Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 1

Low-Order Conditional Independence Graphs for Inferring Genetic Networks

Anja Wille, *ETH Zurich*
Peter Bühlmann, *ETH Zurich*

Recommended Citation:
Wille, Anja and Bühlmann, Peter (2006) "Low-Order Conditional Independence Graphs for Inferring Genetic Networks," *Statistical Applications in Genetics and Molecular Biology*: Vol. 5: Iss. 1, Article 1.
DOI: 10.2202/1544-6115.1170

Low-Order Conditional Independence Graphs for Inferring Genetic Networks

Anja Wille and Peter Bühlmann

Abstract

As a powerful tool for analyzing full conditional (in-)dependencies between random variables, graphical models have become increasingly popular to infer genetic networks based on gene expression data. However, full (unconstrained) conditional relationships between random variables can be only estimated accurately if the number of observations is relatively large in comparison to the number of variables, which is usually not fulfilled for high-throughput genomic data.

Recently, simplified graphical modeling approaches have been proposed to determine dependencies between gene expression profiles. For sparse graphical models such as genetic networks, it is assumed that the zero- and first-order conditional independencies still reflect reasonably well the full conditional independence structure between variables. Moreover, low-order conditional independencies have the advantage that they can be accurately estimated even when having only a small number of observations. Therefore, using only zero- and first-order conditional dependencies to infer the complete graphical model can be very useful. Here, we analyze the statistical and probabilistic properties of these low-order conditional independence graphs (called 0-1 graphs). We find that for faithful graphical models, the 0-1 graph contains at least all edges of the full conditional independence graph (concentration graph). For simple structures such as Markov trees, the 0-1 graph even coincides with the concentration graph. Furthermore, we present some asymptotic results and we demonstrate in a simulation study that despite their simplicity, 0-1 graphs are generally good estimators of sparse graphical models. Finally, the biological relevance of some applications is summarized.

KEYWORDS: Graphical modeling, Gene expression

Author Notes: We would like to thank the reviewers for their helpful comments.

Introduction

Graphical models (Edwards, 2000; Lauritzen, 1996) form a probabilistic tool to analyze and visualize conditional dependence between random variables. Random variables are represented by vertices of a graph and conditional relationships between them are encoded by edges. Based on graph theoretical concepts and algorithms, the multivariate distribution can be often decomposed into simpler distributions which facilitates the detection of direct and indirect relationships between variables.

Due to this property, graphical models have become increasingly popular for inferring genetic regulatory networks based on the conditional dependence structure of gene expression levels (Wang *et al.*, 2003; Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Toh & Horimoto, 2002). However, when analyzing genetic regulatory associations from high-throughput biological data such as gene expression data, the activity of thousands of genes is monitored over relatively few samples. Since the number of variables (genes) largely exceeds the number of observations (chip experiments), inference of the dependence structure is rendered difficult due to computational complexity and inaccurate estimation of high-order conditional dependencies. With an increasing number of variables, only a small subset of the super-exponentially growing number of models can be tested (Wang *et al.*, 2003). More importantly, an inaccurate estimation of conditional dependencies leads to a high rate of false positive and false negative edges. An interpretation of the graph within the Markov property framework (Edwards, 2000; Lauritzen, 1996) is then rather difficult (Husmeier, 2003; Waddell & Kishino, 2000).

These problems may be circumvented using a simpler approach with better estimation properties to characterize the dependence structure between random variables. The simplest method would be to model the marginal dependence structure in a so called covariance graph (Cox & Wermuth, 1993, 1996). The covariance structure of random variables can be accurately estimated and easily interpreted even with a large number of variables and a small sample size. However, the covariance graph contains only limited information since the effect of other variables on the relationship between two variables is ignored.

As a simple yet powerful approach to balance between the independence and covariance graph, zero- and first-order conditional independence graphs have recently gained attention to model genetic networks (Wille *et al.*, 2004; Magwene & Kim, 2004; de la Fuente *et al.*, 2004). Instead of conditioning on all variables at a time, only zero- and first-order conditional dependence relationships are combined for inference on the complete graph. This allows to

study dependence patterns in a more complex and exhaustive way than with only pairwise correlation-based relationships while maintaining high accuracy even for few observations. We here use the notation 0-1 graphs from de Campos & Huete (2000).

In the three aforementioned studies, it has been shown that 0-1 graphs can be quite powerful to discover genetic associations. However, the probability and estimation properties of 0-1 graphs as an alternative to full conditional independence graphs (concentration graphs) have not been studied so far. Here, we demonstrate the usefulness of 0-1 graphs to discover conditional dependence patterns in settings with many variables and few observations. Following the recent studies, we focus on concentration graphs with continuous data, the so called graphical Gaussian models. In the next sections, we first review graphical Gaussian models, covariance graphs and 0-1 graphs before we analyze the estimation properties of 0-1 graphs in comparison with graphical Gaussian models. As our main interest is to apply our approach in gene expression profiling, we study simulated networks with genetic and metabolic topologies, and discuss the biological relevance of the examples presented in Wille *et al.* (2004) and Magwene & Kim (2004).

Graphical Gaussian models

Consider p random variables $X_1, \ldots, X_p$ which we sometimes denote by the random vector $\mathbf{X} = (X_1, \ldots, X_p)$. Full conditional dependence between two variables X_i and X_j refers to the conditional dependence between X_i and X_j given all other variables $X_k, k \in \{1, \ldots, p\} \setminus \{i, j\}$. Conditional independence between X_i and X_j denoted by $X_i \perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$ states that there is no direct relationship between X_i and X_j.

$$X_2 \perp\!\!\!\perp X_3 \mid (X_1, X_4)$$
$$X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$$
$$X_3 \perp\!\!\!\perp X_4 \mid (X_1, X_2)$$

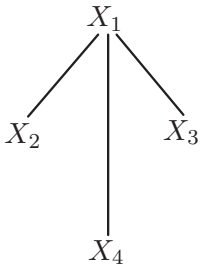

Figure 1: Conditional independence model and associated graph

In graphical modeling, the dependence pattern between variables is associ-

ated with a graph G in which vertices encode the random variables and edges encode conditional dependence between variables. In a concentration graph, two vertices i and j are adjacent if and only if the corresponding variables X_i and X_j are conditionally dependent given all remaining variables. Figure 1 shows an example of the dependence patterns between variables $X_1, \ldots, X_4$ and the corresponding concentration graph. All edges in the graph are undirected.

A set of vertices K is said to separate i and j ($i, j \notin K$) in G if every path between i to j passes through a vertex in K. For random variables $\mathbf{X}$ that follow a multivariate normal distribution, we now have the following definitions (Lauritzen, 1996):

Definition 1 *(**Markov property**)*
A multivariate normal distribution on $\mathbf{X}$ follows the (global) Markov property with respect to G if for all vertices i and j and sets of vertices K ($i, j \notin K$) that separate i and j it holds that $X_i \perp\!\!\!\perp X_j | \{X_k; k \in K\}$.

Definition 2 *(**Faithfulness**)*
A multivariate normal distribution on $\mathbf{X}$ is faithful to G if for all vertices i and j and sets of vertices K ($i, j \notin K$) with $X_i \perp\!\!\!\perp X_j | \{X_k; k \in K\}$ it holds that K separates i and j.

For multivariate normal random variables $\mathbf{X}$ with mean $\mathbb{E}(\mathbf{X}) = \mu$ and covariance matrix $\mathrm{Cov}(\mathbf{X}) = \Sigma$, i.e.

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma),$$

we now give the probabilistic definitions for graphical modeling based on the concentration graph and the covariance graph.

In the concentration graph, an edge between vertex i and j is drawn if and only if X_i and X_j are conditionally dependent given all other variables $\{X_k; k \in \{1, \ldots, p\} \setminus \{i, j\}\}$. Due to the Gaussian assumption, this means that the vertices i and j ($i \neq j$) are adjacent in G if and only if the partial correlation coefficients

$$\omega_{ij} \neq 0, \quad \omega_{ij} = \frac{-\Sigma^{-1}_{ij}}{\sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}} \tag{1}$$

where Σ^{-1}_{ij} are the elements of the inverse covariance matrix (precision or concentration matrix). A family of normal distributions represented by a graph

G is also called a graphical Gaussian model. Graphical Gaussian models follow the Markov property (Lauritzen, 1996) and almost all graphical Gaussian models represented by a graph G are faithful.

To learn the conditional independence structure of the graph, it is necessary to determine which elements of the precision matrix Σ^{-1} are 0. Commonly, this is carried out jointly for all edges in a likelihood approach, where tests for all $2^{p(p-1)/2}$ possible graphical models are conducted to find the best model for the data. For a large number of variables, however, this is hardly feasible so that non-exhaustive search algorithms such as backward and forward selection procedures are used to learn the model (Edwards, 2000). These two selection techniques are the standard modeling procedures although more advanced data adaptive strategies may be applied as well to search through the graph space.

Alternatively, hypothesis testing-based model selection can be pursued in which each edge is tested separately for inclusion ($\frac{p(p-1)}{2}$ hypotheses total). For example, Drton & Perlman (2004) describe an approach where simultaneous conservative confidence intervals are computed for all $\frac{p(p-1)}{2}$ partial correlation coefficients. An edge is included in the model if the corresponding confidence interval does not comprise 0.

In the likelihood-based search, it is necessary to invert the covariance matrix in order to compute the partial correlation coefficients. For the hypothesis-based model selection, confidence intervals increase with larger p and smaller sample size n (Drton & Perlman, 2004) leading to a higher error rate for incorrect edge exclusion. Therefore, both model selection strategies require relative large sample sizes n for a precise estimation of the concentration graph (Lauritzen, 1996, page 128).

