha \supseteq G] = 1,$$

while (2.17) holds trivially if $G = G_\emptyset$. Combining (2.13) and (2.17) we conclude that, under the faithfulness assumption, our procedure is $(1 - \alpha)$ -consistent:

$$(2.18) \qquad \liminf_{n \rightarrow \infty} \Pr_G[\hat{G}_\alpha = G] \geq 1 - \alpha.$$

Thus, for fixed α the correct model is selected with probability $\geq 1 - \alpha$ for large sample size. Note that if the data-generating distribution $\mathcal{N}(\mu, \Sigma)$ is not faithful to G then our procedure is $(1 - \alpha)$ -consistent for the inclusion-minimal graph G^* such that $\mathcal{N}(\mu, \Sigma) \in \mathcal{N}(G^*)$.

If $\Lambda > 0$ is *known or specified* (equivalently, if $\min\{|\rho^{ij}| \mid ij \in E_1\}$ is known or specified), then the sample size can be chosen large enough that $\Pr_G[\hat{G}_\alpha \neq G]$ is arbitrarily small, so in this sense our procedure is *fully consistent*. This is seen as follows. Select $\alpha > 0$ and $\beta > 0$ so that $\alpha + \beta$ is arbitrarily small. Then by (2.13) and (2.15),

$$\begin{aligned}
\Pr_G[\hat{G}_\alpha \neq G] &\leq \Pr_G[\hat{G}_\alpha \not\supseteq G] + \Pr_G[\hat{G}_\alpha \not\subseteq G] \\
&\leq \alpha + [p(p-1)/2] \Pr[\chi_1^2(n_p \Lambda^2) \leq b_p^2(\alpha)] \\
(2.19) \qquad &\leq \alpha + [p(p-1)/2] [1 - \Phi(\sqrt{n_p} \Lambda - b_p(\alpha))].
\end{aligned}$$

Now select $n \equiv n(\alpha, \beta, \Lambda, p)$ large enough that the second term in (2.19) is $\leq \beta$, that is, so that

$$(2.20) \quad n \geq p + \frac{1}{\Lambda^2} \left[b_p(\alpha) + \Phi^{-1} \left(1 - \frac{2\beta}{p(p-1)} \right) \right]^2$$

$$(2.21) \quad = p + \frac{1}{\Lambda^2} \left[\Phi^{-1} \left(\frac{1}{2} [(1-\alpha)^{2/p(p-1)} + 1] \right) + \Phi^{-1} \left(1 - \frac{2\beta}{p(p-1)} \right) \right]^2$$

$$(2.22) \quad \approx p + \frac{1}{\Lambda^2} \left[\sqrt{2 \ln \left(\frac{p(p-1)}{\alpha} \right)} + \sqrt{2 \ln \left(\frac{p(p-1)}{2\beta} \right)} \right]^2$$

as $\alpha, \beta \rightarrow 0$. This guarantees that $\Pr_G[\hat{G}_\alpha \neq G] \leq \alpha + \beta$, as required.

3. EXAMPLES

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}_α . In practice, a simple plot of the entire set of simultaneous p-values $\{\pi^{ij}\}$ 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.

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 $S \cup I$. This holds in all the examples in this section except for Example 3.1, where, perhaps due to the large sample sizes, $I = \emptyset$ so $\hat{G}_S = \hat{G}_{SI}$.

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. (However, see §4.) 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 - see Figure 3.9.

We now apply the SIN model selection method to eight well-known data sets. We compare the selected models \hat{G}_S and \hat{G}_{SI} to the results of alternative model selection procedures from the literature, most often the backward stepwise selection method in Edwards' MIM package with the option of *unrestricted* selection wherein both decomposable and non-decomposable models are considered. In order to avoid confusion, we denote the sample size $n + 1$ by n_1 .

Example 3.1 (Anxiety and Anger). The Anxiety and Anger data consists of $p = 4$ psychological measurements on $n_1 = 684$ students. The variables are “anxiety state”,

“anger state”, “anxiety trait”, and “anger trait”, labeled as 1, 2, 3, and 4, respectively. The data are also treated by Cox and Wermuth [4, Ex.1], and Edwards [10, Ex.3.1.5].

Psychological theory suggests the conditional independences $1 \perp\!\!\!\perp 4 \mid 2, 3$ and $2 \perp\!\!\!\perp 3 \mid 1, 4$ that correspond to the graphical model based on the 4-cycle in Figure 3.1. This model is selected by backward stepwise selection in MIM with individual confidence level $\alpha = 0.05$ for each test of edge presence.

The SIN procedure yields the simultaneous p-values π^{ij} given numerically in Table 1

TABLE 1. Simultaneous p-values π^{ij} for Anxiety and Anger.

	anx st (1)	ang st (2)	anx tr (3)	ang tr (4)
anxiety state (1)				
anger state (2)	0.00			
anxiety trait (3)	0.00	0.99		
anger trait (4)	0.85	0.00	0.00	

and visually in Figure 3.9. This figure shows a clear separation of the π^{ij} into two groups: $S = \{0.00^{(4)}\}$ corresponding to included edges 12, 13, 24, and 34, and $N = \{0.85, 0.99\}$ corresponding to excluded edges 14 and 23. Here $I = \emptyset$, which is not surprising in view of the large sample size. Thus, for $\alpha \in (0.00, 0.85)$, our selected model $\hat{G}_\alpha \equiv \hat{G}_S \equiv \hat{G}_{SI}$ coincides with the 4-cycle obtained via MIM.

Note that since we round the p-values to two decimal digits for presentation a p-value can appear multiple times in one SIN-set. Therefore, we adopt here and in the following examples the notation $\pi^{(m)}$ for a p-value π in S, I, or N which arises for m edges (in two digits of accuracy).

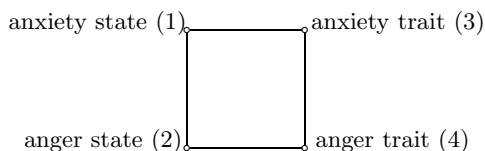


FIGURE 3.1. Anxiety and Anger (MIM and SIN).

Example 3.2 (Cork Borings). The data are presented in Whittaker [25, Exercise 8.6.5] and were originally used by Rao [18]. The $p = 4$ measurements are the weight of cork borings on $n_1 = 28$ trees in the four directions: North, East, South, and West, which we label as 1, 2, 3, and 4, respectively.

The model selected for these data by unrestricted backward stepwise selection in MIM with individual confidence level $\alpha = 0.05$ is shown in Figure 3.2(a). The SIN procedure yields the simultaneous p-values π^{ij} given in Table 2 and illustrated in Figure 3.9. This figure suggests a separation of the π^{ij} into three groups: $S = \{0.00, 0.01\}$ corresponding

TABLE 2. Simultaneous p-values π^{ij} for Cork Borings.

	N (1)	E (2)	S (3)	W (4)
North (1)				
East (2)	0.01			
South (3)	0.71	0.90		
West (4)	0.44	0.95	0.00	

to included edges 34 and 12 respectively, $N = \{0.71, 0.90, 0.95\}$ corresponding to excluded edges 13, 23, and 24 respectively, and $I = \{0.44\}$ corresponding to a possible edge 14. (In view of the small sample size, it is not surprising that $I \neq \emptyset$.)

Thus our two selected models are $\hat{G}_\alpha \equiv \hat{G}_S$ shown in Figure 3.2(b) and $\hat{G}_\alpha \equiv \hat{G}_{SI}$ shown in Figure 3.2(c). We note that neither model agrees with that selected by MIM, the main difference being that MIM includes the edge 13, which would seem to be less biologically significant than the edge 14 included in our \hat{G}_{SI} .

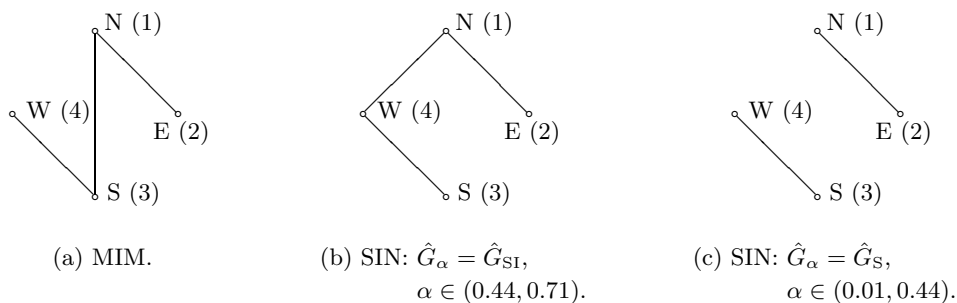


FIGURE 3.2. Cork Borings.

