Journal of Machine Learning Research 10 (2009) 1187-1238 Submitted 4/08; Revised 11/08; Published 6/09

The Hidden Life of Latent Variables: Bayesian Learning with Mixed Graph Models

Ricardo Silva[*] RICARDO@STATS.UCL.AC.UK
Department of Statistical Science
University College London, WC1E 6BT, UK

Zoubin Ghahramani[†] ZOUBIN@ENG.CAM.AC.UK
Department of Engineering
University of Cambridge, CB2 1PZ, UK

Editor: David Maxwell Chickering

Abstract

Directed acyclic graphs (DAGs) have been widely used as a representation of conditional independence in machine learning and statistics. Moreover, hidden or latent variables are often an important component of graphical models. However, DAG models suffer from an important limitation: the family of DAGs is not closed under marginalization of hidden variables. This means that in general we cannot use a DAG to represent the independencies over a subset of variables in a larger DAG. Directed mixed graphs (DMGs) are a representation that includes DAGs as a special case, and overcomes this limitation. This paper introduces algorithms for performing Bayesian inference in Gaussian and probit DMG models. An important requirement for inference is the specification of the distribution over parameters of the models. We introduce a new distribution for covariance matrices of Gaussian DMGs. We discuss and illustrate how several Bayesian machine learning tasks can benefit from the principle presented here: the power to model dependencies that are generated from hidden variables, but without necessarily modeling such variables explicitly.

Keywords: graphical models, structural equation models, Bayesian inference, Markov chain Monte Carlo, latent variable models

1. Contribution

The introduction of graphical models (Pearl, 1988; Lauritzen, 1996; Jordan, 1998) changed the way multivariate statistical inference is performed. Graphical models provide a suitable language to decompose many complex real-world processes through conditional independence constraints.

Different families of independence models exist. The directed acyclic graph (DAG) family is a particularly powerful representation. Besides providing a language for encoding causal statements (Spirtes et al., 2000; Pearl, 2000), it is in a more general sense a family that allows for non-monotonic independence constraints: that is, models where some independencies can be destroyed by conditioning on new information (also known as the "explaining away" effect — Pearl, 1988), a feature to be expected in many real problems.

∗. Part of this work was done while RS was at the Gatsby Computational Neuroscience Unit, UCL, and at the Statistical Laboratory, University of Cambridge.

†. Also affiliated with the Machine Learning Department, Carnegie Mellon University.

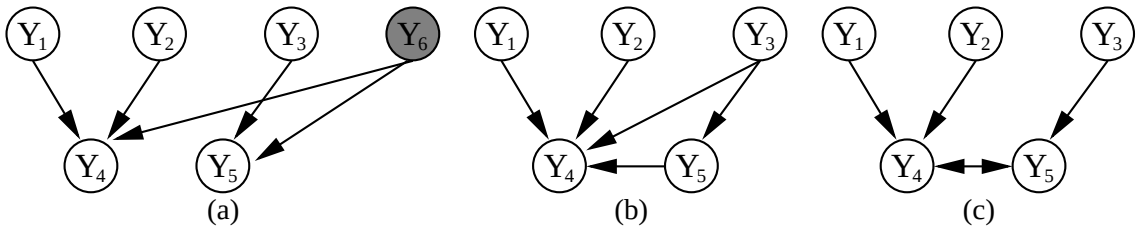


Figure 1: Consider the DAG in (a). Suppose we want to represent the marginal dependencies and independencies that result after marginalizing out Y_6. The simplest resulting DAG (i.e., the one with fewest edges) is depicted in (b). However, notice that this graph does not encode some of the independencies of the original model. For instance, Y_3 and Y_4 are no longer marginally independent in the modified DAGs. A different family of graphical models, encoded with more than one type of edge (directed and *bi-directed*), is the focus of this paper. The graph in (c) depicts the solution using this "mixed" representation.

However, DAG independence models have an undesirable feature: they are not closed under marginalization, as we will illustrate. Consider the regression problem where we want to learn the effect of a cocktail of two drugs for blood pressure, while controlling for a chemotherapy treatment of liver cancer. We refer to Y_1, Y_2 as the dosage for the blood pressure drugs, Y_3 as a measure of chemotherapy dosage, Y_4 as blood pressure, and Y_5 as an indicator of liver status. Moreover, let Y_6 be an hidden physiological factor that affects both blood pressure and liver status. It is assumed that the DAG corresponding to this setup is given by Figure 1(a).

In this problem, predictions concerning Y_6 are irrelevant: what we care is the marginal for $\{Y_1, \ldots, Y_5\}$. Ideally, we want to take such irrelevant hidden variables out of the loop. Yet the set of dependencies within the marginal for $\{Y_1, \ldots, Y_5\}$ cannot be efficiently represented as a DAG model. If we remove the edge $Y_3 \rightarrow Y_4$ from Figure 1(b), one can verify this will imply a model where Y_3 and Y_4 are independent given Y_5, which is not true in our original model. To avoid introducing unwanted independence constraints, a DAG such as the one in Figure 1(b) will be necessary. Notice that in general this will call for extra dependencies that did not exist originally (such as Y_3 and Y_4 now being marginally dependent). Not only learning from data will be more difficult due to the extra dependencies, but specifying prior knowledge on the parameters becomes less intuitive and therefore more error prone.

In general, it will be the case that variables of interest have hidden common causes. This puts the researcher using DAGs in a difficult position: if she models only the marginal comprising the variables of interest, the DAG representation might not be suitable anymore. If she includes all hidden variables for the sake of having the desirable set of independencies, extra assumptions about hidden variables will have to be taken into account. In this sense, the DAG representation is flawed. There is a need for a richer family of graphical models, for which *mixed graphs* are an answer.

Directed mixed graphs (DMGs) are graphs with directed and bi-directed edges. In particular, acyclic directed mixed graphs (ADMGs) have no directed cycle, that is, no sequence of directed edges $X \rightarrow \cdots \rightarrow X$ that starts and ends on the same node. Such a representation encodes a set of conditional independencies among random variables, which can be read off a graph by using a

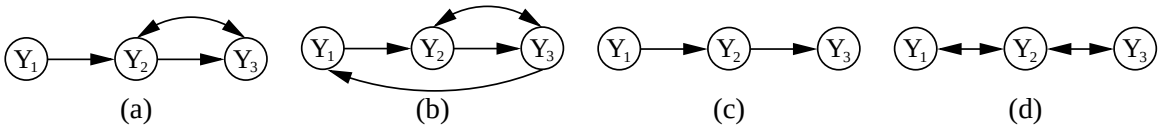


Figure 2: Different examples of directed mixed graphs. The graph in (b) is cyclic, while all others are acyclic. A subgraph of two variables where both edges $Y_1 \rightarrow Y_2$ and $Y_1 \leftrightarrow Y_2$ are present is sometimes known as a "bow pattern" (Pearl, 2000) due to its shape.