For certain applications like genomics, however, such a sample size is typically not available. Concentration graphs learned from such data will then be rather unreliable with a high false positive and high false negative rate. We will show that the much simpler concepts such as the covariance graph and the 0-1 conditional independence graph can be estimated with higher accuracy. However, among the latter two, only the 0-1 graph can capture the more complex conditional independence structure.

In the covariance graph, an edge between vertex i and j ($i \neq j$) is drawn if and only if the correlation coefficient

$$\rho_{ij} \neq 0, \quad \rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{2}$$

The covariance graph as a representation of the marginal dependence structure between variables is simple to interpret and has the advantage that it can

be accurately estimated from finite-sample data even if p is very large in comparison to sample size n, see Proposition 4. However, this graph is often not sufficient to capture more complex conditional dependence patterns.

Zero- and first-order conditional independence graphs

Zero- and first-order conditional independence graphs combine statistical features from the covariance and the concentration graph. In this respect, they can be viewed as striking a balance between the covariance and the concentration graph.

To explore some dependence structure between two variables X_i and X_j, we do not jointly condition on all remaining variables at a time. Instead, we consider separately all pairwise partial correlations

$$\omega_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}$$

of X_i and X_j given one of the remaining variable X_k. These partial correlation coefficients are then combined to draw conclusions on some aspect of the dependence between X_i and X_j.

Definition 3 ***(0-1 conditional independence graph)*** *Draw an edge between vertex i and j ($i \neq j$) if and only if*

$$\rho_{ij} \neq 0 \quad \textit{and} \quad \omega_{ij|k} \neq 0 \textit{ for all } k \in \{1,\ldots,p\} \setminus \{i,j\}.$$

Let $F_{ij} = \rho_{ij} \cup \{\omega_{ij|k}; k \in \{1,\ldots,p\} \setminus \{i,j\}\}$ be the set of the correlation and partial correlation coefficients for X_i and X_j. As parameter ϕ_{ij} for an edge between X_i and X_j, we can use the element of F_{ij} with minimum absolute value. We assign an edge if and only if

$$\phi_{ij} \neq 0, \quad \phi_{ij} = \arg\min_{f \in F_{ij}} (|f|) \tag{3}$$

In general, 0-1 conditional independence graphs are not the same as the concentration graphs. Still, these graphs reflect some measure of conditional dependence. In fact, we can show that for sparse concentration graphs, they can capture the full conditional independence structure well and sometimes even exactly, see Proposition 1 and 2. On the other hand, they are still reasonably simple to interpret. An edge between two variables X_i and X_j represents

$$X_2 \perp\!\!\!\perp X_3 \mid (X_1, X_4)$$
$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$$

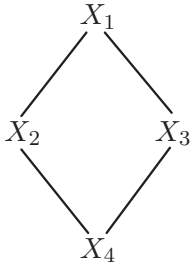

Figure 2: A conditional independence model for which the cyclic concentration graph is contained in the 0-1 graph

a dependence that cannot be explained by any of the other variables X_k. From a statistical perspective, a 0-1 graph can be accurately estimated from data even if p is large relative to sample size n, see Proposition 5 and 6.

Some examples and rigorous properties

We are describing here with some simple examples and two propositions in how far the concentration graph and the 0-1 graph relate to each other.

Example 1: Consider 4 random variables $\mathbf{X} = (X_1, X_2, X_3, X_4) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 2 & 1 & 1 \\ -1 & 1 & 2 & 1 \\ -1 & 1 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \Sigma^{-1} = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Based on the inverted covariance matrix Σ^{-1}, we obtain a conditional independence model as shown in Figure 1. In such a setting, 0-1 graph and concentration graph are exactly the same whereas the covariance graph is the full graph.

Example 2: Consider 4 random variables $\mathbf{X} = (X_1, X_2, X_3, X_4) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 4 & -7 & -5 & 6 \\ -7 & 13 & 9 & -11 \\ -5 & 9 & 7 & -8 \\ 6 & -11 & -8 & 10 \end{pmatrix} \quad \text{and} \quad \Sigma^{-1} = \begin{pmatrix} 5 & 2 & 1 & 0 \\ 2 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

The concentration graph includes all edges except those between the pairs (X_1, X_4) and (X_2, X_3) as shown in Figure 2. From Σ we see that the covariance graph includes all edges. The 0-1 graph also includes all edges since for example, X_2 and X_3 are not conditionally independent on either X_1 or X_4 alone.

Example 3: Consider 4 random variables $\mathbf{X} = (X_1, X_2, X_3, X_4) \sim N(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 4 & -1 & -1 & -1 \\ -1 & 2 & 0 & 0 \\ -1 & 0 & 2 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix} \quad \text{and} \quad \Sigma^{-1} = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.6 & 0.1 \\ 0.2 & 0.1 & 0.1 & 0.6 \end{pmatrix}.$$

Here, the concentration graph includes all edges whereas the 0-1 graph does not contain the edges (X_2,X_3), (X_2,X_4), and (X_3,X_4).

In general, it is difficult to determine to what extent a 0-1 conditional independence graph G_{0-1} represents the structure of the true underlying concentration graph G. However, for faithful concentration graphs, we have the following proposition that the 0-1 conditional independence graph contains all edges of the concentration graph and some more. All proofs are given in the Appendix.

Proposition 1 *If the distribution on* $\mathbf{X}$ *is Gaussian and faithful to the concentration graph* G*, then every edge in* G *is also an edge of the 0-1 graph* G_{0-1}*.*

Furthermore, if $X_i \perp\!\!\!\perp X_j$ and all paths between between i and j lead through a vertex k, we also have $X_i \perp\!\!\!\perp X_j | X_k$ and therefore $\phi_{ij} = 0$. In other words, we have the following proposition:

Proposition 2 *Assume that the distribution on* $\mathbf{X}$ *is Gaussian and let* G *be the corresponding concentration graph. Moreover, assume that if* i *and* j *are not adjacent in* G *then* i *and* j *are either in two different connected components of* G *or there exists a vertex* k *that separates* i *and* j *in* G*. Then, every edge in* G_{0-1} *is also an edge in* G*.*

Due to Proposition 1 and 2, the 0-1 graph and the concentration graph may coincide. In particular, all Gaussian distributions corresponding to a tree are faithful (Becker *et al.*, 2000) so that one obtains:

Corollary 1 *If the concentration graph of a Gaussian distribution is a forest of trees (the graph does not contain any cycles) then the 0-1 graph and the concentration graph coincide.*

0-1 and concentration graphs do also coincide in more complicated scenarios, for example, if the distribution is Gaussian and faithful and if the corresponding concentration graph consists of sets of cliques that (pairwise) share at most one common vertex (Figure 3).

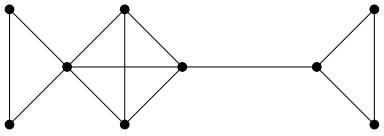

Figure 3: A conditional independence model for which concentration graph and 0-1 graph coincide.

Biological networks such as genetic regulatory networks are sparse. From Propositions 1 and 2, we expect that sparse concentration graphs have fewer edges than the 0-1 conditional independence graph. The number of chordless cycles will be an indicator for the difference between the number of edges in the 0-1 graph and the number of edges in the concentration graph. The larger the number of cycles, the larger the difference in the number of edges.

As distributions are not always faithful (see Example 3), some concentration graphs may also contain more edges than the corresponding 0-1 graph. However, in our simulations for biological networks, this case has only rarely occurred.

Estimation from data

In this section we devise an estimation algorithm for the 0-1 graph and show that it can be accurately estimated even if the number of variables p is large compared to the number of observations n.

In a 0-1 graph, to test whether $\phi_{ij} \neq 0$ (see (3)) for a pair of edges i, j, we first focus on $\omega_{ij|k}$ for all $k \notin \{i, j\}$ and on ρ_{ij}. We can test all null-hypotheses

$$H_0(i,j|k) : \ \omega_{ij|k} = 0 \quad \text{versus} \quad H_1(i,j|k) : \ \omega_{ij|k} \neq 0.$$

with the likelihood ratio test under the Gaussian assumption $X_i, X_j, X_k \sim \mathcal{N}_3(\mu, \Sigma)$. The null hypotheses are $(\Sigma^{-1})_{12} = 0$ (which is equivalent to $\omega_{ij|k} = 0$) and the alternatives are Σ unconstrained. Under the null-hypotheses and

the assumption that the data are i.i.d. realizations from a p-dimensional normal distribution, the log-likelihood ratios are asymptotically χ^2-distributed (Lauritzen, 1996) and every likelihood ratio test of $H_0(i,j|k)$ versus $H_1(i,j|k)$ yields a P-value $P(i,j|k)$. Furthermore, the likelihood ratio test of the null hypothesis for the marginal correlation

$$H_0(i,j|\,\emptyset):\ \rho_{ij}=0 \quad \text{versus} \quad H_1(i,j|\,\emptyset):\ \rho_{ij}\neq 0$$

yields a P-value $P(i,j|\,\emptyset)$.

Recall that an edge in a 0-1 graph between vertex i and j exists if $H_0(i,j|\,\emptyset)$ is rejected and $H_0(i,j|\,k)$ is rejected for all vertices $k \notin \{i,j\}$. Thus, there is evidence for an edge between vertex i and j if

$$\max_{k\in\{\emptyset,1,2,\ldots,p\}\setminus\{i,j\}} P(i,j|\,k) < \alpha,$$

where α is the significance level. For deciding about a single edge between vertices i,j, it is not necessary to correct for the $p-1$ multiple testing over all conditioning vertices k.