Example 3.3 (Fisher’s Iris data). The data is used by Roverato [20] to illustrate a Bayesian methodology for model selection. More specifically, he studies the $n_1 = 50$ flowers of the species “Virginica”. The $p = 4$ variables are “Sepal length”, “Sepal width”, “Petal length”, and “Petal width”, which we label by 1, 2, 3, and 4, respectively.

Roverato’s Bayesian model selection procedure yields the two graphs shown in Figure 3.3(a) and (b) as the two models with highest posterior probability. Backward selection in MIM at individual level $\alpha = 0.05$ leads to the model in Figure 3.3(b).

The models selected by SIN are found using the p-values quoted in Table 3, which are plotted in Figure 3.9. From this figure we find $S = \{0.00^{(2)}\}$ corresponding to included edges 13, 24, $N = \{0.78, 0.95, 1.00\}$ corresponding to excluded edges 34, 14, 23, and $I = \{0.33\}$ corresponding to possible edge 12. Our two selected models are $\hat{G}_\alpha \equiv \hat{G}_S$ shown in Figure 3.3(c) and $\hat{G}_\alpha \equiv \hat{G}_{SI}$ shown in Figure 3.3(b). Our simultaneous p-values show immediately that the length-length edge 13 and the width-width edge 24 are highly significant, with no other edges strongly supported. Neither our selected models nor

TABLE 3. Simultaneous p-values π^{ij} for Fisher's Iris data.

	SL (1)	SW (2)	PL (3)	PW (4)
Sepal length (1)				
Sepal width (2)	0.33			
Petal length (3)	0.00	1.00		
Petal width (4)	0.95	0.00	0.78	

the MIM model agrees with Roverato's highest posterior probability model. The latter includes the edge 34, which might be considered dubious in view of the non-significant simultaneous p-value 0.78 associated with its inclusion.

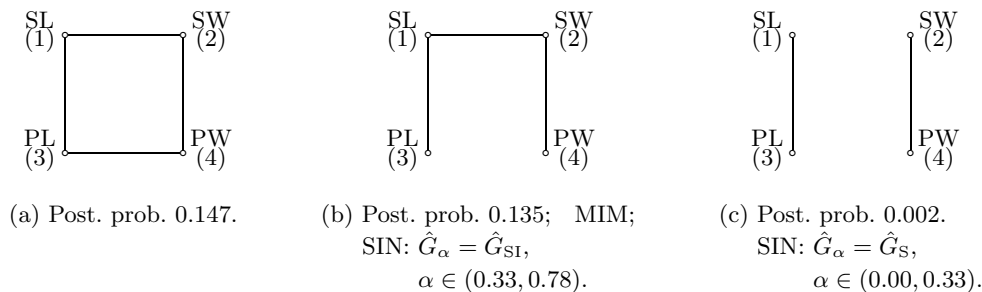


FIGURE 3.3. Fisher's Iris data.

Example 3.4 (Frets' Heads). The $p = 4$ variables in the data are the head length and head breadth of the first son, and the head length and head breadth of the second son, labeled 1, 2, 3, and 4, respectively. The data was collected on $n_1 = 25$ pairs of first and second sons and can be found in Whittaker [25, Exercise 8.6.1]. It is also studied by Dellaportas, Giudici, and Roberts [6], Giudici [11], and Brooks, Giudici, and Roberts [2].

Backward selection in MIM at level 0.05 yields the 4-cycle shown in Figure 3.4(a). This non-decomposable model also receives the largest posterior probability in Dellaportas, Giudici, and Roberts [6]. Giudici [11], and Brooks, Giudici, and Roberts [2] restrict the Bayesian model search to decomposable models and find that the two models that add a single chord to the 4-cycle have the largest posterior probabilities.

From the simultaneous p-values in Table 4 and Figure 3.9 we find $S = \{0.01\}$ corresponding to included edge 34, $N = \{0.89^{(2)}, 0.98, 0.99\}$ corresponding to excluded edges 13, 24, 14, 23, and $I = \{0.23\}$ corresponding to possible edge 12. SIN selects $\hat{G}_\alpha \equiv \hat{G}_S$ in Figure 3.4(b) and $\hat{G}_\alpha \equiv \hat{G}_{SI}$ in Figure 3.4(c). The former model corresponds to the easily interpretable independence (HLI, HBI) $\perp\!\!\!\perp$ (HLII, HBII), that is, the two characteristics of the first son are independent of those of the second son. Our simultaneous p-values strongly suggest that the 13 and 24 edges present in the model selected by MIM and Bayesian methods are highly non-significant and not supported by the data.

TABLE 4. Simultaneous p-values π^{ij} for Frets’ Heads.

	HLI (1)	HBI (2)	HLII (3)	HBII (4)
HLI (1)				
HBI (2)	0.23			
HLII (3)	0.89	0.99		
HBII (4)	0.98	0.89	0.01	

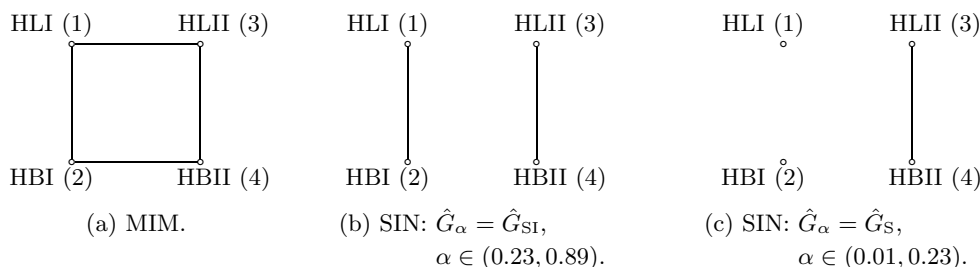


FIGURE 3.4. Frets’ Heads.

Example 3.5 (Math Marks). Mardia, Kent, and Bibby [17, pp. 3–4] present the marks of $n_1 = 88$ students in $p = 5$ examinations in mechanics, vectors, algebra, analysis, and statistics, labeled by 1, 2, 3, 4, and 5, respectively. The data also appear in Edwards [10].

Backward selection in MIM at individual level 0.05 gives the “butterfly” graph in Figure 3.5(a). From the p-values in Table 5 and Figure 3.9 we take $S = \{0.00, 0.01, 0.02\}$ corresponding to included edges 34, 35, and 12, and $N = \{1.00^{(4)}\}$ corresponding to excluded edges 14, 15, 24, and 25. The indeterminate set of p-values is $I = \{0.09, 0.18, 0.29\}$ corre-

TABLE 5. Simultaneous p-values π^{ij} for Math Marks.

	mec (1)	vec (2)	alg (3)	ana (4)	stat (5)
mechanics (1)					
vectors (2)	0.02				
algebra (3)	0.29	0.09			
analysis (4)	1.00	1.00	0.00		
statistics (5)	1.00	1.00	0.01	0.18	

sponding to possible edges 23, 45, and 13. Thus SIN selects $\hat{G}_\alpha \equiv \hat{G}_S$ for $\alpha \in (0.02, 0.09)$ and $\hat{G}_\alpha \equiv \hat{G}_{SI}$ for $\alpha \in (0.29, 1.00)$, which are shown in Figure 3.5(a) and (b).

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),

which may also be cognitively interpretable. This illustrates the ease with which SIN reveals alternative models that may convey scientifically meaningful information.