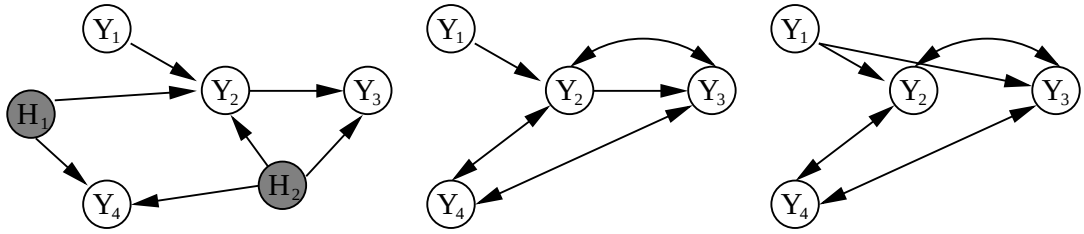


Figure 3: After marginalizing variables H_1 and H_2 from the DAG on the left, one possible DMG representation of the same dependencies is shown by the graph in the middle. Notice that there are multiple DMGs within a same Markov equivalence class, that is, encoding the same set of conditional independencies (Richardson and Spirtes, 2002). The two last graphs above are on the same class.

criterion known as m-separation, a natural extension of the d-separation criterion used for directed acyclic graphs (Richardson, 2003).

In a ADMG, two adjacent nodes might be connected by up to two edges, where in this case one has to be bi-directed and the other directed. A cyclic model can in principle allow for two directed edges of opposite directions. Figure 2 provides a few examples of DMGs. The appeal of this graphical family lies on the representation of the marginal independence structure among a set of observed variables, assuming they are part of a larger DAG structure that includes hidden variables. This is illustrated in Figure 3.[1] More details on DMGs are given in Sections 2 and 8. In our blood pressure\liver status multiple regression problem, the suitable directed mixed graph is depicted in Figure 1(c).

The contribution of this paper is how to perform Bayesian inference on two different families of mixed graph models: Gaussian and probit. Markov chain Monte Carlo (MCMC) and variational approximations will be discussed. Current Bayesian inference approaches for DMG models have limitations, as discussed in Section 2, despite the fact that such models are widely used in several sciences.

The rest of the paper is organized as follows. Section 3 describes a special case of Gaussian mixed graph models, where only bi-directed edges are allowed. Priors and a Monte Carlo algorithm are described. This case will be a building block for subsequent sections, such as Section 4, where

1. Notice that it is not necessarily the case that the probability model itself is closed under marginalization. This will happen to some models, including the Gaussian model treated in this paper. But the basic claim of closure concerns the graph, that is, the representation of independence constraints.

Gaussian DMG models are treated. Section 5 covers a type of discrete distribution for binary and ordinal data that is Markov with respect to an acyclic DMG. In Section 6 we discuss more sophisticated algorithms that are useful for scaling up Bayesian learning to higher-dimensional problems. Section 7 presents several empirical studies. Since the use of mixed graph models in machine learning applications is still in its early stages, we briefly describe in Section 8 a variety of possible uses of such graphs in machine learning applications.

2. Basics of DMGs, Gaussian Models and Related Work

In this section, we describe the Gaussian DMG model and how it complements latent variable models. At the end of the section, we also discuss a few alternative approaches for the Bayesian inference problem introduced in this paper.

2.1 Notation and Terminology

In what follows, we will use standard notions from the graphical modeling literature, such as vertex (node), edge, parent, child, ancestor, descendant, DAG, undirected graph, induced subgraph, Markov condition and d-separation. Refer to Pearl (1988) and Lauritzen (1996) for the standard definitions if needed. Less standard definitions will be given explicitly when appropriate. A useful notion is that of m-separation (Richardson, 2003) for reading off which independencies are entailed by a DMG representation. This can be reduced to d-separation (Pearl, 1988) by the following trick: for each bi-directed edge $Y_i \leftrightarrow Y_j$, introduce a new hidden variable X_{ij} and the edges $X_{ij} \rightarrow Y_i$ and $X_{ij} \rightarrow Y_j$. Remove then all bi-directed edges and apply d-separation to the resulting directed graph.

As usual, we will refer to vertices (nodes) in a graph and the corresponding random variables in a distribution interchangeably. Data points are represented by vectors with an upper index, such as $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}$. The variable corresponding to node Y_i in data point $\mathbf{Y}^{(j)}$ is represented by $Y_i^{(j)}$.

2.2 Gaussian Parameterization

The origins of mixed graph models can be traced back to Sewall Wright (Wright, 1921), who used special cases of mixed graph representations in genetic studies. Generalizing Wright's approach, many scientific fields such as psychology, social sciences and econometrics use linear mixed graph models under the name of *structural equation models* (Bollen, 1989). Only recently the graphical and parametrical aspects of mixed graph models have been given a thorough theoretical treatment (Richardson and Spirtes, 2002; Richardson, 2003; Kang and Tian, 2005; Drton and Richardson, 2008a). In practice, many structural equation models today are Gaussian models. We will work under this assumption unless stated otherwise.

For a DMG $\mathcal{G}$ with a set of vertices $\mathbf{Y}$, a standard parameterization of the Gaussian model is given as follows. For each variable Y_i with a (possibly empty) parent set $\{Y_{i1}, \ldots, Y_{ik}\}$, we define a "structural equation"

$$Y_i = \alpha_i + b_{i1}Y_{i1} + b_{i2}Y_{i2} + \cdots + b_{ik}Y_{ik} + \varepsilon_i$$

where ε_i is a Gaussian random variable with zero mean and variance v_{ii}. Notice that this parameterization allows for cyclic models.

Unlike in standard Gaussian DAG models, the error terms $\{\varepsilon_i\}$ are not necessarily mutually independent. Independence is asserted by the graphical structure: given two vertices Y_i and Y_j,

the respective error terms ε_i and ε_j are marginally independent if Y_i and Y_j are not connected by a bi-directed edge.