Proposition 3 *For some fixed pair (i,j), consider the single hypothesis,*

$H_0(i,j)$*: at least one $H_0(i,j|\,k^*)$ is true for some $k^* \in \{\emptyset,1,2,\ldots,p\}\setminus\{i,j\}$.*

Assume that for all $k \in \{\emptyset,1,2,\ldots,p\}\setminus\{i,j\}$ the individual test satisfies

$$\mathbb{P}_{\tilde{H}_0(i,j|k)}\left[H_0(i,j|\,k)\ \textit{rejected}\ \right] \leq \alpha,$$

where $\tilde{H}_0(i,j|k) = \{H_0(i,j|k)\ \textit{true}\ \} \cap \{H_0(i,j|k')\ \textit{true or false and compatible with}\ H_0(i,j|k)\ \textit{true for all}\ k' \neq k\}$. Then, the type-I error

$$\mathbb{P}_{H_0(i,j)}\left[H_0(i,j|\,k)\ \textit{are rejected for all}\ k \in \{\emptyset,1,2,\ldots,p\}\setminus\{i,j\}\right] \leq \alpha.$$

Note that the log-likelihood ratio test described above satisfies asymptotically the assumption of Proposition 3. It will be necessary though to correct over the $p(p-1)/2$ multiple tests over all pairs of vertices (i,j). The estimation algorithm is as follows.

Estimation algorithm

1. For all $i, j \in \{1, \ldots, p\}$, $i \neq j$ and $k \in \{1, 2, \ldots, p\} \setminus \{i, j\}$, compute P-values $P(i, j|k)$ from the log-likelihood ratio test with respect to the model $X_i, X_j, X_k \sim \mathcal{N}(0, \Sigma)$ with null hypothesis $H_0(i, j|k)$: $\Sigma^{-1}_{ij} = 0$ and alternative $H_1(i, j|k)$: $\Sigma^{-1}_{ij} \neq 0$. Also, compute $P(i, j|\emptyset)$ from the log-likelihood ratio test with null hypothesis $H_0(i, j|\emptyset)$: $\Sigma_{ij} = 0$ and alternative $H_1(i, j|\emptyset)$: $\Sigma_{ij} \neq 0$. Note the symmetry $P(i, j|k) = P(j, i|k)$.

2. For all pairs $(i, j) = (j, i)$ compute the maximum P-values (note the correspondence to Proposition 3)

$$P_{max}(i, j) = \max_{k \in \{\emptyset, 1, 2, \ldots, p\} \setminus \{i, j\}} P(i, j|k).$$

3. Correct the maximum P-values $P_{max}(i, j)$ over the $p(p-1)/2$ multiple tests for all pairs of vertices. For example, use the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) for controlling the false discovery rate. Alternatively, the family-wise error rate could be controlled. Denote the corrected maximal P-values by

$$P_{max,corr}(i, j).$$

4. Draw an edge between vertex i and j if and only if

$$P_{max,corr}(i, j) < \alpha,$$

for some pre-specified significance level such as $\alpha = 0.05$.

The corrected maximum P-values $P_{max,corr}(i, j)$ can be used as a measure of significance for an edge between nodes i and j. The maximum P-value $P_{max}(i, j)$ may often be an over-conservative estimate of the type I error for edge i, j. It should be noted, however, that we test the null hypothesis that at least one $H_0(i, j|k)$ is true versus the alternative that none $H_0(i, j|k)$ is true. Therefore, less conservative approaches (Holm, 1979; Simes, 1986) are not applicable. For a fixed pair of nodes, Proposition 2 and 3 imply the following.

Corollary 2 *Let G be the concentration graph representing a Gaussian distribution $\mathbf{X}$. For some fixed pair of nodes (i, j), assume that the conditions of Propositions 2 (about the separateness of i and j) and Proposition 3 hold. Then,*

$$\mathbb{P}[\textit{an edge between nodes } i \textit{ and } j \textit{ is estimated in the 0-1 graph but there is no edge between } i \textit{ and } j \textit{ in } G] \leq \alpha.$$

DOI: 10.2202/1544-6115.1170

It is worth pointing out that our estimation for a 0-1 graph is done in an exhaustive manner where each edge is tested separately for inclusion in the graph. The number of $\frac{p(p-1)}{2}$ tests that have to be conducted is feasible even for a large number of vertices p. Our approach is in line with the hypothesis testing-based model selection for concentration graphs (Drton & Perlman, 2004) and is in contrast to searching the huge graph space of $2^{p(p-1)/2}$ models with non-exhaustive computational methods such as random search methods, greedy stepwise algorithms, or stochastic simulation in the Bayesian framework (Madigan & Raftery, 1994; Giudici & Green, 1999; Dobra *et al.*, 2004).

Asymptotic consistency for large number of variables

We present here some theory which reflects at least from an asymptotic point of view that 0-1 graphs can be accurately estimated even if the number p of vertices is large relative to sample size.

Denote the data by $\mathbf{X}_1, \ldots, \mathbf{X}_n$ ($\mathbf{X}_i \in \mathrm{R}^p$) which are assumed to be i.i.d. random vectors. The estimators for the mean $\mu = \mathbb{E}[\mathbf{X}]$, the covariance matrix $\Sigma = \mathrm{Cov}(\mathbf{X})$, the correlation coefficients ρ_{ij} and the partial correlation coefficients $\omega_{ij|k}$ are as follows:

$$\begin{aligned}
\hat{\mu}(n) &= n^{-1} \sum_{i=1}^{n} \mathbf{X}_i, \\
\hat{\Sigma}(n) &= n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\mu})(\mathbf{X}_i - \hat{\mu})^T \\
\hat{\rho}(n)_{ij} &= \frac{\hat{\Sigma}(n)_{ij}}{\sqrt{\hat{\Sigma}(n)_{ii}\hat{\Sigma}(n)_{jj}}} \\
\hat{\omega}(n)_{ij|k} &= \frac{\hat{\rho}_{ij} - \hat{\rho}_{ik}\hat{\rho}_{jk}}{\sqrt{(1-\hat{\rho}_{ik}^2)(1-\hat{\rho}_{jk}^2)}}, \ 1 \le i < j \le p, \ k \ne i, j. \qquad (4)
\end{aligned}$$

We are giving below some uniform consistency results for these estimators when the dimensionality p is large relative to sample size. The set-up is as follows. We assume that the data are realizations from a triangular array of random vectors of dimension $p = p_n$ where p_n is allowed to grow as sample size $n \to \infty$:

$$\mathbf{X}_{(n),1}, \ldots, \mathbf{X}_{(n),n} \text{ i.i.d. } \sim P_{(n)}, \qquad (5)$$

where $P_{(n)}$ denotes some probability distribution in R^{p_n}. We denote by $\mu(n) =$

$\mathbb{E}[\mathbf{X}_{(n)}]$ and by $\Sigma(n) = \mathrm{Cov}(\mathbf{X}_{(n)})$; these moments exist by the following assumption.

(A1) $\quad \sup_{n\in\mathrm{N},1\le j\le p_n} \mathbb{E}|(\mathbf{X}_{(n)})_j|^{4s} < \infty$ for some $s \ge 1/2$.

Proposition 4 *The data are as in (5), satisfying assumption (A1) for some $s \ge 1/2$. Assume that $p_n = o(n^{s/2})$ $(n \to \infty)$. Then,*

$$\max_{1\le j\le p_n} |\hat{\mu}(n)_j - \mu(n)_j| = o_P(n^{-3s/2}) \ (n \to \infty),$$
$$\max_{1\le i<j\le p_n} |\hat{\Sigma}(n)_{ij} - \Sigma(n)_{ij}| = o_P(1) \ (n \to \infty).$$

In case where $\mathbf{X} \sim \mathcal{N}_{p_n}(\mu(n), \Sigma(n))$, we could allow of a faster growth rate p_n satisfying $\log(p_n)/n \to 0$.

For uniform consistency of partial correlations, we make an additional assumption:

(A2) $\quad \inf_{n\in\mathrm{N},1\le j\le p_n} \Sigma(n)_{jj} > 0$, and $\sup_{n\in\mathrm{N},1\le i<j\le p_n} |\rho(n)_{ij}| < 1$.

The first assumption in (A2) means that none of the variables becomes degenerate as $n \to \infty$, i.e. having a variance tending to zero. The second assumption says that all the variables are linearly identifiable, i.e. none of the variables is an exact linear function of another one.

Proposition 5 *The data are as in (5), satisfying assumption (A1) for some $s \ge 1/2$ and (A2). Assume that $p_n = o(n^{s/2})$ $(n \to \infty)$. Then,*

$$\max_{1\le i<j\le p_n} |\hat{\rho}(n)_{ij} - \rho(n)_{ij}| = o_P(1) \ (n \to \infty),$$
$$\max_{1\le i<j\le p_n; 1\le k\le p_n, k\ne i,j} |\hat{\omega}(n)_{ij|k} - \omega(n)_{ij|k}| = o_P(1) \ (n \to \infty).$$

Also here, in case where $\mathbf{X} \sim \mathcal{N}_{p_n}(\mu(n), \Sigma(n))$, we could allow of a p_n satisfying $\log(p_n)/n \to 0$. Proposition 5 describes a *uniform* convergence result for the ϕ_{ij} parameters in (3): for a small number $\delta > 0$ and with high probability, all estimated marginal and partial correlations are within δ-distance from the true partial correlations if the sample size is sufficiently large. This is much stronger than a pointwise result. Since a 0-1 graph involves all marginal and partial correlations, see Definition 3, our uniform consistency result, saying that we can *simultaneously* estimate *all* of them reasonably well, implies that we can estimate a 0-1 graph reasonably well even if the number of vertices p is much larger than sample size n. In fact, consistent estimation of

high-dimensional 0-1 graphs is possible if true non-zero partial and marginal correlations are bounded away from zero.