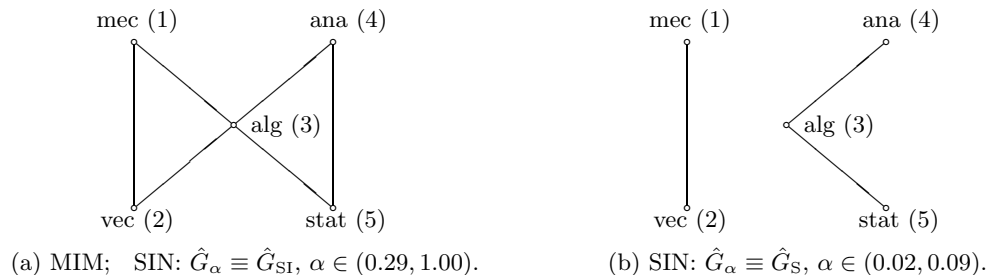


FIGURE 3.5. Math Marks.

Example 3.6 (Fowl Bones). The Fowl Bones data consists of $n_1 = 276$ measurements on $p = 6$ bones. The variables are “skull length”, “skull breadth”, “humerus”, “ulna”, “femur”, and “tibia”, which we label as 1 through 6. The correlation matrix for these data is given in Whittaker [25, Exercise 8.6.3]. In the recent literature, the data are employed by Roverato and Whittaker [22], and Brooks, Giudici, and Roberts [2].

Roverato and Whittaker [22] perform a stepwise model search using multiple individual Wald tests (level 0.05) and find the model in Figure 3.6(a). The same model is found by MIM’s backward selection at the individual level 0.05. In fact, Roverato and Whittaker’s procedure is identical to MIM’s backward selection except for the use of the Wald test instead of likelihood ratio tests.

The simultaneous p-values for SIN are given in Table 6 and illustrated in Figure 3.9. We suggest taking $S = \{0.00^{(4)}, 0.03\}$ corresponding to included edges 12, 34, 46, 56, and 23, $N = \{0.59, 0.68, 0.82, 0.92, 0.98, 0.99^{(2)}, 1.00^{(2)}\}$ corresponding to excluded edges 45, 24, 26, 16, 36, 13, 14, 15, and 25, and $I = \{0.07\}$ corresponding to the possible edge 35. (We do not assign 0.07 to S due to the relatively large sample size, but this is subjective.) The graphs selected by SIN are $\hat{G}_\alpha = \hat{G}_S$ for $\alpha \in (0.03, 0.07)$ and $\hat{G}_\alpha = \hat{G}_{SI}$ for $\alpha \in (0.07, 0.59)$, illustrated in Figure 3.6(b) and (c). Both SIN models are

TABLE 6. Simultaneous p-values π^{ij} for Fowl Bones.

	sl (1)	sb (2)	hum (3)	ul (4)	fem (5)	tib (6)
skull length (1)						
skull breadth (2)	0.00					
humerus (3)	0.99	0.03				
ulna (4)	0.99	0.68	0.00			
femur (5)	1.00	1.00	0.07	0.59		
tibia (6)	0.92	0.82	0.98	0.00	0.00	

more conservative than the Roverato/Whittaker and MIM model, and appear easier to interpret anatomically.

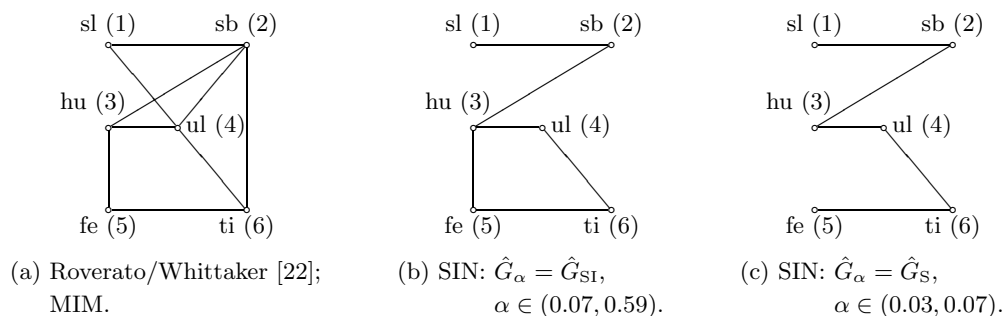


FIGURE 3.6. Fowl Bones.

Example 3.7 (HIV). Roverato and Whittaker [21] consider HIV data consisting of the $p = 6$ variables “immunoglobulin G”, “immunoglobulin A”, “lymphocyte B”, “platelet count”, “lymphocyte T4”, and “T4/T8 lymphocyte ratio”, labeled as 1 through 6, measured on $n_1 = 107$ babies. Roverato and Whittaker [21] state that a reasonable model

TABLE 7. Simultaneous p-values π^{ij} for HIV.

	G (1)	A (2)	B (3)	plat (4)	T4 (5)	T4/T8 (6)
imm. G (1)						
imm. A (2)	0.00					
lymph. B (3)	1.00	1.00				
plat. count (4)	1.00	1.00	1.00			
lymph. T4 (5)	0.00	0.29	0.00	1.00		
T4/T8 ratio (6)	0.22	0.99	0.01	1.00	0.00	

according to expert opinion is given in Figure 3.7(a). This model is also obtained by MIM’s backward selection at individual level 0.05. The SIN p-values are shown in Table

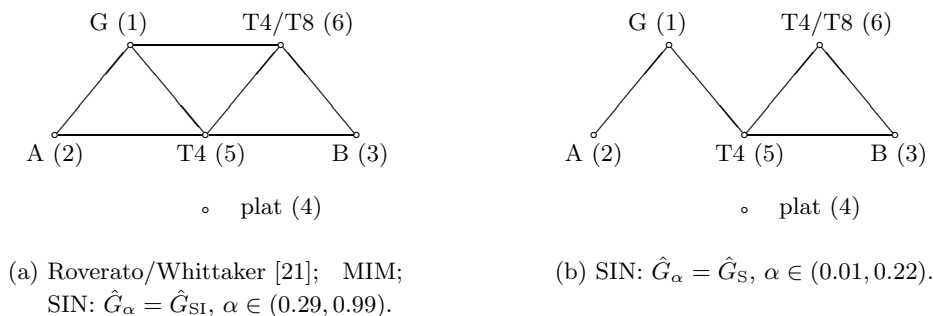


FIGURE 3.7. HIV.

7 and Figure 3.9. The significant p-values are $S = \{0.00^{(4)}, 0.01\}$ for included edges 12, 15, 35, 56, and 36. The non-significant p-values are $N = \{0.99, 1.00^{(7)}\}$ for excluded edges 26, 13, 14, 23, 24, 34, 45, and 46. The indeterminate p-values are $I = \{0.22, 0.29\}$ for possible edges 16 and 25. Thus, SIN selects $\hat{G}_\alpha = \hat{G}_{SI}$ shown in Figure 3.7(a) for $\alpha \in (0.29, 0.99)$, which agrees with the model selected by experts and by MIM, and, more conservatively, SIN selects $\hat{G}_\alpha = \hat{G}_S$ shown in Figure 3.7(b) for $\alpha \in (0.01, 0.22)$.

Example 3.8 (School Grades). Whittaker [25, Exercise 8.6.2] gives data on $n_1 = 220$ boys tested on $p = 6$ school subjects: “Gaelic”, “English”, “history”, “arithmetic”, “algebra”, and “geometry”, here labeled as 1 through 6.

A model search in MIM at individual level 0.05 finds the model in Figure 3.8(a). The simultaneous p-values for SIN are given in Table 8. Their illustration in Figure

TABLE 8. Simultaneous p-values π^{ij} for School Grades.

	Gae (1)	Eng (2)	his (3)	ari (4)	alg (5)	geo (6)
Gaelic (1)						
English (2)	0.00					
history (3)	0.00	0.26				
arithmetic (4)	1.00	0.94	0.00			
algebra (5)	0.12	1.00	0.74	0.00		
geometry (6)	1.00	0.34	1.00	0.01	0.01	

3.9 suggests the partition $S = \{0.00^{(4)}, 0.01^{(2)}\}$ corresponding to included edges 12, 13, 34, 45, 46, and 56, $N = \{0.74, 0.94, 1.00^{(4)}\}$ corresponding to excluded edges 35, 24, 14, 16, 25, and 36, and $I = \{0.12, 0.26, 0.34\}$ corresponding to possible edges 15, 23, and 26. Hence, SIN chooses the model $\hat{G}_\alpha = \hat{G}_{SI}$ for $\alpha \in (0.34, 0.74)$ and $\hat{G}_\alpha = \hat{G}_S$ for $\alpha \in (0.01, 0.12)$, shown in Figure 3.8(b) and (c).