By this parameterization, each directed edge $Y_i \leftarrow Y_j$ in the graph corresponds to a parameter b_{ij}. Each bi-directed edge $Y_i \leftrightarrow Y_j$ in the graph is associated with a covariance parameter v_{ij}, the covariance of ε_i and ε_j. Each vertex Y_j in the graph is associated with variance parameter v_{jj}, the variance of ε_j. Algebraically, let $\mathbf{B}$ be a $m \times m$ matrix, m being the number of observed variables. This matrix is such that $(\mathbf{B})_{ij} = b_{ij}$ if $Y_i \leftarrow Y_j$ exists in the graph, and 0 otherwise. Let $\mathbf{V}$ be a $m \times m$ matrix, where $(\mathbf{V})_{ij} = v_{ij}$ if $i = j$ or if $Y_i \leftrightarrow Y_j$ is in the graph, and 0 otherwise. Let $\mathbf{Y}$ be the column vector of observed variables, α the column vector of intercept parameters, and ε be the corresponding vector of error terms. The set of structural equations can be given in matrix form as

$$\begin{aligned} & \mathbf{Y} = \mathbf{B}\mathbf{Y} + \alpha + \varepsilon \Rightarrow \mathbf{Y} = (\mathbf{I} - \mathbf{B})^{-1}(\varepsilon + \alpha) \\ \Rightarrow \quad & \Sigma(\Theta) = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{V}(\mathbf{I} - \mathbf{B})^{-\mathsf{T}} \end{aligned} \tag{1}$$

where $\mathbf{A}^{-\mathsf{T}}$ is the transpose of $\mathbf{A}^{-1}$ and $\Sigma(\Theta)$ is the *implied covariance matrix* of the model, $\Theta \equiv \{\mathbf{B}, \mathbf{V}, \alpha\}$.

2.2.1 COMPLETENESS OF PARAMETERIZATION AND ANCESTRAL GRAPHS

An important class of ADMGs is the directed ancestral graph. Richardson and Spirtes (2002) provide the definition and a thorough account of the Markov properties of ancestral graphs. One of the reasons for the name "ancestral graph" is due to one of its main properties: if there is a directed path $Y_i \rightarrow \cdots \rightarrow Y_j$, that is, if Y_i is an ancestor of Y_j, then there is no bi-directed edge $Y_i \leftrightarrow Y_j$. Thus directed ancestral graphs are ADMGs with this constraint.[2]

In particular, they show that any Gaussian distribution that is Markov with respect to a given ADMG can be represented by some Gaussian ancestral graph model that is parameterized as above. For the ancestral graph family, the given parameterization is *complete*: that is, for each Markov equivalence class, it is always possible to choose an ancestral graph where the resulting parameterization imposes no further constraints on the distribution besides the independence constraints of the class. Since the methods described in this paper apply to general DMG models, they also apply to directed ancestral graphs.

In principle, it is possible to define and parameterize a Gaussian DAG model that entails exactly the same independence constraints encoded in an directed ancestral graph. One possibility, as hinted in the previous Section, is to replace each bi-directed edge $Y_i \leftrightarrow Y_j$ by a new path $Y_i \leftarrow X_{ij} \rightarrow Y_j$. Variables $\{X_{ij}\}$ are "ancillary" hidden variables, in the sense that they are introduced for the sake of obtaining the same independence constraints of an ancestral graph. Standard Bayesian methodology can then be applied to perform inference in this Gaussian DAG model.

However, this parameterization might have undesirable consequences, as discussed in Section 8.6 of Richardson and Spirtes (2002). Moreover, when Markov chain Monte Carlo algorithms are applied to compute posteriors, the "ancillary" hidden variables will have to be integrated out numerically. The resulting Markov chain can suffer from substantial autocorrelation when compared to a model with no ancillary variables. We illustrate this behavior in Section 7.

Further constraints beyond independence constraints are certainly desirable depending on the context. For instance, general ADMGs that are not ancestral graphs may impose other constraints (Richardson and Spirtes, 2002), and such graphs can still be sensible models of, for example, the

2. Notice this rules out the possibility of having both edges $Y_i \rightarrow Y_j$ and $Y_i \leftrightarrow Y_j$ in the same ancestral graph.

causal processes for the problem at hand. When many observed variables are confounded by a same hidden common cause, models based on factor analysis are appropriate (Silva et al., 2006). However, it is useful to be able to build upon independence models that are known to have a complete parameterization. In any case, even the latent variables in any model might have dependencies that arise from other latent variables that were marginalized, and a latent variable ADMG model will be necessary. When it comes to solving a problem, it is up to the modeler (or learning algorithm) to decide if some set of latent variables should be included, or if they should be implicit, living their hidden life through the marginals.

Richardson and Spirtes (2002) provide further details on the advantages of a complete parameterization. Drton and Richardson (2004) provide an algorithm for fitting Gaussian ancestral graph models by maximum likelihood.

2.3 Bayesian Inference

The literature on Bayesian structural equation models is extensive. Scheines et al. (1999) describe one of the first approaches, including ways of testings such models. Lee (2007) provides details on many recent advances. Standard Bayesian approaches for Gaussian DMG models rely on either attempting to reduce the problem to inference with DAG models, or on using rejection sampling.

In an application described by Dunson et al. (2005), the "ancillary latent" trick is employed, and Gibbs sampling for Gaussian DAG models is used. This parameterization has the disadvantages mentioned in the previous section. Scheines et al. (1999) use the complete parameterization, with a single parameter corresponding to each bi-directed edge. However, the global constraint of positive-definiteness in the covariance matrix is enforced only by rejection sampling, which might be inefficient in models with moderate covariance values. The prior is setup in an indirect way. A Gaussian density function is independently defined for each error covariance v_{ij}. The actual prior, however, is the result of multiplying all of such functions and the indicator function that discards non-positive definite matrices, which is then renormalized.

In contrast, the Bayesian approach delineated in the next sections uses the complete parameterization, does not appeal to rejection sampling, makes use of a family of priors which we believe is the natural choice for the problem, and leads to convenient ways of computing marginal likelihoods for model selection. We will also see that empirically they lead to much better behaved Markov chain Monte Carlo samplers when compared to DAGs with ancillary latent variables.