The 0-1 graph can be consistently estimated under the following additional assumption:

(A3) $\inf_{1\le i<j\le p_n, n\in\mathbb{N}}\{|\rho(n)_{ij}|; \rho(n)_{ij}\neq 0\} > C_1 > 0$, and

$\inf_{1\le i<j\le p_n, 1\le k\le p_n, k\neq i,j, n\in\mathbb{N}}\{|\omega(n)_{ij|k}|; \omega(n)_{ij|k}\neq 0\} > C_2 > 0.$

Proposition 6 *Consider the following 0-1 graph estimate $\hat{G}_{0-1}(n, K)$ which is a theoretical simplified version of our algorithm described above:*

draw an edge between nodes i and j if and only if $\hat{\phi}(n)_{ij} > K$,

where $\hat{\phi}(n)_{ij}$ is the estimate of ϕ_{ij} in (3). Assume the conditions from Proposition 5 and assumption (A3). Then, the 0-1 graph can be estimated consistently, i.e. for some suitable K,

$$\mathbb{P}[\hat{G}_{0-1}(n, K) \;=\; \textit{true 0-1 graph}] \to 1 \;(n\to\infty).$$

The estimation method in the proof of Proposition 6 is non-constructive since we do not know the constants C_1 and C_2 in (A3). Nevertheless, Proposition 6 indicates the potential of estimating the correct underlying 0-1 graph with probability tending to one as sample size increases.

It should be stated clearly that the bound in Proposition 5 is generally worse, although still $o_P(1)$, than in Proposition 4 for the covariances. Clearly, the result from Proposition 5 could be generalized to partial correlations $\omega(X_i, X_j|\{X_{k_1}, \ldots, X_{k_m}\})$ $(k_1, \ldots k_m \neq i, j)$ for a *fixed* m with respect to sample size n (although a uniform bound for such partial correlations is expected to become worse as the the value of m increases). If $m = m_n$ would grow with sample size, we would have to further restrict the growth of the dimensionality p_n.

The extreme case is the estimate of $\Sigma(n)^{-1}$ when inverting the estimate $\hat{\Sigma}(n)$ from (4). This can only be done if $p_n < n$ and pointwise consistency $|(\hat{\Sigma}(n))^{-1}_{ij} - \Sigma(n)^{-1}_{ij}| = o_P(1)$ $(1 \le i < j \le p_n)$ only holds if $p_n = o(n)$ $(n\to\infty)$ (Lauritzen, 1996). Thus, the unconstrained graphical Gaussian model can only be estimated if the dimensionality is "small" relative to the sample size. This is in sharp contrast to 0-1 graphs, where p_n is allowed to grow much faster than n, as described in Proposition 5. For example, by neglecting the constants in Proposition 5, the following dimensionalities are allowed for $n = 100$ and $4s$ existing moments for the components of $\mathbf{X}$:

$n = 100$	$4s = 8$	$4s = 12$	$4s = 16$	$4s = 20$
$p = o(n^{s/2})$	$o(100)$	$o(1'000)$	$o(10'000)$	$o(100'000)$

.

In a graph with p vertices, the maximum number of edges is $\frac{p(p-1)}{2}$. If the number of actually present edges is much smaller than $\frac{p(p-1)}{2}$, a graph is generally referred to as being sparse. For example, it can be assumed that the number of edges grows only linearly (or even less) in the number of vertices p. Alternatively, the number of neighbors per vertex could be restricted (Dobra *et al.*, 2004; Meinshausen & Bühlmann, 2004).

Under such sparsity assumptions for the true concentration graph, regularization methods could be used to cope with large p in the estimation of the concentration graph (Dobra *et al.*, 2004; Meinshausen & Bühlmann, 2004). In comparison, consistent 0-1 graph estimation is not subject to such sparsity assumptions.

Numerical results for simulated data

In the previous section, we have shown that the 0-1 graph can be consistently estimated. Furthermore, we have shown that for sparse concentration graphs that are trees or fulfill the conditions of Proposition 2, the 0-1 and the concentration graph coincide. For faithful concentration graphs, the edges form a subset of the edges of the corresponding 0-1 graph.

In this section we show in simulations that a focus on simpler aspects of conditional independence in combination with good estimation properties make 0-1 graphs a good estimator for full conditional independence relationships in sparse graphs.

For metabolic, genetic regulatory or protein interaction networks, it has been repeatedly suggested that the connectivity of the vertices follows a power law with exponents γ between 2 and 3 (Jeong *et al.*, 2000; Maslov & Sneppen, 2002). In our simulations of Gaussian graphs with many nodes, we adopt this network structure by sampling the number of edges for each node independently from a power-law distribution $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$ with exponent $\gamma = 2.5$. The normalization constant $\zeta(\gamma)$ is the Riemann zeta function. The graphs that we obtain by this method are very sparse and usually contain fewer edges than the number of nodes (see Table 1). In order to simulate graphs with more edges, we also generate graphs with exponent $\gamma = 1.5$ and 0.5.

Edges are then randomly assigned to other nodes (with equal probabilities). This random graph structure is used to define the zeros in the precision matrix: $\Sigma^{-1}_{ij} = 0$ if there is no edge between i and j. In order to model the non-

$X_i \perp\!\!\!\perp X_j \mid T$

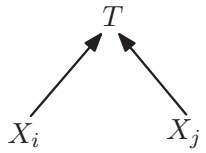

Figure 4: Conditional independence model and associated graph for X_i, X_j and the selection variable T.

zero elements of Σ^{-1} (and the partial correlation coefficients), we introduce a selection variable T_{ij} for each pair of adjacent vertices i and j. X_i and X_j are assumed to be marginally independent and to have an effect on the variable T_{ij} (Figure 4). The corresponding graph, also called the canonical directed acyclic graph (Richardson & Spirtes, 2002), only comprises directed edges $i \to T_{ij}$ and $j \to T_{ij}$. Selection for specific values for T_{ij} corresponds to conditioning on the selection variables T_{ij} yielding a graph with the desired graph structure.

If we only consider the three variables X_i, X_j and T_{ij}, we could model the effect of X_i and X_j on T_{ij} with a covariance matrix

$$\Sigma_{X_i,X_j,T_{ij}} = \begin{pmatrix} 1 & 0 & \beta_{ij} \\ 0 & 1 & \beta_{ji} \\ \beta_{ij} & \beta_{ji} & 1+\beta_{ij}^2+\beta_{ji}^2 \end{pmatrix}.$$

Magnitude and sign of the coefficients β_{ij} and β_{ji} determine how strong the effect of X_i and X_j is on T_{ij} respectively. After conditioning on T_{ij}, we obtain (considering T_{ij} unobserved now)

$$\Sigma^{-1}_{X_i,X_j} = \begin{pmatrix} 1+\beta_{ij}^2 & \beta_{ij}\beta_{ji} \\ \beta_{ij}\beta_{ji} & 1+\beta_{ji}^2 \end{pmatrix}.$$

We can therefore write

$$\Sigma^{-1}_{X_i,X_j} = Id + BB^t \tag{6}$$

with

$$B = \begin{pmatrix} \beta_{ij} \\ \beta_{ji} \end{pmatrix}.$$

If we model partial correlation coefficients for all variables $X_1, \ldots, X_p$, we use the complete canonical directed acyclic graph (DAG) and extend the scheme described in (6). Let $\{e_{kl}\}$ be the edges in the graph where the indices $k < l$

refer to the indices of the variables X_k and X_l that are connected by e_{kl}. Let further e be the total number of edges and B a $p \times e$ matrix with elements

$$b_{ie_{kl}} = \begin{cases} \beta_{il} & \text{if } i = k \\ \beta_{ik} & \text{if } i = l \\ 0 & \text{otherwise.} \end{cases}$$

Then we find

$$\begin{aligned}(BB^t)_{ij} &= \sum_{e_{kl}} b_{ie_{kl}} b_{je_{kl}} \\ &= \begin{cases} \sum_{e_{ik}} \beta_{ik}^2 & \text{if } i = j \\ \beta_{ij}\beta_{ji} & \text{if } i \neq j \text{ and there is an edge between } i \text{ and } j \\ 0 & \text{if } i \neq j \text{ and there is no edge between } i \text{ and } j \end{cases}\end{aligned}$$

and the partial correlation coefficient for two conditionally dependent variables X_i and X_j can be modeled as (Equations (1) and (6))

$$\omega_{ij} = \frac{-\beta_{ij}\beta_{ji}}{\sqrt{1 + \sum_{e_{ik}} \beta_{ik}^2}\sqrt{1 + \sum_{e_{jk}} \beta_{jk}^2}}.$$

The random graph structure and B define a normal distribution $N(0, \Sigma)$. The magnitude and sign of the coefficients β_{ij} determine the magnitude and sign of the partial correlation coefficients. In our simulations, we sampled the coefficients β_{ij} from three different uniform distributions $U(-\beta_{\max}, \beta_{\max})$ with $\beta_{\max} = 1, 5, 100$.

The use of canonical DAGs to generate partial correlation coefficient for a pre-specified concentration graph has the advantage that it is very flexible while keeping the sampled precision matrix always positive definite. The assumption that a dependence is due to a unobserved random variable generates a particular parametrization. Not all multivariate normal distributions represented by a graph, can be parameterized by a canonical DAG (Richardson & Spirtes, 2002). Still, this parametrization can display various scenarios that seem relevant in biological studies. From the various factors that play a role in genetic regulation, many will be unknown. Our parametrization scheme suits particularly well to account for these factors as well as the sparse structure of the graphs.