Moving from left to right in Figure 3.8 the selected models become progressively more conservative – and probably more accurate, since in the final model the dubious Gaelic-algebra and English-geometry edges are excluded.

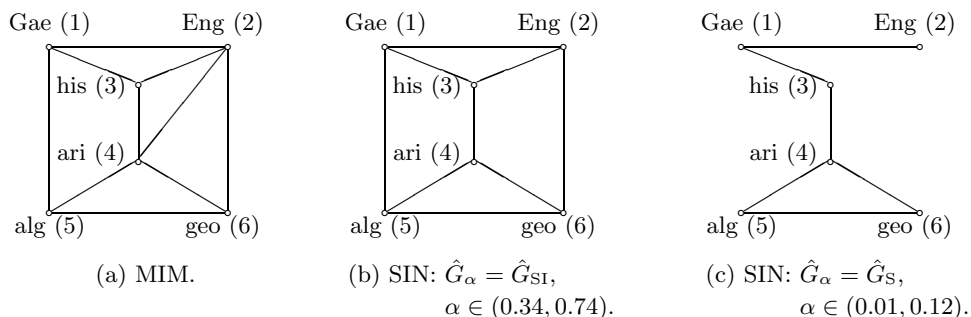
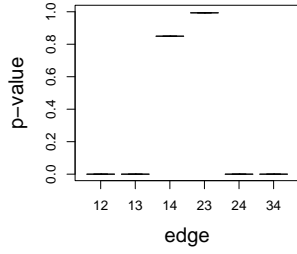
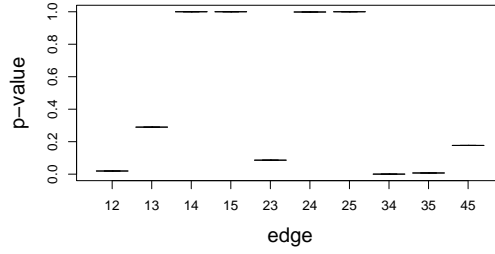


FIGURE 3.8. School Grades.

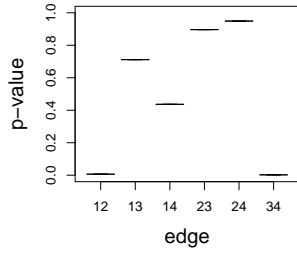
Ex.3.1: Anxiety and Anger.



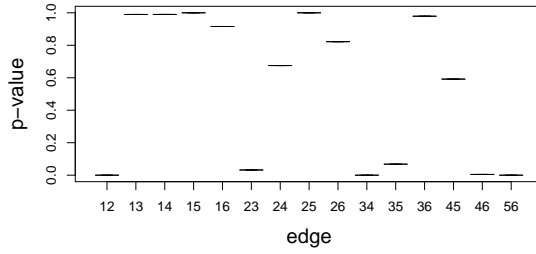
Ex.3.5: Math Marks.



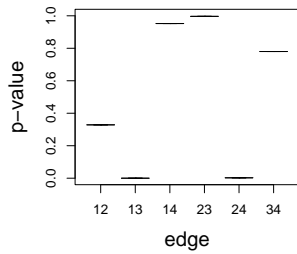
Ex.3.2: Corkborings.



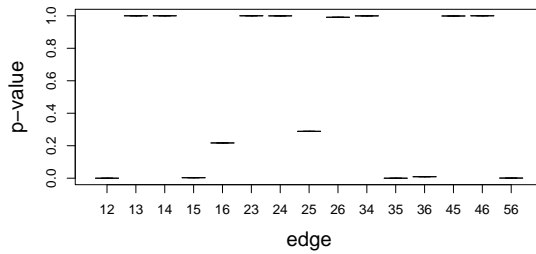
Ex.3.6: Fowl Bones.



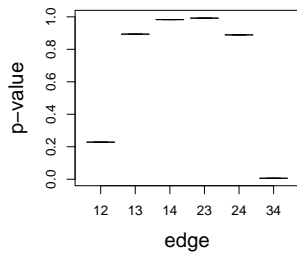
Ex.3.3: Fisher's Iris.



Ex.3.7: HIV.



Ex.3.4: Fret's Heads.



Ex.3.8: School Grades.

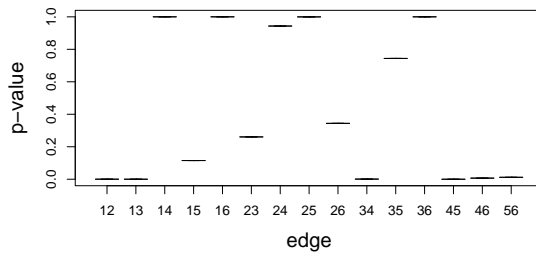


FIGURE 3.9. Simultaneous p-values in the considered examples.

4. CHOOSING THE SIGNIFICANCE LEVEL—THEORY AND SIMULATIONS

It follows from (2.6) that $b_p(\alpha)$ increases from 0 to ∞ as α decreases from 1 to 0. Thus for $\alpha \approx 1$, the simultaneous intervals (2.7) are very narrow, so all edges are deemed significant and retained, that is, $\hat{G}_\alpha = \bar{G}$. As α decreases more and more edges are excluded from \hat{G}_α , until for $\alpha \approx 0$, $\hat{G}_\alpha = G_\emptyset$. Hopefully, $\hat{G}_\alpha = G$ for some nondegenerate interval $\alpha^* < \alpha < \alpha^{**}$, which may be estimated or approximated.

Precisely, if we define

$$(4.1) \quad \alpha^* = \inf \{ \alpha \mid |z^{ij}| > n_p^{-1/2} b_p(\alpha) \forall ij \in E_1 \} \equiv b_p^{-1}(\sqrt{n_p} L),$$

$$(4.2) \quad \alpha^{**} = \sup \{ \alpha \mid |z^{ij}| \leq n_p^{-1/2} b_p(\alpha) \forall ij \in E_0 \} \equiv b_p^{-1}(\sqrt{n_p} \ell),$$

where $L = \min \{ |z^{ij}| \mid ij \in E_1 \}$ and $\ell = \max \{ |z^{ij}| \mid ij \in E_0 \}$, then $\hat{G}_\alpha = G$ iff

- (i) $\alpha^* < \alpha^{**}$ (equivalently, $\ell < L$), and
- (ii) $\alpha \in (\alpha^*, \alpha^{**}]$.

From (2.6) and (2.9),

$$(4.3) \quad \alpha^* = 1 - [2\Phi(\sqrt{n_p} L) - 1]^{p(p-1)/2} \equiv \pi(L),$$

$$(4.4) \quad \alpha^{**} = 1 - [2\Phi(\sqrt{n_p} \ell) - 1]^{p(p-1)/2} \equiv \pi(\ell).$$

Because E_1 and E_0 are determined by the true model G , they are unknown, so L and ℓ are unobservable. However, $L \xrightarrow{a.s.} \Lambda > 0$ as $n \rightarrow \infty$ (since $\{z^{ij}\}$ is the MLE of $\{\zeta^{ij}\}$) so $\alpha^* \rightarrow 0$. Furthermore, it is possible to obtain an approximate upper bound for ℓ and thereby an approximate lower bound for α^{**} , as follows.

By (2.2) and Šidák's inequality,

$$(4.5) \quad \sqrt{n_p} \ell \stackrel{\text{st}}{\leq} \max \{ |u^{ij}| \mid ij \in E_0 \}$$

$$(4.6) \quad \leq \max \{ |u^{ij}| \mid 1 \leq i < j \leq p \} \equiv u$$

(to the accuracy of the normal approximations), where the $\{u^{ij}\}$ are i.i.d. $\mathcal{N}(0, 1)$ variates. (We remark that (4.6) is an approximate equality if G is a sparse graph.) Because $u \approx (2 \ln[p(p-1)/2])^{1/2}$ even for moderately large values of p (compare [16, Theorem 1.5.3]), we have that

$$(4.7) \quad \alpha^{**} \stackrel{\text{st}}{\geq} 1 - [2\Phi((2 \ln[p(p-1)/2])^{1/2}) - 1]^{p(p-1)/2}$$

$$(4.8) \quad \equiv \alpha(p).$$

Note that $\alpha(p)$ does not depend on n , but we expect the lower bound $\alpha^{**} > \alpha(p)$ to be sharpest for large values of n

Table 9 lists the values of $\alpha(p)$ for $p = 3, \dots, 10, 20, 50$. For small or moderate p , these values are in rough agreement with our rule-of-thumb to use 0.25 as the upper indeterminate p-value for the class I for large n and to use 0.4 for smaller n (see §3), but also suggest that the p-values 0.25 and 0.4 should be reduced for larger p .