3. Gaussian Models of Marginal Independence

This section concerns priors and sampling algorithms for zero-mean Gaussian models that are Markov with respect to a bi-directed graph, that is, a DMG with no directed edges. Focusing on bi-directed graphs simplifies the presentation, while providing a convenient starting point to solve the full DMG case in the sequel.

Concerning the notation: the distribution we introduce in this section is a distribution over covariance matrices. In the interest of generality, we will refer to the random matrix as Σ. In the context of the previous section, $\Sigma \equiv \Sigma(\Theta) = \mathbf{V}$, since we are assuming $\mathbf{B} = 0, \alpha = 0$.

3.1 Priors

Gaussian bi-directed graph models are sometimes called *covariance graph models*. Covariance graphs are models of marginal independence: each edge corresponds to a single parameter in the covariance matrix (the corresponding covariance); the absence of an edge $Y_i \leftrightarrow Y_j$ is a statement that $\sigma_{Y_iY_j} = 0$, σ_{XY} being the covariance of random variables X and Y. More precisely, if Σ is a random covariance matrix generated by a covariance model, a distribution of Σ is the distribution over the (non-repeated) entries corresponding to variances and covariances of adjacent nodes.[3]

In a model with a fully connected bi-directed graph, this reduces to a space of unrestricted covariance matrices. A common distribution for covariance matrices is the inverse Wishart $IW(\delta, \mathbf{U})$. In this paper, we adopt the following inverse Wishart parameterization:

$$p(\Sigma) \propto |\Sigma|^{-(\delta+2m)/2} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1}\mathbf{U})\right\}, \Sigma \text{ positive definite},$$

$p(\cdot)$ being the density function, $tr(\cdot)$ the trace function, and m the number of variables (nodes) in our model.[4] We will overload the symbol $p(\cdot)$ wherever it is clear from the context which density function we are referring to. It is assumed that $\delta > 0$ and $\mathbf{U}$ is positive definite.

Following Atay-Kayis and Massam (2005), let $M^+(\mathcal{G})$ be the cone of positive definite matrices such that, for a given bi-directed graph $\mathcal{G}$ and $\Sigma \in M^+(\mathcal{G})$, $\sigma_{ij} = 0$ if nodes Y_i and Y_j are not adjacent in $\mathcal{G}$. It is convenient to choose a distribution that is conjugate to the Gaussian likelihood function, since one can use the same algorithms for performing inference both in the prior and posterior. In a zero-mean Gaussian model, the likelihood function for a fixed data set $\mathcal{D} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}\}$ is defined by the sufficient statistic $\mathbf{S} = \sum_{d=1}^{n} (\mathbf{Y}^{(d)})(\mathbf{Y}^{(d)})^\mathsf{T}$ as follows:

$$\mathcal{L}(\Sigma; \mathcal{D}) = (2\pi)^{-nm/2} |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1}\mathbf{S})\right\}. \tag{2}$$

We extend the inverse Wishart distribution to the case of constrained covariance matrices in order to preserve conjugacy. This define the following distribution:

$$p(\Sigma) = \frac{1}{I_{\mathcal{G}}(\delta, \mathbf{U})} |\Sigma|^{-(\delta+2m)/2} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1}\mathbf{U})\right\}, \Sigma \in M^+(\mathcal{G}) \tag{3}$$

which is basically a re-scaled inverse Wishart prior with a different support and, consequently, different normalizing constant $I_{\mathcal{G}}(\delta, \mathbf{U})$. An analogous concept exists for undirected graphs, where $\Sigma^{-1} \in M^+(\mathcal{G})$ is given a Wishart-like prior: the "$\mathcal{G}$-Wishart" distribution (Atay-Kayis and Massam, 2005). We call the distribution with density function defined as in Equation (3) the *$\mathcal{G}$-Inverse Wishart* distribution ($\mathcal{G}$-*IW*). It will be the basis of our framework. There are no analytical formulas for the normalizing constant.

3. As such, the density function for Σ is defined with respect to the Lebesgue measure of the non-zero, independent elements of this matrix.
4. We adopt this non-standard parameterization of the inverse Wishart because it provides a more convenient reparameterization used in the sequel. Notice this is the parameterization used by Brown et al. (1993) and Atay-Kayis and Massam (2005), which developed other distributions for covariance matrices.

3.2 The Normalizing Constant

We now derive a Monte Carlo procedure to compute $I_{\mathcal{G}}(\delta, \mathbf{U})$. In the sequel, this will be adapted into an importance sampler to compute functionals of a $\mathcal{G}$-*IW* distribution. The core ideas are also used in a Gibbs sampler to obtain samples from its posterior.

The normalizing constant is essential for model selection of covariance graphs. By combining the likelihood equation (2) with the prior (3), we obtain the joint

$$p(\mathcal{D}, \Sigma \mid \mathcal{G}) = (2\pi)^{-\frac{nm}{2}} I_{\mathcal{G}}(\delta, \mathbf{U})^{-1} \times |\Sigma|^{-\frac{\delta+2m+n}{2}} \exp\left\{-\frac{1}{2} tr[\Sigma^{-1}(\mathbf{S}+\mathbf{U})]\right\}$$

where we make the dependency on the graphical structure $\mathcal{G}$ explicit. By the definition of $I_{\mathcal{G}}$, integrating Σ out of the above equation implies the following marginal likelihood:

$$p(\mathcal{D} \mid \mathcal{G}) = \frac{1}{(2\pi)^{\frac{nm}{2}}} \frac{I_{\mathcal{G}}(\delta+n, \mathbf{S}+\mathbf{U})}{I_{\mathcal{G}}(\delta, \mathbf{U})}$$

from which a posterior $\mathcal{P}(\mathcal{G} \mid \mathcal{D})$ can be easily derived as a function of quantities of the type $I_{\mathcal{G}}(\cdot, \cdot)$.

The normalizing constant $I_{\mathcal{G}}(\delta, \mathbf{U})$ is given by the following integral:[5]

$$I_{\mathcal{G}}(\delta, \mathbf{U}) = \int_{M^+(\mathcal{G})} |\Sigma|^{-\frac{\delta+2m}{2}} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1}\mathbf{U})\right\} d\Sigma. \tag{4}$$

The space $M^+(\mathcal{G})$ can be described as the space of positive definite matrices conditioned on the event that each matrix has zero entries corresponding to non-adjacent nodes in graph $\mathcal{G}$. We will reduce the integral (4) to an integral over random variables we know how to sample from. The given approach follows the framework of Atay-Kayis and Massam (2005) using the techniques of Drton and Richardson (2003).