Our parametrization can also nearly represent direct relationships between nodes (directed edges). If for example the latent random variable has a strong

number of variables	γ	number of edges in the independence graph	0-1 graph	covariance graph
p=5	2.5	3.53(0.70)	3.56(0.81)	6.52(2.98)
	1.5	4.14(1.04)	4.38(1.57)	7.84(2.89)
	0.5	5.59(1.19)	6.47(2.07)	9.82(1.03)
p=10	2.5	7.46(1.51)	7.76(2.31)	20.68(14.38)
	1.5	10.87(2.67)	15.02(7.70)	38.46(11.42)
	0.5	18.86(3.82)	33.02(7.45)	45.00(0.00)
p=20	2.5	15.48(2.91)	16.87(8.08)	56.68(46.09)
	1.5	24.42(4.32)	51.27(23.96)	166.45(43.82)
	0.5	44.97(6.70)	130.00(26.17)	190.00(0.00)
p=40	2.5	30.45(3.80)	31.08(7.40)	115.66(91.62)
	1.5	49.70(6.79)	173.03(85.08)	680.35(162.73)
	0.5	88.34(8.74)	498.33(82.76)	780.00(0.00)

Table 1: Mean number of edges (and standard deviation) for the three different graphical models in Section as a function of γ and p.

effect on X_i, i.e. β_{ij} is large, the latent random variable can be merged with X_i and β_{ji} measures the direct effect of X_i on X_j. The edge then represents a directed edge.

As an alternative, a parametrization using hyper inverse Wishart distribution could have been applied to simulate concentration matrices. However, this approach is most useful in the context of conjugate Bayesian inference, since a prior concentration matrix would have to be specified. Also, it is rather tedious to sample large sparse non-decomposable models (Roverato, 2002).

With our parametrization scheme, we generated 100 graphs and covariance matrices each for graphs with p =5, 10, 20, and 40 vertices and connectivity parameter γ =2.5, 1.5 and 0.5. For each p and each γ, we compared the structure of the concentration graph, the covariance graph and the 0-1 graph. In Table 1, the mean and standard deviations for the number of edges per graph is shown. For decreasing γ, the number of edges increases in the concentration graphs. The edges of the concentration graph almost always formed a subset

number of variables	γ	RMSE covariance graph $\omega_{ij}=0$	covariance graph $\omega_{ij}\neq 0$	0-1 graph $\omega_{ij}=0$	0-1 graph $\omega_{ij}\neq 0$
	2.5	0.221	0.144	0.002	0.029
p=5	1.5	0.268	0.189	0.009	0.04
	0.5	0.251	0.178	0.02	0.056
	2.5	0.161	0.16	0.004	0.046
p=10	1.5	0.168	0.151	0.01	0.046
	0.5	0.118	0.1	0.018	0.042
	2.5	0.105	0.155	0.001	0.044
p=20	1.5	0.111	0.136	0.007	0.05
	0.5	0.065	0.066	0.011	0.031
	2.5	0.075	0.162	0.001	0.045
p=40	1.5	0.076	0.127	0.005	0.051
	0.5	0.046	0.058	0.006	0.028

Table 2: RMSE averaged over all $i<j$ with $\omega_{ij}=0$ and averaged over all $i<j$ with $\omega_{ij}\neq 0$ between correlation coefficients ρ_{ij} and partial correlation coefficient ω_{ij} (right columns) and RMSE between 0-1 graph coefficients ϕ_{ij} and partial correlation coefficients ω_{ij} (left columns). $\beta_{\max}=5$.

of the 0-1 graph. For graphs with low connectivity ($\gamma=2.5$), the 0-1 graph contained only few additional edges indicating that mostly trees were sampled. However, for $\gamma=1.5$ and $\gamma=0.5$, the 0-1 graphs were considerably larger than the corresponding concentration graphs. Although being sparse, the concentration graphs must therefore contain a considerable number of cycles, see Proposition 2.

We also monitored the difference between the correlation and partial correlation coefficients ($\rho_{ij}-\omega_{ij}$) and the difference between 0-1 graph and partial correlation coefficients ($\phi_{ij}-\omega_{ij}$) for unconnected ($\omega_{ij}=0$) and connected ($\omega_{ij}\neq 0$) vertices i and j (see Table 2 for the root mean squared errors (RMSE) averaged over all $i<j$). Most edges in the 0-1 graph that are not part of the concentration graph have coefficients in the vicinity of 0. In fact, for $\omega_{ij}=0$ the 5%- and 95%-quantile of the distribution of 0-1 graph coefficients were

number of variables p	number of observations n
5	10,20,30,50,100,500,1000,5000
10	20,30,50,100,500,1000,5000
20	30,50,100,500,1000,5000
40	50,100,500,1000,5000

Table 3: Number of observations n used to sample data from the original graphs with p vertices

located within the interval [-0.05, 0.05] for all simulation settings. For $\omega_{ij} \neq 0$, the 5%-95%-quantile ranges were always larger. This indicates that the 0-1 graph can capture the conditional independence structure quite well, and much better than the covariance graph.

Estimation results with sampled data

From each of the simulated models, we sampled i.i.d. data from $\mathcal{N}(0, \Sigma)$ where Σ is the covariance matrix of the corresponding model parameters as described by equation (6). Depending on the size of the graph, we sampled data with few and many observations (see Table 3). The effect of the sample sizes on the estimates of the partial correlation coefficients $\hat{\omega}_{ij}$, 0-1 graph coefficients $\hat{\phi}_{ij}$ and correlation coefficients $\hat{\rho}_{ij}$ can be seen in Figures 5-8.

Figure 5 shows the root mean squared error (RMSE) between true coefficients and the corresponding estimates of the different graphical modeling approaches. Results are shown for $\gamma = 1.5$ and $\beta_{\max} = 5$. It can be seen that for small n, the RMSE of the coefficients $\omega_{ij|k}$ does not differ much from the RMSE of the correlation coefficients ρ_{ij} and that both coefficients can be more accurately estimated than the full partial correlation coefficients ω_{ij}. As the number of observations n increases, the RMSEs for all coefficients decrease to 0. In all simulation settings, we found the same underlying pattern as in Figure 5. For $\beta_{\max} = 1$, however, the RMSE of the coefficients differed only slightly, even when n was small. Interestingly, estimates of the 0-1 graph coefficients ϕ_{ij} (see (3)) are even better than the estimates of the coefficients ρ_{ij} and $\omega_{ij|k}$. This indicates that the minimum of ρ_{ij} and $\omega_{ij|k}$ for $k \in \{1, 2, \ldots, p\} \setminus \{i, j\}$ can be much more reliably estimated than each of the coefficients ρ_{ij} and $\omega_{ij|k}$

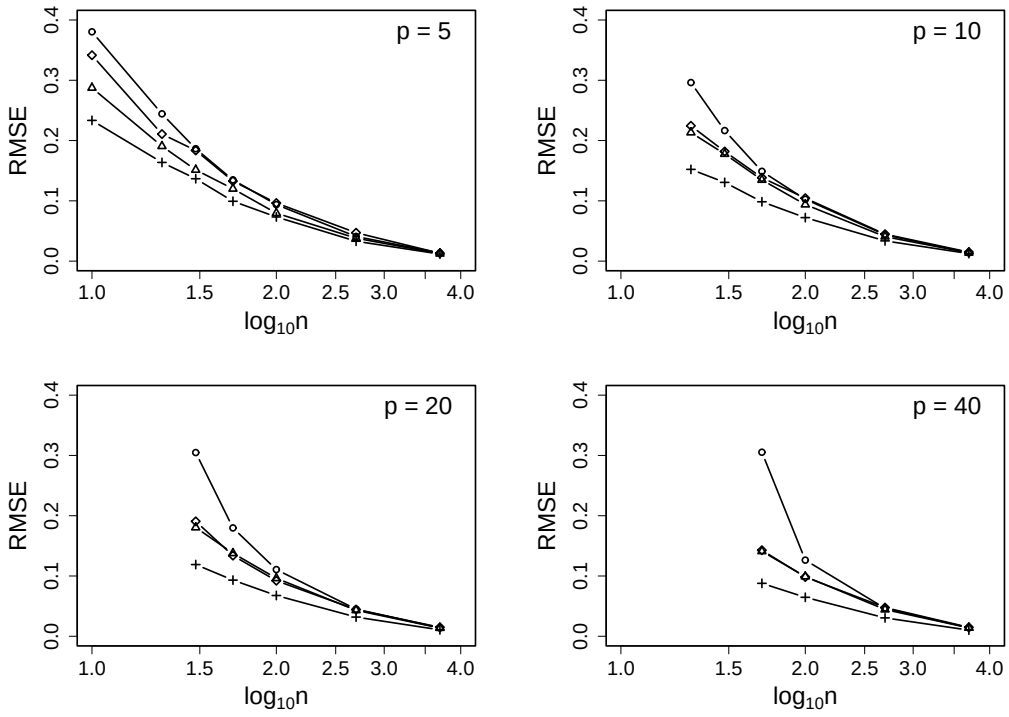

Figure 5: Root mean squared error (RMSE) averaged over all $i < j$ between the sampled and true partial correlation coefficients $\hat{\omega}_{ij}$ and ω_{ij} ($\circ$), sampled and true correlation coefficients $\hat{\rho}_{ij}$ and ρ_{ij} ($\triangle$), $\hat{\omega}_{ij|k}$ and $\omega_{ij|k}$ ($\diamond$) and sampled and true 0-1 graph coefficients $\hat{\phi}_{ij}$ and ϕ_{ij} ($+$) for different network sizes p and different number of observations n.

for $k \in \{1, 2, \ldots, p\} \setminus \{i, j\}$ separately. Proposition 5 can therefore be viewed as providing a conservative upper bound for the estimation accuracy of the 0-1 graph coefficients.