TABLE 9. Values of $\alpha(p)$ for $p = 3, \dots, 10$.

p	3	4	5	6	7	8	9	10	20	50
$\alpha(p)$	0.36	0.30	0.28	0.26	0.25	0.24	0.24	0.23	0.20	0.18

In order to judge how well our procedure is able to detect the presence or absence of edges in a concentration graph model $N(G)$, we selected the two graphs G_3 ($p = 3$) and G_4 ($p = 4$) shown in Figure 4.1. Then we simulated samples of size $n_1 \equiv n + 1$ from the

FIGURE 4.1. The graphs G_3 and G_4 .

distributions $\mathcal{N}_3(\mu_3, \Sigma_3) \in N(G_3)$ and $\mathcal{N}_4(\mu_4, \Sigma_4) \in N(G_4)$, respectively. The selected mean vectors are zero, i.e. $\mu_3 = 0$ and $\mu_4 = 0$, and the selected covariance matrices Σ_3 and Σ_4 are determined from the corresponding concentration matrices shown in (4.9), where σ^{12} ranges over 0.2, 0.3, 0.4, and 0.5:

$$(4.9) \quad \Sigma_3^{-1} = \begin{pmatrix} 1 & \sigma^{12} & 0 \\ \sigma^{12} & 1 & 0.55 \\ 0 & 0.55 & 1 \end{pmatrix}, \quad \Sigma_4^{-1} = \begin{pmatrix} 1 & \sigma^{12} & 0 & -0.6 \\ \sigma^{12} & 1 & 0.65 & 0 \\ 0 & 0.65 & 1 & 0.55 \\ -0.6 & 0 & 0.55 & 1 \end{pmatrix}.$$

Note that by (1.3) the off-diagonal entries of Σ_3^{-1} and Σ_4^{-1} are also the negated partial correlations.

The results from 10,000 simulations are presented in Figures 4.2 and 4.3. The box-plots indicate the (marginal) distributions of the $p(p-1)/2$ simultaneous p-values $\{\pi^{ij}\}$ in (2.9). The figures for both $G = G_3$ and $G = G_4$ show that as the sample size increases a clear separation appears between the p-values for edges that are present in G and the p-values for edges that are absent in G . Furthermore, the separation is more pronounced for larger values of $\sigma^{12} = -\rho^{12}$. The SIN procedure is based on this separation, hence will successfully detect the true model with high probability in these examples.

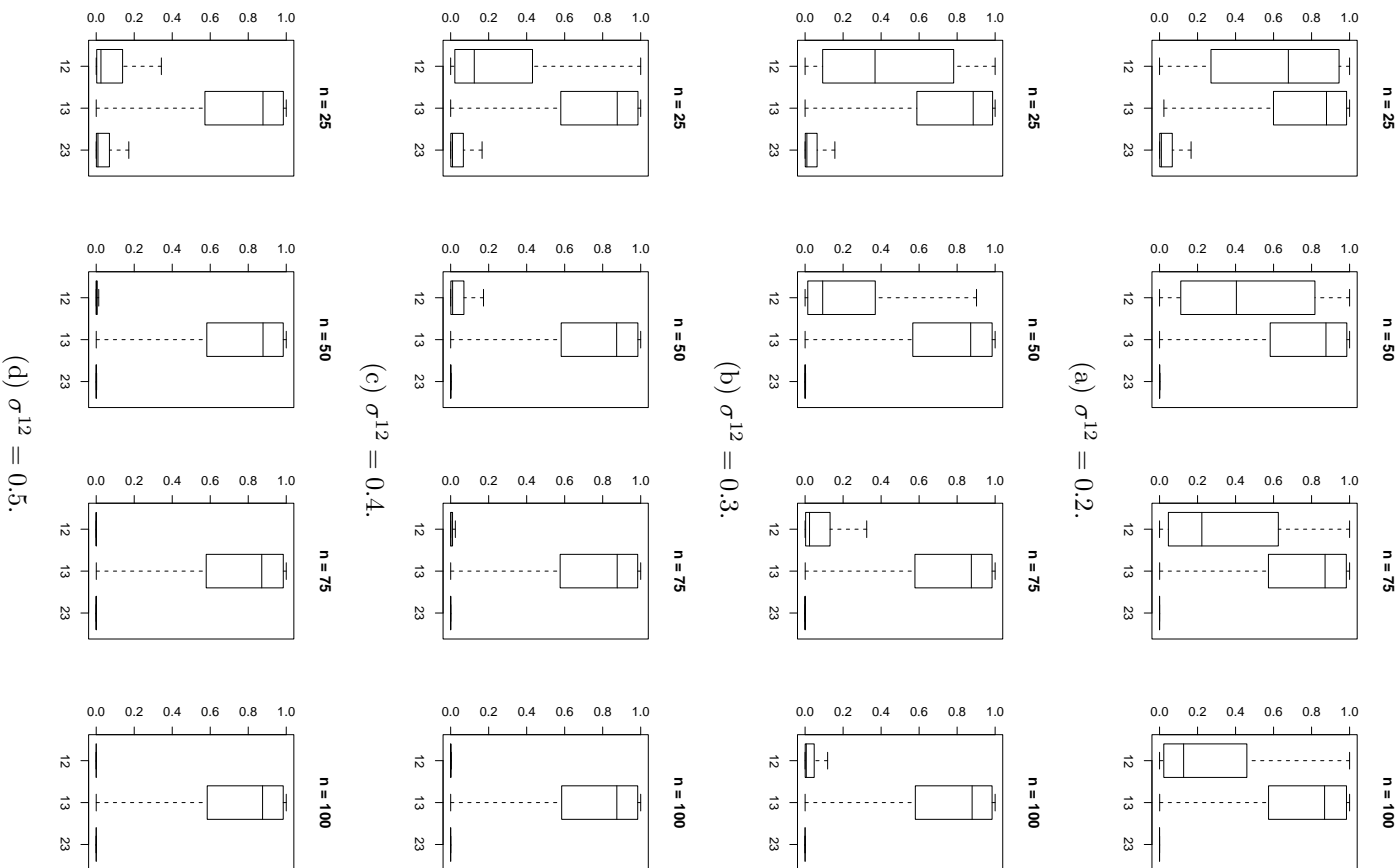
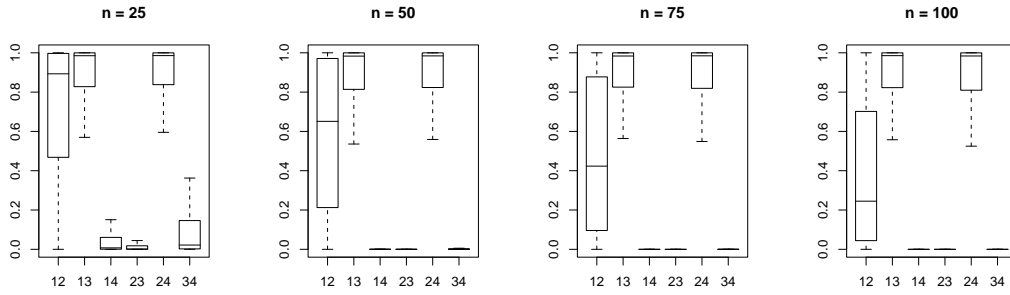
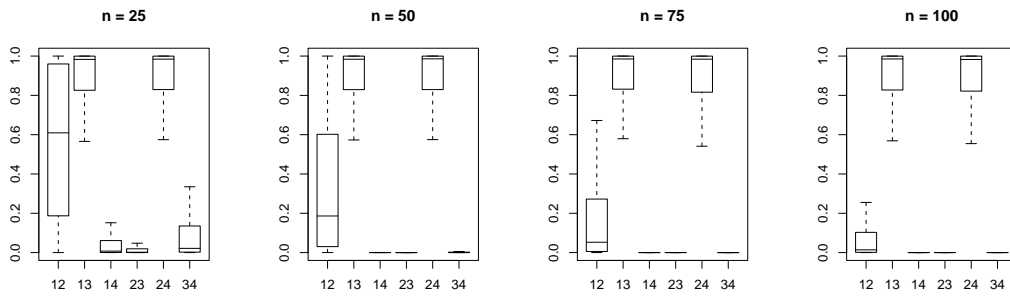