Atay-Kayis and Massam (2005) show how to compute the marginal likelihood of non-decomposable undirected models by reparameterizing the precision matrix through the Cholesky decomposition. The zero entries in the inverse covariance matrix of this model correspond to constraints in this parameterization, where part of the parameters can be sampled independently and the remaining parameters calculated from the independent ones.

We will follow a similar framework but with a different decomposition. It turns out that the Cholesky decomposition does not provide an easy reduction of (4) to an integral over canonical, easy to sample from, distributions. We can, however, use Bartlett's decomposition to achieve this reduction.

3.2.1 BARTLETT'S DECOMPOSITION

Before proceeding, we will need a special notation for describing sets of indices and submatrices.

Let $\{i\}$ represent the set of indices $\{1, 2, \ldots, i\}$. Let $\Sigma_{i,\{i-1\}}$ be the row vector containing the covariance between Y_i and all elements of $\{Y_1, Y_2, \ldots, Y_{i-1}\}$. Let $\Sigma_{\{i-1\},\{i-1\}}$ be the marginal covariance matrix of $\{Y_1, Y_2, \ldots, Y_{i-1}\}$. Let σ_{ii} be the variance of Y_i. Define the mapping

$$\Sigma \rightarrow \Phi \equiv \{\gamma_1, \mathcal{B}_2, \gamma_2, \mathcal{B}_3, \gamma_3, \ldots, \mathcal{B}_m, \gamma_m\},$$

5. Notice this integral is always finite for any choice of $\delta > 0$ and positive definite $\mathbf{U}$, since it is no greater than the normalizing constant of the inverse Wishart.

such that $\mathcal{B}_i$ is a row vector with $i-1$ entries, γ_i is a scalar, and

$$\begin{array}{rcll} \gamma_1 & = & \sigma_{11}, & \\ \mathcal{B}_i & = & \Sigma_{i,\{i-1\}}\Sigma^{-1}_{\{i-1\},\{i-1\}}, & i>1, \\ \gamma_i & = & \sigma_{ii.\{i-1\},\{i-1\}} \equiv \sigma_{ii} - \Sigma_{i,\{i-1\}}\Sigma^{-1}_{\{i-1\},\{i-1\}}\Sigma_{\{i-1\},i}, & i>1. \end{array} \tag{5}$$

The set Φ provides a parameterization of Σ, in the sense that the mapping (5) is bijective. Given that $\sigma_{11} = \gamma_1$, the inverse mapping is defined recursively by

$$\begin{array}{rcl} \Sigma_{i,\{i-1\}} & = & \mathcal{B}_i\Sigma_{\{i-1\},\{i-1\}},\ i>1, \\ \sigma_{ii} & = & \gamma_i + \mathcal{B}_i\Sigma_{\{i-1\},i},\ i>1. \end{array} \tag{6}$$

We call the set $\Phi \equiv \{\gamma_1, \mathcal{B}_2, \gamma_2, \mathcal{B}_3, \gamma_3, \ldots, \mathcal{B}_m, \gamma_m\}$ the *Bartlett parameters* of Σ, since the decomposition (6) is sometimes known as Bartlett's decomposition (Brown et al., 1993).

For a random inverse Wishart matrix, Bartlett's decomposition allows the definition of its density function by the joint density of $\{\gamma_1, \mathcal{B}_2, \gamma_2, \mathcal{B}_3, \gamma_3, \ldots, \mathcal{B}_m, \gamma_m\}$. Define $\mathbf{U}_{\{i-1\},\{i-1\}}$, $\mathbf{U}_{\{i-1\},i}$ and $u_{ii.\{i-1\},\{i-1\}}$ in a way analogous to the Σ definitions. The next lemma follows directly from Lemma 1 of Brown et al. (1993):

Lemma 1 *Suppose Σ is distributed as $IW(\delta, \mathbf{U})$. Then the distribution of the corresponding Bartlett parameters $\Phi \equiv \{\gamma_1, \mathcal{B}_2, \gamma_2, \mathcal{B}_3, \gamma_3, \ldots, \mathcal{B}_m, \gamma_m\}$ is given by:*

1. *γ_i is independent of $\Phi \backslash \{\gamma_i, \mathcal{B}_i\}$*
2. *$\gamma_i \sim IG((\delta+i-1)/2, u_{ii.\{i-1,i-1\}}/2)$, where $IG(\alpha,\beta)$ is the inverse gamma distribution*
3. *$\mathcal{B}_i \mid \gamma_i \sim N(\mathbf{U}^{-1}_{\{i-1\},\{i-1\}}\mathbf{U}_{\{i-1\},i}, \gamma_i\mathbf{U}^{-1}_{\{i-1\},\{i-1\}})$, where $N(\mathbf{M},\mathbf{C})$ is a multivariate Gaussian distribution and $\mathbf{U}^{-1}_{\{i-1\},\{i-1\}} \equiv (\mathbf{U}_{\{i-1\},\{i-1\}})^{-1}$.*

3.2.2 BARTLETT'S DECOMPOSITION OF MARGINAL INDEPENDENCE MODELS

What is interesting about Bartlett's decomposition is that it provides a simple parameterization of the inverse Wishart distribution with variation independent parameters. This decomposition allows the derivation of new distributions. For instance, Brown et al. (1993) derive a "Generalized Inverted Wishart" distribution that allows one to define different degrees of freedom for different submatrices of an inverse Wishart random matrix. For our purposes, Bartlett's decomposition can be used to reparameterize the *G-IW* distribution. For that, one needs to express the independent elements of Σ in the space of Bartlett parameters.

The original reparameterization maps Σ to $\Phi \equiv \{\gamma_1, \mathcal{B}_2, \gamma_2, \mathcal{B}_3, \gamma_3, \ldots, \mathcal{B}_d, \gamma_d\}$. To impose the constraint that Y_i and Y_j are uncorrelated, for $i>j$, is to set $\left(\mathcal{B}_i\Sigma_{\{i-1\},\{i-1\}}\right)_j = \sigma_{Y_iY_j}(\Phi) = 0$. For a fixed $\Sigma_{\{i-1\},\{i-1\}}$, this implies a constraint on $(\mathcal{B}_i)_j \equiv \beta_{ij}$.