We also monitored how well the estimates of the full partial correlation coefficients $\hat{\omega}_{ij}$, the 0-1 graph coefficients $\hat{\phi}_{ij}$ and the correlation coefficients $\hat{\rho}_{ij}$ represent the true full partial correlation coefficients ω_{ij} of the original concentration graph. In Figure 6, the RSME between the sampled partial correlation coefficients, the sampled 0-1 graph coefficients, the sampled correlation coefficients and the true partial correlation coefficients are shown. For small to moderate n, the concentration graph is better represented by the estimated 0-1 graph coefficients than the estimated partial correlation coefficients. Therefore, although being a rather simple estimator of complex dependence

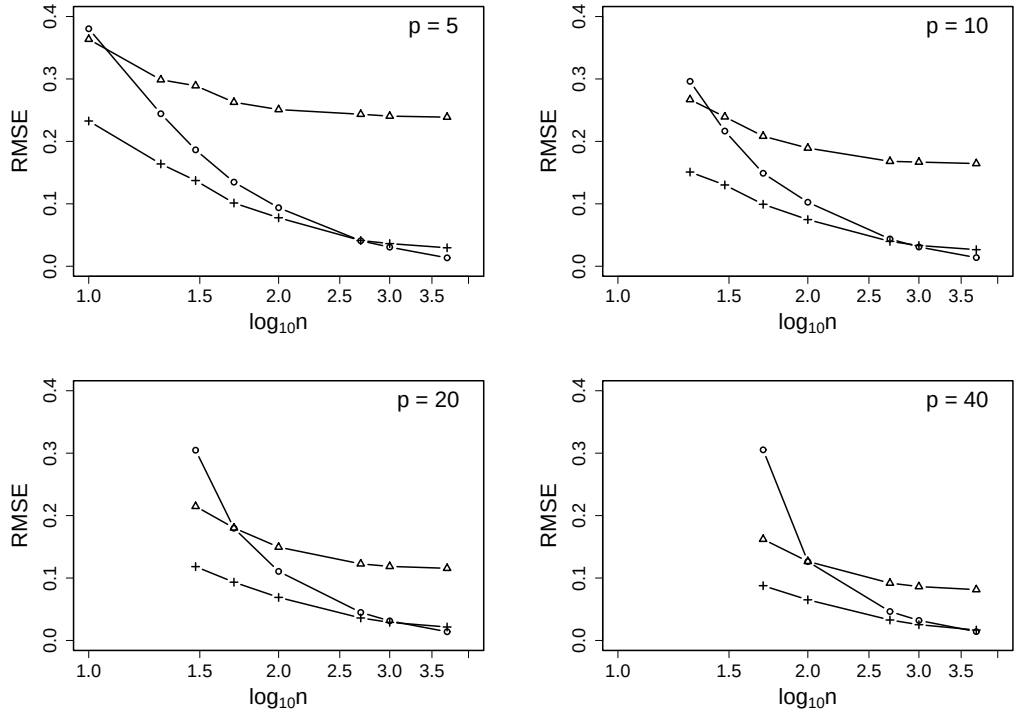

Figure 6: Root mean squared error (RMSE) averaged over all $i < j$ between sampled partial correlation coefficients $\hat{\omega}_{ij}$ and true partial correlation coefficients ω_{ij} ($\circ$), sampled correlation coefficients $\hat{\rho}_{ij}$ and ω_{ij} ($\triangle$) and 0-1 graph coefficients $\hat{\phi}_{ij}$ and ω_{ij} ($+$) for different network sizes p and different number of observations n.

patterns, 0-1 graph coefficients can outperform partial correlation coefficients in detecting conditional dependence/independence.

Figure 7 shows the cumulative distribution functions (CDF) of the different coefficients for pairs of vertices with and without edges. Again, one can clearly see that a small to moderate sample size ($n = 50$) leads to rather unreliable estimates $\hat{\omega}_{ij}$ for the concentration graph (reflected by a gradual slope of the CDF of $\hat{\omega}_{ij} - \omega_{ij}$ at 0) whereas estimates of the 0-1 graph coefficients $\hat{\phi}_{ij}$ are much more stable (steeper slope of the CDF of $\hat{\phi}_{ij} - \omega_{ij}$).

In graphs with many nodes, the main purpose of a study may not be to find all connections between nodes but to find some true connections, hopefully the most important ones. In such a procedure, one would only consider gene pairs whose absolute partial correlation coefficient or 0-1 graph coefficient would be above a certain threshold t. By counting the number of true and false

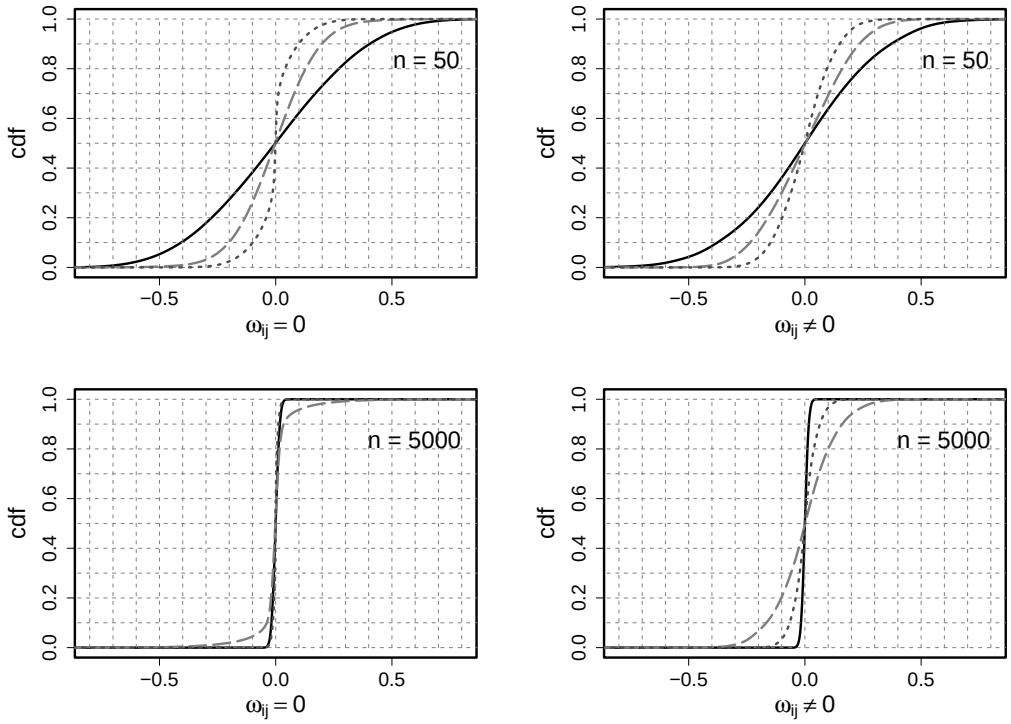

Figure 7: Cumulative distribution function (CDF) of the difference between sampled partial correlation coefficient $\hat{\omega}_{ij}$ and true partial correlation coefficients ω_{ij} (black line), between sampled correlation coefficients $\hat{\rho}_{ij}$ and ω_{ij} (dashed pale grey line) and sampled 0-1 graph coefficients $\hat{\phi}_{ij}$ and ω_{ij} (dotted grey line) for $p = 40$ and $n = 50$ (upper panel) or $n = 5000$ (lower panel) observations.

positives, true and false negatives for all values $t \in [0, 1]$, one obtains the so called ROC curves by plotting the sensitivity (true positive rate) against the complementary specificity (false positive rate) for each t. The upper panel of Figure 8 displays the average ROC curves for the concentration graph, the covariance graph and the 0-1 graph for p=40 and $\beta_{\max} = 100$. We also included the ROC curves for learning the concentration graph based on backward selection within the maximum likelihood framework, as implemented in the MIM package (2003). For small complementary specificities, the ROC curve of the 0-1 graph has a steeper slope than the other ROC curves suggesting the best performance in detecting true positive edges of the concentration graph.

The 0-1 graph outperforms all the other methods (including the backward selection approach) for a small (n=100) and a large (n=1000) number

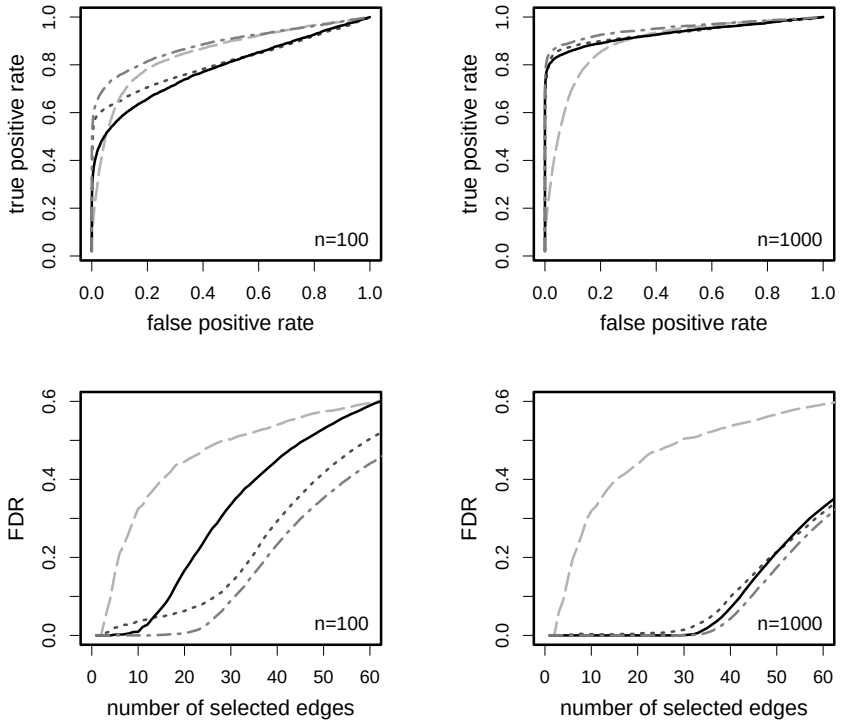

Figure 8: ROC curves (upper panel) and the False Discovery Rate (FDR) as a function of the number of selected edges (lower panel) for the covariance graph (dashed pale grey line), the 0-1 graph (dash-dotted grey line), the concentration graph (black line) and the concentration graph learned under backward selection (dotted dark grey line). Here, $p = 40$.

of observation. For n=1000 observations, however, the ROC curves of the 0-1 graph, the concentration graph and the backward selection approach differ only marginally. Our findings are further substantiated when we look at the false discovery rate (FDR) as a function of the selected edges. Again, the FDR of the 0-1 graph is smaller than the ones of the other methods.