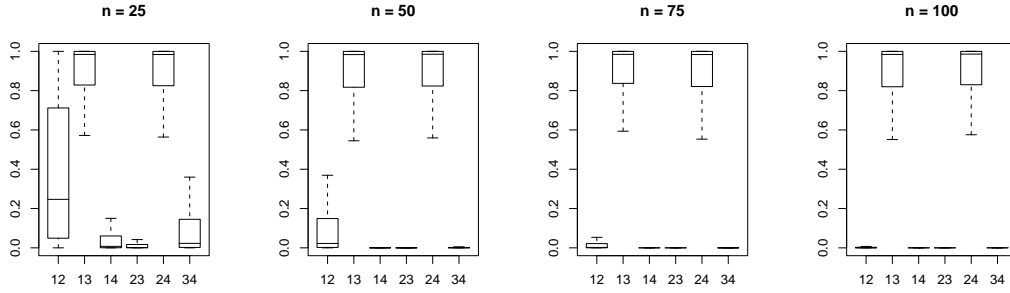
FIGURE 4.2. Boxplots of p-values from 10,000 simulations from $\mathcal{N}_3(\mu_3, \Sigma_3)$.



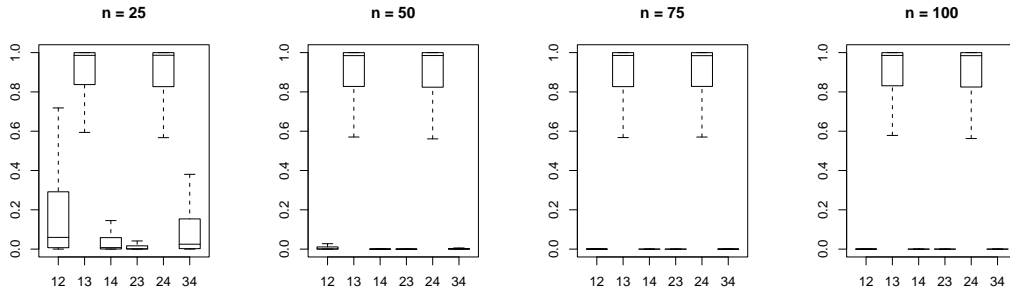
(a) $\sigma^{12} = 0.2$.



(b) $\sigma^{12} = 0.3$.



(c) $\sigma^{12} = 0.4$.



(d) $\sigma^{12} = 0.5$.

FIGURE 4.3. Boxplots of p-values from 10,000 simulations from $\mathcal{N}_4(\mu_4, \Sigma_4)$.

5. INCORPORATING PRIOR INFORMATION ABOUT EDGES

Suppose it is known that certain edges E_0^d of the true graph $G \equiv (V, E)$ are definitely absent and that certain other edges E_1^d are definitely present. More precisely, we assume that in the data-generating distribution $\sigma^{ij} = 0$ for all $ij \in E_0^d$ and $\sigma^{ij} > 0$ for all $ij \in E_1^d$. Thus, $E = E_0^d \dot{\cup} E_1^d \dot{\cup} E^u$, 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 $\bar{G}^u \equiv (V, E_0^d \dot{\cup} E_1^d \dot{\cup} \bar{E}^u)$ denote the *upper graph*, where \bar{E}^u replaces all uncertain edges by present edges, and let $G_\emptyset^u \equiv (V, E_0^d \dot{\cup} E_1^d \dot{\cup} E_\emptyset^u)$ denote the *lower graph*, where E_\emptyset^u replaces all uncertain edges by absent edges. Thus the true graph G satisfies $G_\emptyset^u \subseteq G \subseteq \bar{G}^u$, where G_\emptyset^u and \bar{G}^u are known. (This notation agrees with that in the preceding sections, where the upper and lower graphs are the complete graph \bar{G} and the empty graph G_\emptyset , respectively. Consonni and Leucari [3] call the upper graph the “full graph”, although in the context of directed, rather than undirected, graphs.)

The methodology and SIN approach presented in Sections 2 and 3 extend readily to the present case with only minor modifications. First, the $p(p-1)/2$ simultaneous testing problems in (1.5) are reduced to the $q \equiv |E^u|$ testing problems corresponding to the uncertain edges only. By the local Markov property for \bar{G}^u and the relation $G \subseteq \bar{G}^u$, for each uncertain edge $ij \in E^u$ the partial correlation ρ^{ij} is equal to the conditional correlation $\bar{\rho}^{ij}$ between Y_i and Y_j given $Y_{\text{nb}_{\bar{G}^u}\{i,j\}}$, where $\text{nb}_{\bar{G}^u}\{i,j\}$ denotes the set of neighbors of $\{i,j\}$ in \bar{G}^u , so the hypothesis H^{ij} in (1.5) can be tested by means of the corresponding sample conditional correlation \bar{r}^{ij} , or equivalently by its sample z -transform \bar{z}^{ij} . The approximation (2.2) is now replaced by

$$(5.1) \quad \sqrt{\bar{n}^{ij}} (\bar{z}^{ij} - \bar{\zeta}^{ij}) \rightarrow \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty,$$

where $\bar{n}^{ij} = n - |\text{nb}_{\bar{G}^u}\{i,j\}| - 2$ and $\bar{\zeta}^{ij}$ is the z -transform of $\bar{\rho}^{ij}$. Šidák’s inequality now yields that

$$(5.2) \quad \Pr_G[|\bar{z}^{ij} - \bar{\zeta}^{ij}| \leq (\bar{n}^{ij})^{-1/2} \bar{b}_q(\alpha) \forall ij \in E^u] \geq 1 - \alpha$$

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

$$(5.3) \quad [2\Phi(\bar{b}_q(\alpha)) - 1]^q = 1 - \alpha,$$

or equivalently by

$$(5.4) \quad \bar{b}_q(\alpha) = \Phi^{-1} \left(\frac{1}{2} [(1 - \alpha)^{1/q} + 1] \right),$$

From (5.2) we obtain the following set of conservative simultaneous $1 - \alpha$ confidence intervals for $\{\bar{\zeta}^{ij} \mid ij \in E^u\}$:

$$(5.5) \quad \bar{z}^{ij} - (\bar{n}^{ij})^{-1/2} \bar{b}_q(\alpha) \leq \bar{\zeta}^{ij} \leq \bar{z}^{ij} + (\bar{n}^{ij})^{-1/2} \bar{b}_q(\alpha), \quad ij \in E^u.$$

Because $\bar{n}^{ij} \geq n_p$ and $\bar{b}_q(\alpha) \leq b_p(\alpha)$, these intervals are narrower than those in (2.7) and the corresponding tests are more powerful.