Following the terminology used by Richardson and Spirtes (2002), let a *spouse* of node Y in a mixed graph be any node adjacent to Y by a bi-directed edge. The set of spouses of Y_i is denoted by $sp(i)$. The set of spouses of Y_i *according to order* $Y_1, Y_2, \ldots, Y_m$ is defined by $sp_\prec(i) \equiv sp(i) \cap \{Y_1, \ldots, Y_{i-1}\}$. The set of non-spouses of Y_i is denoted by $nsp(i)$. Analogously, $nsp_\prec(i) \equiv \{Y_1, \ldots, Y_{i-1}\} \backslash sp_\prec(i)$. Let $\mathcal{B}_{i,sp_\prec(i)}$ be the subvector of $\mathcal{B}_i$ corresponding to the the respective spouses of Y_i. Define $\mathcal{B}_{i,nsp_\prec(i)}$ analogously.

Given the constraint $\mathcal{B}_i\Sigma_{\{i-1\},nsp_\prec(i)} = 0$, it follows that

$$\mathcal{B}_{i,sp_\prec(i)}\Sigma_{sp_\prec(i),nsp_\prec(i)} + \mathcal{B}_{i,nsp_\prec(i)}\Sigma_{nsp_\prec(i),nsp_\prec(i)} = 0 \Rightarrow$$

$$\mathcal{B}_{i,nsp_\prec(i)} = -\mathcal{B}_{i,sp_\prec(i)}\Sigma_{sp_\prec(i),nsp_\prec(i)}\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}. \tag{7}$$

Identity (7) was originally derived by Drton and Richardson (2003). A property inherited from the original decomposition for unconstrained matrices is that $\mathcal{B}_{i,sp_\prec(i)}$ is functionally independent of $\Sigma_{\{i-1\},\{i-1\}}$. From (7), we obtain that the free Bartlett parameters of Σ are $\Phi_\mathcal{G} \equiv \{\gamma_1, \mathcal{B}_{2,sp_\prec(2)}, \gamma_2, \mathcal{B}_{3,sp_\prec(3)}, \gamma_3, \ldots, \mathcal{B}_{m,sp_\prec(m)}, \gamma_m\}$.

Notice that, according to (5), Φ corresponds to the set of parameters of a fully connected, zero-mean, Gaussian DAG model. In such a DAG, Y_i is a child of $\{Y_1, \ldots, Y_{i-1}\}$, and

$$Y_i = \mathcal{B}_i\mathbf{Y}_{i-1} + \zeta_j, \qquad \zeta_j \sim N(0, \gamma_j)$$

where $\mathbf{Y}_{i-1}$ is the $(i-1)\times 1$ vector corresponding to $\{Y_1, \ldots, Y_{i-1}\}$.

As discussed by Drton and Richardson (2003), this interpretation along with Equation (7) implies

$$Y_i = \mathcal{B}_{i,sp_\prec(i)}\mathbf{Z}_i + \zeta_j \tag{8}$$

where the entries in $\mathbf{Z}_i$ are the corresponding residuals of the regression of $sp_\prec(i)$ on $nsp_\prec(i)$.

The next step in solving integral (4) is to find the Jacobian $J(\Phi_\mathcal{G})$ of the transformation $\Sigma \to \Phi_\mathcal{G}$. This is given by the following Lemma:

Lemma 2 *The determinant of the Jacobian for the change of variable $\Sigma \to \Phi_\mathcal{G}$ is*

$$|J(\Phi_\mathcal{G})| = \prod_{i=2}^{m}|R_i| = \frac{1}{\prod_{i=2}^{m}|\Sigma_{nsp_{\prec(i)},nsp_{\prec(i)}}|}\prod_{i=1}^{m-1}\gamma_i^{m-i}$$

where $R_i \equiv \Sigma_{sp_{\prec(i)},sp_{\prec(i)}} - \Sigma_{sp_{\prec(i)},nsp_{\prec(i)}}\Sigma^{-1}_{nsp_{\prec(i)},nsp_{\prec(i)}}\Sigma_{nsp_{\prec(i)},sp_{\prec(i)}}$, that is, the covariance matrix of the respective residual $\mathbf{Z}_i$ (as parameterized by $\Phi_\mathcal{G}$). If $nsp_{\prec(i)} = \emptyset$, R_i is defined as $\Sigma_{sp_{\prec(i)},sp_{\prec(i)}}$ and $|\Sigma_{nsp_{\prec(i)},nsp_{\prec(i)}}|$ is defined as 1.

The proof of this Lemma is in Appendix C. A special case is the Jacobian of the unconstrained covariance matrix (i.e., when the graph has no missing edges):

$$|J(\Phi)| = \prod_{i=1}^{m-1}\gamma_i^{m-i}. \tag{9}$$

Now that we have the Jacobian, the distribution over Bartlett's parameters given by Lemma 1, and the identities of Drton and Richardson (2003) given in Equation (7), we have all we need to provide a Monte Carlo algorithm to compute the normalizing constant of a $\mathcal{G}$-*IW* with parameters $(\delta, \mathbf{U})$.

Let $\Sigma(\Phi_\mathcal{G})$ be the implied covariance matrix given by our set of parameters $\Phi_\mathcal{G}$. We start from the integral in (4), and rewrite it as a function of $\Phi_\mathcal{G}$. This can be expressed by substituting Σ for $\Sigma(\Phi_\mathcal{G})$ and multiplying the integrand by the determinant of the Jacobian. Notice that the parameters in $\Sigma(\Phi_\mathcal{G})$ are variation independent: that is, their joint range is given by the product of

their individual ranges (positive reals for the γ variables and the real line for the β coefficients). This range will replace the original $M^+(\mathcal{G})$ space, which we omit below for simplicity of notation:

$$I_{\mathcal{G}}(\delta, \mathbf{U}) = \int |J(\Phi_{\mathcal{G}})||\Sigma(\Phi_{\mathcal{G}})|^{-\frac{\delta+2m}{2}} \exp\left\{-\frac{1}{2} tr(\Sigma(\Phi_{\mathcal{G}})^{-1}\mathbf{U})\right\} d\Phi_{\mathcal{G}}.$$