All the simulations were based on 100 graphs. For $p = 40$ genes, a single computation of the 0-1 graph could be completed in the order of seconds whereas the computation of the concentration graph with backward selection (with MIM) took approximately 15 minutes (on a 2.6GHz Pentium 4 machine). Simulations that included forward selection was computationally not feasible.

Application to gene expression microarray data

In this section, we will further discuss and motivate the usefulness of 0-1 graphs for the inference of genetic regulatory networks. We will here focus on the applications presented in Magwene & Kim (2004) and Wille *et al.* (2004).

Magwene & Kim (2004) estimated the coexpression network of 5007 yeast open reading frames (ORFs) based on 87 microarrays. Their inferred network contained 11450 edges most of which (11416) were included in one single giant connected component. To further analyze their network, the authors compared their network with 38 metabolic pathways and also studied the biological relevance of locally distinct subgraphs.

They found that 99% of vertex pairs in the 0-1 network were separated by a shortest path with more than 2 edges. In order to evaluate the coherence between metabolic pathways and the estimated 0-1 network, starting from the set P of genes assigned to one pathway, they searched for connected components in which no vertex was more than 2 edges away from at least one other node in that component. If O denotes the maximum overlap between the genes of each single component and the pathway genes P, the ratio $\frac{|O|}{|P|}$ was taken as measure for the coherence between 0-1 network and metabolic network. 19 of the 38 metabolic pathways had coherence values that were significant when compared to random pathways of the same size.

Another way to validate the biological relevance of a genetic network is to search for functional enrichment based on Gene Ontology annotation (Gene Ontology Consortium, 2001) in dense subgraphs of the network. The authors used an unsupervised graph algorithm to determine subgraphs whose network topology differs from the neighboring nodes with respect to density. They could find 32 locally distinct subgraphs 24 of which were enriched for biological function (Gene Ontology annotation).

Whereas Magwene & Kim (2004) focused on the properties of the 0-1 network comprising the majority of yeast genes, our group (Wille *et al.*, 2004) applied 0-1 graphs to a smaller group of 40 isoprenoid genes to study in more detail the regulatory network of isoprenoid biosynthesis in *Arabidopsis thaliana*. Isoprenoids comprehend the most diverse class of natural products and have been identified in many different organisms including viruses, bacteria, fungi, yeasts, plants, and mammals. In plants, isoprenoids play important roles in a variety of processes such as photosynthesis, respiration, regulation of growth and development, and in protecting plants against herbivores and pathogens.

In higher plants such as *Arabidopsis thaliana*, two distinct pathways for the formation of isoprenoids exist, one in the cytosol (MVA pathway) and the other in the chloroplast (MEP pathway). Although both pathways operate fairly

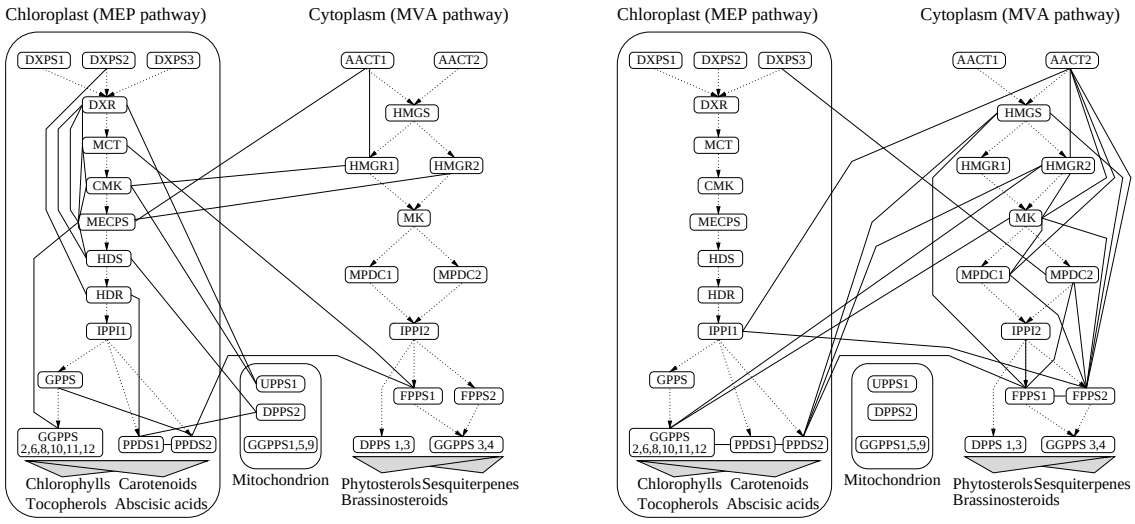

Figure 9: 0-1 graph of the isoprenoid pathways. Left panel: subgraph of the gene module in the MEP pathway, right panel: subgraph of the gene module in the MVA pathway.

independently under normal conditions, interaction between them has been repeatedly reported (Laule *et al.*, 2003; Rodriguez-Concepcion *et al.*, 2004). In order to gain better insight into the crosstalk between both pathways on the transcriptional level, gene expression patterns were monitored under various experimental conditions using 118 microarrays.

Figure 9 shows the network model obtained from the 0-1 graph. Since we find a module with strongly interconnected genes in each of the two pathways, we split up the graph into two subgraphs each displaying the subnetwork of one module and its neighbors.

In the MEP pathway, the genes DXR, MCT, CMK, and MECPS are nearly fully connected (left panel of Figure 9). From this group of genes, there are a few edges to genes in the MVA pathway. Similarly, the genes AACT2, HMGS, HMGR2, MK, MPDC1, FPPS1 and FPPS2 share many edges in the MVA pathway (right panel of Figure 9). The subgroup AACT2, MK, MPDC1, FPPS2 is completely interconnected. From these genes, we find edges to IPPI1 and GGPPS12 in the MEP pathway.

In the conventional graphical modeling with backward selection, we could only identify the gene module in the MEP pathway. The genes in the MVA pathway did not form a separate regulatory structure, even when more edges were included in the model. In the 0-1 graph, the detection of the additional gene module in the MVA pathway is in good agreement with earlier find-

ings that within a pathway, potentially many consecutive or closely positioned genes are jointly regulated (Ihmels *et al.*, 2004). Also, a high level of coexpression between the genes AACT2, MK, MPDC1, FPPS2 suggests a separate regulatory module in the MVA pathway.

In addition to that we attached 795 genes from 56 other metabolic pathways to the inferred network. We found that genes from downstream pathways that use isoprenoids as substrates attach significantly better to the 0-1 network than genes from other (unrelated) pathways. This provides an additional biological validation for the network.

Conclusions

Graphical Gaussian modeling suffers from unreliable estimates of the full partial correlation coefficients if the number of observations is relatively small in comparison with the number of random variables in the model. In order to still be able to analyze the conditional dependence structure between variables, one can focus on zero- and first-order conditional dependencies as a simplified measure of dependence.

The 0-1 graph coefficients proved to be powerful in two ways. First, we showed theoretically that the 0-1 graph coefficients have nearly the same good estimation properties as the more simple correlation coefficients. Second, for small sample sizes in our simulation framework, the estimated 0-1 graph coefficients were on average better estimators of the full partial correlation coefficients than the estimated full partial correlation coefficients themselves. This finding indicates that although full partial correlation coefficients take the effect of many other variables into account, only few of these variables have a large effect on the dependence structure. For sparse graphs, modeling approaches based on low-order conditional dependencies can therefore be generally preferable to methods based on full conditional dependencies. Proposition 1 and 2 give some additional theoretical underpinning why the 0-1 graph works so well.

The 0-1 graph approach carries resemblance to the first two steps in the SGS and PC algorithm (Spirtes *et al.*, 2000) and the algorithm presented by de Campos & Huete (2000). These algorithms use low-order conditional independencies as a first step to infer the concentration graph. In the 0-1 graph, modeling is limited to zero- and first-order independencies only. By this simplification, we completely avoid to carry out the statistically unreliable and computationally costly search for conditional independence in large subsets. We have shown that this can be a good strategy to model sparse graphical

models with many nodes and only few observations.

By generating the number of edges in a graph according to a power law, we aimed at simulating network topologies found in biological networks. Other examples include computer and social interaction networks (Barabasi & Albert, 1999). With this restriction, only a subclass of sparse conditional independence models is considered. However, the restriction enabled us to consistently study the effect of the sample size, the number of vertices, the level of sparsity and the level of conditional dependencies on the various graphical modeling approaches.