As before, these intervals can be used for model selection. From the simultaneous intervals (5.5) we obtain the set of simultaneous p-values $\{\bar{\pi}^{ij} \equiv \bar{\pi}(\bar{z}^{ij}) \mid ij \in E^u\}$ for the q testing problems $\{H^{ij} \mid ij \in E^u\}$, namely,

$$(5.6) \quad \bar{\pi}(\bar{z}^{ij}) = 1 - [2\Phi(\sqrt{\bar{n}^{ij}}|\bar{z}^{ij}|) - 1]^q.$$

The selected graph $\hat{G}_\alpha \equiv (V, E_0^d \dot{\cup} E_1^d \dot{\cup} \hat{E}_\alpha)$, where $\hat{E}_\alpha \equiv \{\hat{e}_\alpha^{ij} \mid ij \in E^u\}$, is now given by

$$(5.7) \quad \hat{e}_\alpha^{ij} = \begin{cases} 0 & \text{if } \bar{\pi}^{ij} \geq \alpha, \\ 1 & \text{if } \bar{\pi}^{ij} < \alpha. \end{cases}$$

The results (2.13), (2.17), and (2.18) concerning the overall error rate of the selection procedure \hat{G}_α remain valid here. Also, the approximate bound (2.15) can be sharpened as follows:

$$(5.8) \quad \Pr_G[\hat{G}_\alpha \supseteq G] \geq 1 - |E_1^u| \cdot \Pr[\chi_1^2(\bar{n}\bar{\Lambda}^2) \leq \bar{b}_q^2(\alpha)],$$

where E_1^u denotes the set of edges in E^u that are actually present in G , $\bar{n} = \min\{\bar{n}^{ij} \mid ij \in E_1^u\}$, and

$$(5.9) \quad \bar{\Lambda} \equiv \bar{\Lambda}(G, \Sigma) = \min\{|\bar{\zeta}^{ij}| \mid ij \in E_1^u\}.$$

Note that $\bar{\Lambda} > 0$ iff the data-generating distribution is faithful to G . Furthermore, if $\bar{\Lambda} > 0$ is *known or specified* (equivalently, if $\min\{|\bar{\rho}^{ij}| \mid ij \in E_1^u\}$ is known or specified), then as in (2.19), (2.21), and (2.22), $\Pr_G[\hat{G}_\alpha \neq G] \leq \alpha + \beta$ for arbitrarily small α, β provided that

$$(5.10) \quad n \geq \bar{m} + \frac{1}{\bar{\Lambda}^2} \left[\Phi^{-1} \left(\frac{1}{2} [(1 - \alpha)^{1/q} + 1] \right) + \Phi^{-1} \left(1 - \frac{\beta}{q} \right) \right]^2$$

$$(5.11) \quad \approx \bar{m} + \frac{1}{\bar{\Lambda}^2} \left[\sqrt{2 \ln \left(\frac{2q}{\alpha} \right)} + \sqrt{2 \ln \left(\frac{q}{\beta} \right)} \right]^2$$

where $\bar{m} = \max\{|\text{nb}_{\bar{G}^u}\{i, j\}| \mid ij \in E_1^u\} + 2$. (This sample size requirement is less stringent than that given by (2.21) and (2.22) for two reasons. First, the prior edge information guarantees that $q < p(p-1)/2$. Second, $\bar{m} \leq p$ with strict inequality if the true G is a sparse graph.)

The SIN approach to model selection is applied here by simply replacing the entire set of p-values $\{\pi^{ij} \mid 1 \leq i < j \leq p\}$ by the reduced set $\{\bar{\pi}^{ij} \mid ij \in E^u\}$, obtaining two selected graphs \hat{G}_S and \hat{G}_{SI} as in Section 3.

Lastly, the discussion at the beginning of Section 4 also remains valid with minor modifications, most notably, $p(p-1)/2$ should be replaced by q in the stochastic lower bound (4.7) for α^{**} .

6. MODEL SELECTION FOR COVARIANCE GRAPHS

Let $G^{bi} \equiv (V, E^{bi})$ be a *bidirected* graph with vertex set $V \equiv \{1, \dots, p\}$ and edge set $E^{bi} \equiv \{e_{ij}\}$, where $e_{ij} = 1$ (resp., 0) indicates the presence (resp., absence) of a bidirected edge $i \leftrightarrow j$ between vertices i and j ($1 \leq i < j \leq p$). The *covariance graph model* $N(G^{bi})$ is defined as the family of all p -variate normal distributions $\mathcal{N}_p(\mu, \Sigma)$ with unknown mean vector μ and unknown covariance matrix $\Sigma = \{\sigma_{ij}\}$ (assumed nonsingular) that satisfies the linear restrictions:

$$(6.1) \quad e_{ij} = 0 \quad \implies \quad \sigma_{ij} = 0 \quad (\iff \rho_{ij} = 0).$$

Here, ρ_{ij} denotes the ij -th (ordinary) correlation. Since $\sigma_{ij} = 0$ iff $Y_i \perp\!\!\!\perp Y_j$, the absence of an edge between two vertices denotes marginal independence of the corresponding variables.

The name ‘‘covariance graph’’ was introduced by Cox and Wermuth [4, 5], who used dashed lines instead of bidirected edges. These models correspond to the special case of Richardson and Spirtes’ [19] *ancestral graph* models where the graph has bidirected edges only. We adopt Richardson and Spirtes’ use of bidirected edges, since this choice allows the global Markov properties of the graph to be readily expressed in terms of its pathwise separation properties.

Inference and model selection for graphical models for marginal independence have not been developed as thoroughly as for graphical models for conditional independence. In particular, algorithms for maximum likelihood (ML) estimation are currently under development (see Drton and Richardson [9]) and are not yet implemented in software packages; e.g no methods for maximum likelihood (ML) estimation are available in Edwards’ [10] MIM package. However, Kauermann [13] developed a heuristic inference method based on a ‘‘dual likelihood’’, which can be carried out in MIM as described in Edwards [10, §7.4]. In this method, the inverse sample covariance matrix is treated as if it were the sample covariance matrix and a Gaussian concentration graph model fitted to the altered sufficient statistics. Thus, one can perform a backward stepwise model selection by this dualization. Note, however, that, in general, the dual likelihood approach and direct ML estimation will not give the same results. In particular, the actual likelihood of a covariance graph model can be multimodal whereas the dual likelihood is always unimodal. (For an example of such a multimodal likelihood see Drton and Richardson [8].)

A simple way to bypass these complications for covariance graph model selection, i.e. determination of the graph G^{bi} , is suggested by (6.1), namely, to apply the methods developed in Section 2 to the sample correlations $\{r_{ij}\}$ rather than to the sample partial correlations $\{r^{ij}\}$. Denote the z -transforms of r_{ij} and ρ_{ij} by z_{ij} and ζ_{ij} , respectively. Then as shown by Larntz and Perlman [14, p.297], the following are conservative

simultaneous $1 - \alpha$ confidence intervals for $\{\zeta_{ij}\}$:

$$(6.2) \quad z_{ij} - (n - 2)^{-1/2}b_p(\alpha) \leq \zeta_{ij} \leq z_{ij} + (n - 2)^{-1/2}b_p(\alpha) \quad (1 \leq i < j \leq p),$$

where $b_p(\alpha)$ is defined by (2.5) and (2.6). Recall that $\zeta_{ij} = 0$ ($\neq 0$) iff $\rho_{ij} = 0$ ($\neq 0$).

Using these intervals, our proposed model selection procedure for covariance graph models proceeds in the same way as that described in Sections 2 and 3. From (6.2), a set of conservative simultaneous p-values $\{\pi_{ij}\}$ is obtained for the $p(p - 1)/2$ testing problems

$$(6.3) \quad H_{ij} : \rho_{ij} = 0 \quad \text{vs.} \quad K_{ij} : \rho_{ij} \neq 0 \quad (1 \leq i < j \leq p),$$

namely,

$$(6.4) \quad \pi_{ij} = 1 - [2\Phi(\sqrt{n - 2}|z_{ij}|) - 1]^{p(p-1)/2},$$

the only change from (2.9) being that the sample size adjustment is $n - 2$ rather than $n_p \equiv n - p$. The selected graph $\hat{G}_\alpha^{bi} = (V, \hat{E}_\alpha^{bi})$ is then given by

$$(6.5) \quad \hat{e}_{ij}^\alpha = \begin{cases} 0 & \text{if } \pi_{ij} \geq \alpha, \\ 1 & \text{if } \pi_{ij} < \alpha, \end{cases}$$

where $\hat{E}_\alpha^{bi} \equiv \{\hat{e}_{ij}^\alpha\}$ is the edge set of the selected graph. Thus, the edge $i \leftrightarrow j$ is included in \hat{G}_α^{bi} iff π_{ij} is significantly small at overall level α .