We now multiply and divide the above expression by the normalizing constant of an inverse Wishart $(\delta, \mathbf{U})$, which we denote by $I_{IW}(\delta, \mathbf{U})$:

$$I_{\mathcal{G}}(\delta, \mathbf{U}) = I_{IW}(\delta, \mathbf{U}) \int |J(\Phi_{\mathcal{G}})| \times I_{IW}^{-1}(\delta, \mathbf{U})|\Sigma(\Phi_{\mathcal{G}})|^{-\frac{\delta+2m}{2}} \exp\left\{-\frac{1}{2} tr(\Sigma(\Phi_{\mathcal{G}})^{-1}\mathbf{U})\right\} d\Phi_{\mathcal{G}}. \quad (10)$$

The expression

$$I_{IW}^{-1}(\delta, \mathbf{U})|\Sigma|^{-\frac{\delta+2m}{2}} \exp\left\{-\frac{1}{2} tr(\Sigma^{-1}\mathbf{U})\right\}$$

corresponds to the density function of an inverse Wishart Σ. Lemma 1 allows us to rewrite the inverse Wishart density function as the density of Bartlett parameters, but this is assuming no independence constraints. We can easily reuse the result of Lemma 1 as follows:

1. write the density of the inverse Wishart as the product of gamma-normal densities given in Lemma 1;

2. this expression contains the original Jacobian determinant $|J(\Phi)|$. We have to remove it, since we are plugging in our own Jacobian determinant. Hence, we divide the reparameterized density by the expression in Equation (9).

   This ratio $|J(\Phi_{\mathcal{G}})|/|J(\Phi)|$ can be rewritten as

   $$\frac{|J(\Phi_{\mathcal{G}})|}{|J(\Phi)|} = \prod_{i=1}^{m} \frac{|R_i|}{\gamma_i^{m-i}} = \frac{1}{\prod_{i=2}^{m} |\Sigma_{nsp_{\prec(i)}, nsp_{\prec(i)}}|}$$

   where $|\Sigma_{nsp_{\prec(i)}, nsp_{\prec(i)}}| \equiv 1$ if $nsp_{\prec}(i) = \emptyset$;

3. substitute each vector $\mathcal{B}_{i, nsp_{\prec}(i)}$, which is not a free parameter, by the corresponding expression $-\mathcal{B}_{i, sp_{\prec}(i)} \Sigma_{sp_{\prec}(i), nsp_{\prec}(i)} \Sigma^{-1}_{nsp_{\prec}(i), nsp_{\prec}(i)}$.

This substitution takes place into the original factors given by Bartlett's decomposition, as introduced in Lemma 1:

$$\begin{aligned} p(\mathcal{B}_i, \gamma_i) &= (2\pi)^{-(i-1)/2} \gamma_i^{-(i-1)/2} |\mathbf{U}_{\{i-1\},\{i-1\}}|^{1/2} \\ &\times \exp\left(-\frac{1}{2\gamma_i} (\mathcal{B}_i^{\mathsf{T}} - \mathbf{M}_i)^{\mathsf{T}} \mathbf{U}_{\{i-1\},\{i-1\}} (\mathcal{B}_i^{\mathsf{T}} - \mathbf{M}_i)\right) \\ &\times \frac{(u_{ii.\{i-1\},\{i-1\}}/2)^{(\delta+i-1)/2}}{\Gamma((\delta+i-1)/2)} \gamma_i^{-\left(\frac{\delta+i-1}{2}+1\right)} \exp\left(-\frac{1}{2\gamma_i} u_{ii.\{i-1\},\{i-1\}}\right) \end{aligned} \quad (11)$$

where $\mathbf{M}_i \equiv \mathbf{U}^{-1}_{\{i-1\},\{i-1\}} \mathbf{U}_{\{i-1\},i}$. Plugging in this in (10) results in

$$I_{\mathcal{G}}(\delta,\mathbf{U}) = I_{IW}(\delta,\mathbf{U}) \int \frac{1}{\prod_{i=2}^{m} |\Sigma_{nsp_{\prec}(i),nsp_{\prec}(i)}|} \times p(\gamma_1) \prod_{i=2}^{m} p(\mathcal{B}_i,\gamma_i)\, d\Phi_{\mathcal{G}}.$$

However, after substitution, each factor $p(\mathcal{B}_i,\gamma_i)$ is not in general a density function for $\{\mathcal{B}_{i,sp_{\prec}(i)}, \gamma_i\}$ and will include also parameters $\{\mathcal{B}_{j,sp_{\prec}(j)},\gamma_j\}, j<i$. Because of the non-linear relationships that link Bartlett parameters in a marginal independence model, we cannot expect to reduce this expression to a tractable distribution we can easily sample from. Instead, we rewrite each original density factor $p(\mathcal{B}_i,\gamma_i)$ such that it includes all information about $\mathcal{B}_{i,sp_{\prec}(i)}$ and γ_i within a canonical density function. That is, factorize $p(\mathcal{B}_i,\gamma_i)$ as

$$p(\mathcal{B}_i,\gamma_i|\Phi_{i-1}) = p_b(\mathcal{B}_{i,sp_{\prec}(i)}|\gamma_i,\Phi_{i-1}) p_g(\gamma_i|\Phi_{i-1}) \times f_i(\Phi_{i-1}) \tag{12}$$

where we absorb any occurrence of $\mathcal{B}_{i,sp_{\prec}(i)}$ within the sampling distribution and factorize the remaining dependence on previous parameters $\Phi_{i-1} \equiv \{\gamma_1,\gamma_2,\mathcal{B}_{2,sp_{\prec}(2)},\ldots,\gamma_{i-1},\mathcal{B}_{i-1,sp_{\prec}(i-1)}\}$ into a separate function.[6] We derive the functions $p_b(\cdot), p_g(\cdot)$ and $f_i(\cdot)$ in Appendix A. The result is as follows.