Appendix

Proof of Proposition 1:
Assume that the edge i, j is not in the 0-1 conditional independence graph G_{0-1}. Then we either have $\rho_{ij} = 0$ or $\omega_{ij|k} = 0$ for some $k \in \{1, \ldots, p\} \setminus \{i, j\}$. In the first case, X_i and X_j are marginally independent, i.e. i and j are in different connectivity components of G, since G is faithful. In the latter case, $X_i \perp\!\!\!\perp X_j | X_k$, i.e. k separates i and j in G since G is faithful. Therefore, there is no direct edge between i and j. □

Proof of Proposition 2:
Assume that i and j are not adjacent in G. Then we either have
1) i and j are in different connectivity components. X_i and X_j are therefore marginally independent, which implies $\rho_{ij} = 0$ and that there is no edge between i and j in G_{0-1}, or
2) There exists some $k \in \{1, \ldots, p\} \setminus \{i, j\}$ that separates i and j. Due to the Markov property, we have $X_i \perp\!\!\!\perp X_j | X_k$ and therefore $\omega_{i,j|k} = 0$, which further implies that i and j are not adjacent in G_{0-1}. □

Proof of Proposition 3: Consider the hypothesis

$$H_0 = H_0(i, j): \text{ at least one } H_0(i, j|k^*) \text{ is true for some } k^*.$$

The probability for a type I error is

$$\begin{aligned} & \mathbb{P}_{H_0}[H_0(i,j|k) \text{ rejected for all } k] = \mathbb{P}_{H_0}[\cap_k \{H_0(i,j|k) \text{ rejected}\}] \\ \leq \; & \min_k \mathbb{P}_{H_0}[H_0(i,j|k) \text{ rejected}] \leq \mathbb{P}_{H_0}[H_0(i,j|k^*) \text{ rejected}] \leq \alpha, \end{aligned}$$

where the last inequality follows from the assumption in Proposition 3. □

Proof of Proposition 4: We follow the notation from Section . Consider

$$\hat{\mu}(n)_j = n^{-1}\sum_{i=1}^{n} X_{(n),ij},\ X_{(n),ij} = (\mathbf{X}_{(n),i})_j.$$

By Markov's inequality, for $\gamma > 0$,

$$\mathbb{P}[|\hat{\mu}(n)_j - \mu(n)_j| > \gamma] \leq \gamma^{-4s}\mathbb{E}|n^{-1}\sum_{i=1}^{n} X_{(n),ij} - \mu(n)_j|^{4s},$$

and then by Rosenthal's inequality (cf Petrov, 1975) and our assumption (A1),

$$\mathbb{E}|n^{-1}\sum_{i=1}^{n} X_{(n),ij} - \mu(n)_j|^{4s} \leq Cn^{-2s},$$

where $C > 0$ is a constant independent from j and n. Therefore, for $\gamma > 0$,

$$\mathbb{P}[\max_{1\leq j\leq p_n} |\hat{\mu}(n)_j - \mu(n)_j| > \gamma] \leq p_n\gamma^{-4s}Cn^{-2s} = o(n^{-3s/2}),$$

due to our assumption about p_n, which proves the first claim.

For the second assertion, note that

$$\hat{\Sigma}(n)_{ij} = n^{-1}\sum_{r=1}^{n}(X_{(n),ri} - \hat{\mu}(n)_i)(X_{(n),rj} - \hat{\mu}(n)_j)$$

can be asymptotically replaced by

$$\tilde{\Sigma}(n)_{ij} = n^{-1}\sum_{r=1}^{n}(X_{(n),ri} - \mu(n)_i)(X_{(n),rj} - \mu(n)_j),$$

since by the first assertion of Proposition 4, it can be easily shown that

$$\max_{1\leq i<j\leq p_n} |\hat{\Sigma}(n)_{ij} - \tilde{\Sigma}(n)_{ij}| = o_P(1). \tag{7}$$

Similarly as for the mean, we get for $\gamma > 0$,

$$\mathbb{P}[|\tilde{\Sigma}(n)_{ij} - \Sigma(n)_{ij}| > \gamma] \leq \gamma^{-2s}\mathbb{E}|n^{-1}\sum_{r=1}^{n} Y_r(i,j)|^{2s},$$
$$Y_r(i,j) = (X_{(n),ri} - \mu(n)_i)(X_{(n),rj} - \mu(n)_j) - \Sigma(n)_{ij},$$

DOI: 10.2202/1544-6115.1170

and by Rosenthal's inequality (cf Petrov, 1975) and assumption (A1),

$$\mathbb{E}|n^{-1}\sum_{r=1}^{n} Y_r(i,j)|^{2s} \leq Cn^{-s},$$

where $C > 0$ is a constant, independent of j. Note that our assumption (A1) implies that the moments of order $2s$ of the $Y_r(i,j)$ variables are uniformly bounded. Therefore

$$\mathbb{P}[\max_{1\leq i<j\leq p_n} |\tilde{\Sigma}(n)_{ij} - \Sigma(n)_{ij}| > \gamma] \leq p_n^2 \gamma^{-2s} C n^{-s} = o(1),$$

by our assumption about p_n. This, together with (7) completes the proof for the second assertion of the Proposition. □

Proof of Proposition 5: The first assumption in (A2) and the uniform convergence from Proposition 4 imply that

$$\max_{1\leq i<j\leq p_n} |\hat{\rho}(n)_{ij} - \rho(n)_{ij}| = o_P(1) \ (n \to \infty). \tag{8}$$

Furthermore, we can use a Taylor expansion for the partial correlations:

$$\begin{aligned}\hat{\omega}(n)_{ij|k} - \omega(n)_{ij|k} &= \frac{x - yz}{uv} - \frac{x_0 - y_0 z_0}{u_0 v_0} \\ &= \frac{x - x_0}{u_0 v_0} - \frac{yz - y_0 z_0}{u_0 v_0} - \frac{1}{\tilde{u}^2 \tilde{v}^2}(uv - u_0 v_0)(x - yz),\end{aligned}$$

where $|\tilde{u}\tilde{v} - u_0 v_0| \leq |uv - u_0 v_0|$, and $x = \hat{\rho}(n)_{ij}, y = \hat{\rho}(n)_{ik}, z = \hat{\rho}(n)_{jk}$, $u = \sqrt{1 - \hat{\rho}(n)_{ik}^2}, v = \sqrt{1 - \hat{\rho}(n)_{jk}^2}$ and x_0, y_0, z_0, u_0, v_0 the corresponding true population quantities. We now get the assertion of Proposition 5 by the uniform convergence of the correlations in (8) and by using the second assumption in (A2) which guarantees that the denominator in $1/(u_0 v_0)$ is bounded and that $\frac{1}{\tilde{u}^2\tilde{v}^2} = o_P(1)$ uniformly with respect to i, j, k. □

Proof of Proposition 6: Choose $K = \frac{1}{2}\min(C_1, C_2)$. Then, by Proposition 5 and assumption (A3), the assertion follows. □

References

Barabasi, A. & Albert, R. (1999) Emergence of scaling in random networks. *Science,* **286** (5439), 509–512.

Becker, A., Geiger, D. & Meek, C. (2000) Perfect tree-like markovian distributions. In *UAI* pp. 19–23.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B,* **57**, 289–300.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res,* **11** (8), 1425–33.

Cox, D. R. & Wermuth, N. (1993) Linear dependencies represented by chain graphs (with discussion). *Statist Sci,* **8**, 204–218.

Cox, D. R. & Wermuth, N. (1996) *Multivariate dependencies: models analysis and interpretation.* Chapman & Hall, London.

de Campos, L. & Huete, J. (2000) A new approach for learning belief networks using independence criteria. *Internat J Approx Reasoning,* **24**, 11–37.

de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics,* **20** (18), 3565–3574.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G. & West, M. (2004) Sparse graphical models for exploring gene expression data. *J Mult Analysis,* **90**, 196–212.

Drton, M. & Perlman, M. D. (2004) Model selection for Gaussian Concentration Graphs. *Biometrika,* **91** (3), 591–602.

Edwards, D. (2000) *Introduction to Graphical Modelling.* Springer Verlag; 2nd edition.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000) Using bayesian networks to analyze expression data. *J Comput Biol,* **7** (3-4), 601–620.

Giudici, P. & Green, P. (1999) Decomposable graphical gaussian model determination. *Biometrika,* **86**, 785–801.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pac Symp Biocomput* PSB01 pp. 422–433.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat,* **6**, 65–70.

Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics,* **19** (17), 2271–2282.

Ihmels, J., Levy, R. & Barkai, N. (2004) Principles of transcriptional control in the metabolic network of saccharomyces cerevisiae. *Nat Biotechnol,* **22** (1), 86–92.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000) The large-scale organization of metabolic networks. *Nature,* **407** (6804), 651–654.

Laule, O., Fürholz, A., Chang, H. S., Zhu, T., Wang, X., Heifetz, P. B., Gruissem, W. & Lange, M. (2003) Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in arabidopsis thaliana. *Proc Natl Acad Sci U S A,* **100** (11), 6866–6871.

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

Madigan, D. & Raftery, A. (1994) Model selection and accounting for model uncertainty in graphical models using occam's window. *J Amer Statist Assoc,* **89**, 1535–1546.

Magwene, P. & Kim, J. (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol,* **5** (12), R100.

Maslov, S. & Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science,* **296** (5569), 910–913.

Meinshausen, N. & Bühlmann, P. (2004). High-dimensional graphs and variable selection with the Lasso. To appear in *Ann Stat*.

MIM (2003). Student version 3.1. http://www.hypergraph.dk.

Petrov, V. (1975) *Sums of independent random variables*. Springer, Berlin.

Richardson, T. & Spirtes, P. (2002) Ancestral graph Markov models. *Ann Stat,* **30** (4), 962–1030.

Rodriguez-Concepcion, M., Fores, O., Martinez-Garcia, J. F., Gonzalez, V., Phillips, M., Ferrer, A. & Boronat, A. (2004) Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during arabidopsis seedling development. *Plant Cell,* **16** (1), 144–156.

Roverato, A. (2002) Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scand J Stat,* **29** (3), 391–411.

Simes, R. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika,* **73**, 751–754.

Spirtes, P., Glymour, C. & Scheines, R. (2000) *Causation, Prediction, and Search*. 2nd edition, MIT Press.

Toh, H. & Horimoto, K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics,* **18** (2), 287–297.

Waddell, P. J. & Kishino, H. (2000) Cluster inference methods and graphical models evaluated on nci60 microarray gene expression data. *Genome Informatics,* **11**, 129–140.

Wang, J., Myklebost, O. & Hovig, E. (2003) Mgraph: graphical models for microarray data analysis. *Bioinformatics,* **19** (17), 2210–2211.

Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. & Bühlmann, P. (2004) Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biol,* **5** (11), R92.