It is easy to show that \hat{G}_α^{bi} satisfies consistency properties analogous to those developed above for \hat{G}_α . Note that faithfulness in covariance graph models consists of the equivalence $\sigma_{ij} = 0$ iff $e_{ij} = 0$. For applications we again advocate the SIN approach, whereby the simultaneous p-values $\{\pi_{ij}\}$ are partitioned into groups S, I, and N, leading to two selected models \hat{G}_{SI}^{bi} and \hat{G}_S^{bi} , the latter being more conservative with respect to edge inclusion.

We illustrate the SIN approach to covariance graph model selection by the following example.

Example 6.1 (Stressful Events). Cox and Wermuth [4, Example 3] report data from a study on stressful events of $n_1 = 72$ students. The $p = 4$ measured variables are “cognitive avoidance”, “vigilance”, “blunting”, and “monitoring”, labeled as 1 to 4.

Backward model selection in MIM using Kauermann’s [13] dual likelihood approach yields, at the individual significance level $\alpha = 0.05$, the graph depicted in Figure 6.1(a). The simultaneous p-values π_{ij} given by (6.4) are listed in Table 10 and depicted in Figure 6.2. Clearly $S = \{0.00^{(2)}\}$ corresponding to including the (bidirected) edges 13 and 24, $N = \{0.76, 1.00^{(2)}\}$ corresponding to excluding edges 34, 14, and 23, and $I = \{0.44\}$ corresponding to the possible (bidirected) edge 12 (although for a sample of size 72, a simultaneous p-value as large as 0.44 is probably better assigned to the non-significant class N.) Hence, SIN selects $\hat{G}_\alpha^{bi} = \hat{G}_{SI}^{bi}$ for $\alpha \in (0.44, 0.76)$, as illustrated in Figure 6.1(a), and, here more appropriately conservative, $\hat{G}_\alpha^{bi} = \hat{G}_S^{bi}$ for $\alpha \in (0.00, 0.44)$, as illustrated in

TABLE 10. Simultaneous p-values π_{ij} for Stressful Events.

	cog av (1)	vigil (2)	blunt (3)	monit (4)
cogn.avoidance (1)				
vigilance (2)	0.44			
blunting (3)	0.00	1.00		
monitoring (4)	1.00	0.00	0.76	

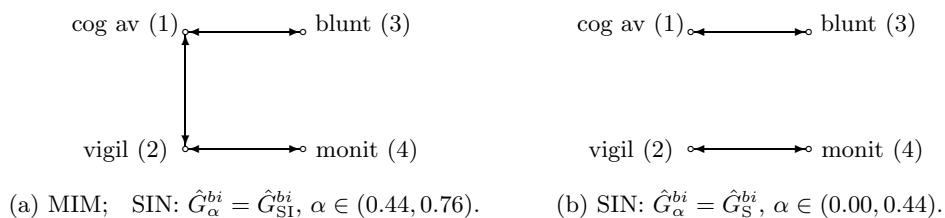


FIGURE 6.1. Stressful Events.

Figure 6.1(b). The latter model entails the marginal independence (cognitive avoidance, blunting) $\perp\!\!\!\perp$ (vigilance, monitoring).

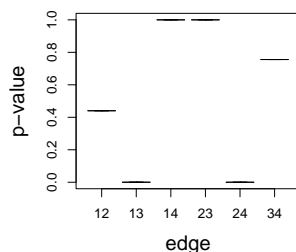


FIGURE 6.2. Simultaneous p-values in Ex.6.1 (Stressful Events).

As in Section 5, prior edge information also can be incorporated into covariance graph model selection. Here we gain the advantage of reducing the $p(p-1)/2$ simultaneous edge tests to q tests only, but it is no longer apparent how/whether to alter the original sample correlations to take advantage of the smaller boundary of $\{i, j\}$ in the upper graph \tilde{G}^u .

7. CONCLUDING REMARKS

We have introduced a new method for model selection in Gaussian concentration graph models and covariance graph models based on the partial and ordinary correlations, respectively. Applying the variance-stabilizing z -transform and using a Šidák's inequality

we find a conservative rectangular simultaneous confidence region for the (partial) correlations. The acceptance or rejection of simultaneous hypotheses of zero (partial) correlations leads to the exclusion or inclusion of the corresponding edge in the selected model. This provides a simple, one-step procedure for selecting a model. Since our method is based on a simultaneous test it avoids problems of multiple testing and permits the control of the overall error in model selection. Our method tends to select sparser (more parsimonious) models than standard procedures such as backward stepwise selection. These parsimonious models are often more easily interpretable.

In the case of Gaussian covariance graph models, our method is attractive from a technical point of view since inference for these models is not as fully developed as inference for concentration graph models. In particular, if the selected covariance graph model does not exhibit a multimodal likelihood, our method avoids problems with the possibly multimodal likelihood of covariance graph models altogether. Stepwise selection procedures, however, might have to go through the fit of inappropriate models which in turn might cause problems with a multimodal likelihood (compare e.g. Drton and Richardson [8] who find a higher frequency of a multimodal likelihood if data is simulated from a misspecified model).

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

REFERENCES

1. T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 1984.
2. S.P. Brooks, P. Giudici, and Roberts G.O., *Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions*, J. Royal Stat. Soc. B **65** (2003), 1–37.
3. G. Consonni and V. Leucari, *Model determination for directed acyclic graphs*, The Statistician **50** (2001), no. 3, 243–256.
4. D.R. Cox and N. Wermuth, *Linear dependencies represented by chain graphs*, Statist. Sci. **8** (1993), no. 3, 204–218, 247–283.
5. ———, *Multivariate Dependencies*, Chapman & Hall, London, 1996.
6. P. Dellaportas, P. Giudici, and G.O. Roberts, *Bayesian inference for nondecomposable graphical Gaussian models*, Sankhyā **65** (2003), 43–55.
7. A.P. Dempster, *Covariance selection*, Biometrics **28** (1972), 157–175.
8. M. Drton and T.S. Richardson, *Multimodality of the likelihood in the bivariate seemingly unrelated regression model*, Tech. Report 410, Department of Statistics, University of Washington, 2002.
9. ———, *A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence*, Uncertainty in Artificial Intelligence: Proceedings of the 19th Conference (Uffe Kjærulff and Christopher Meek, eds.), Morgan Kaufmann, 2003, pp. 184–191.
10. D.M. Edwards, *Introduction to Graphical Modelling*, second ed., Springer-Verlag, New York, 2000.
11. P. Giudici, *Learning in graphical Gaussian models*, Bayesian Statistics, 5 (Alicante, 1994), Oxford Univ. Press, New York, 1996, pp. 621–628.

12. P. Giudici and P.J. Green, *Decomposable graphical Gaussian model determination*, *Biometrika* **86** (1999), 785–801.
13. G. Kauermann, *On a dualization of graphical Gaussian models*, *Scand. J. Statist.* **23** (1996), 105–116.
14. K. Larntz and M.D. Perlman, *A simple test for the equality of correlation matrices*, *Statistical decision theory and related topics, IV, Vol. 2* (West Lafayette, Ind., 1986), Springer, New York, 1988, pp. 289–298.
15. S.L. Lauritzen, *Graphical Models*, Oxford University Press, New York, 1996.
16. M.R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York, 1983.
17. K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
18. C.R. Rao, *Tests of significance in multivariate analysis*, *Biometrika* **35** (1948), 58–79.
19. T.S. Richardson and P. Spirtes, *Ancestral graph Markov models*, *Ann. Statist.* **30** (2002), 962–1030.
20. A. Roverato, *Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models*, *Scand. J. Statist.* **29** (2002), 391–411.
21. A. Roverato and J. Whittaker, *A hyper normal prior distribution for approximate Bayes factor calculations on non-decomposable graphical Gaussian models*, unpublished manuscript, 1996.
22. ———, *Standard errors for the parameters of graphical Gaussian models*, *Statistics and Computing* **6** (1996), 297–302.
23. G.R. Shorack, *Probability for Statisticians*, Springer-Verlag, New York, 2000.
24. Z. Šidák, *Rectangular confidence regions for the means of multivariate normal distributions*, *J. Amer. Statist. Assoc.* **62** (1967), 626–633.
25. J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons Ltd., Chichester, 1990.

DEPARTMENT OF STATISTICS, UNIVERSITY OF WASHINGTON, SEATTLE, WASHINGTON, U.S.A.
E-mail address: drton@stat.washington.edu, michael@ms.washington.edu