The density $p_b(\mathcal{B}_{i,sp_{\prec}(i)}|\gamma_i,\Phi_{i-1})$ is the density of a Gaussian $N(\mathbf{K}_i\mathbf{m}_i,\gamma_i\mathbf{K}_i)$ such that

$$\begin{aligned}
\mathbf{m}_i &= (\mathbf{U}_{ss}-\mathbf{A}_i\mathbf{U}_{ns})\mathbf{M}_{sp_{\prec}(i)} + (\mathbf{U}_{sn}-\mathbf{A}_i\mathbf{U}_{nn})\mathbf{M}_{nsp_{\prec}(i)},\\
\mathbf{K}_i^{-1} &= \mathbf{U}_{ss}-\mathbf{A}_i\mathbf{U}_{ns}-\mathbf{U}_{sn}\mathbf{A}_i^{\mathsf{T}}+\mathbf{A}_i\mathbf{U}_{nn}\mathbf{A}_i^{\mathsf{T}},\\
\mathbf{A}_i &= \Sigma_{sp_{\prec}(i),nsp_{\prec}(i)}\Sigma^{-1}_{nsp_{\prec}(i),nsp_{\prec}(i)}
\end{aligned} \tag{13}$$

where

$$\begin{bmatrix} \mathbf{U}_{ss} & \mathbf{U}_{sn} \\ \mathbf{U}_{ns} & \mathbf{U}_{nn} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{U}_{sp_{\prec}(i),sp_{\prec}(i)} & \mathbf{U}_{sp_{\prec}(i),nsp_{\prec}(i)} \\ \mathbf{U}_{nsp_{\prec}(i),sp_{\prec}(i)} & \mathbf{U}_{nsp_{\prec}(i),nsp_{\prec}(i)} \end{bmatrix}. \tag{14}$$

The density $p_g(\gamma_i|\Phi_{i-1})$ is the density of an inverse gamma $IG(g_1,g_2)$ such that

$$\begin{aligned}
g_1 &= \frac{\delta+i-1+\#nsp_{\prec}(i)}{2},\\
g_2 &= \frac{u_{ii.\{i-1\},\{i-1\}}+\mathcal{U}_i}{2},\\
\mathcal{U}_i &= \mathbf{M}_i^{\mathsf{T}}\mathbf{U}_{\{i-1\},\{i-1\}}\mathbf{M}_i - \mathbf{m}_i^{\mathsf{T}}\mathbf{K}_i\mathbf{m}_i.
\end{aligned}$$

where $u_{ii.\{i-1\},\{i-1\}}$ was originally defined in Section 3.2.1.

Finally,

$$\begin{aligned}
f_i(\Phi_{i-1}) &\equiv (2\pi)^{-\frac{(i-1)-\#sp_{\prec}(i)}{2}} |\mathbf{K}_i|^{1/2} |\mathbf{U}_{\{i-1\},\{i-1\}}|^{1/2}\\
&\times \frac{(u_{ii.\{i-1\},\{i-1\}}/2)^{(\delta+i-1)/2}}{\Gamma((\delta+i-1)/2)} \frac{\Gamma((\delta+i-1+\#nsp_{\prec}(i))/2)}{((u_{ii.\{i-1\},\{i-1\}}+\mathcal{U}_i)/2)^{(\delta+i-1+\#nsp_{\prec}(i))/2}}.
\end{aligned}$$

6. A simpler decomposition was employed by Silva and Ghahramani (2006) (notice however that paper used an incorrect expression for the Jacobian). The following derivation, however, can be adapted with almost no modification to define a Gibbs sampling algorithm, as we show in the sequel.

Density function $p_b(\mathcal{B}_{i,sp_\prec(i)}|\cdot,\cdot)$ and determinant $|\mathbf{K}_i|^{1/2}$ are defined to be 1 if $sp_\prec(i)=\emptyset$. $\mathcal{U}_i$ is defined to be zero if $nsp_\prec(i)=\emptyset$, and $\mathcal{U}_i=\mathbf{M}_i^\mathsf{T}\mathbf{U}_{\{i-1\},\{i-1\}}\mathbf{M}_i$ if $sp_\prec(i)=\emptyset$.

The original normalizing constant integral is the expected value of a function of $\Phi_\mathcal{G}$ over a factorized inverse gamma-normal distribution. The density function of this distribution is given below:

$$p_{I(\delta,\mathbf{U})}(\Phi_\mathcal{G})=\left(\prod_{i=1}^m p_g(\gamma_i|\Phi_{i-1})\right)\left(\prod_{i=2}^m p_b(\mathcal{B}_{i,sp_\prec(i)}|\gamma_i,\Phi_{i-1})\right).$$

We summarize the main result of this section through the following theorem:

Theorem 3 *Let $\langle f(\mathbf{X})\rangle_{p(\mathbf{X})}$ be the expected value of $f(\mathbf{X})$ where $\mathbf{X}$ is a random vector with density $p(\mathbf{X})$. The normalizing constant of a $\mathcal{G}$-Inverse Wishart with parameters $(\delta,\mathbf{U})$ is given by*

$$I_\mathcal{G}(\delta,\mathbf{U})=I_{IW}(\delta,\mathbf{U})\times\left\langle\prod_{i=1}^m\frac{f_i(\Phi_{i-1})}{|\Sigma_{nsp_\prec(i),nsp_\prec(i)}|}\right\rangle_{p_{I(\delta,\mathbf{U})}(\Phi_\mathcal{G})}.$$

This can be further simplified to

$$I_\mathcal{G}(\delta,\mathbf{U})=\left\langle\prod_{i=1}^m\frac{f_i'(\Phi_{i-1})}{|\Sigma_{nsp_\prec(i),nsp_\prec(i)}|}\right\rangle_{p_{I(\delta,\mathbf{U})}(\Phi_\mathcal{G})} \tag{15}$$

where

$$f_i'(\Phi_{i-1})\equiv(2\pi)^{\frac{\#sp_\prec(i)}{2}}|\mathbf{K}_i(\Phi_{i-1})|^{1/2}\frac{\Gamma((\delta+i-1+\#nsp_\prec(i))/2)}{((u_{ii.\{i-1\},\{i-1\}}+\mathcal{U}_i)/2)^{(\delta+i-1+\#nsp_\prec(i))/2}}$$

which, as expected, reduces $I_\mathcal{G}(\delta,\mathbf{U})$ to $I_{IW}(\delta,\mathbf{U})$ when the graph is complete.

A Monte Carlo estimate of $I_\mathcal{G}(\delta,\mathbf{U})$ is then given from (15) by obtaining samples $\{\Phi_\mathcal{G}^{(1)},\Phi_\mathcal{G}^{(2)},\ldots,\Phi_\mathcal{G}^{(M)}\}$ according to $p_{I(\delta,\mathbf{U})}(\cdot)$ and computing:

$$I_\mathcal{G}(\delta,\mathbf{U})\approx\frac{1}{M}\sum_{s=1}^M\prod_{i=1}^m\frac{f_i'(\Phi_{i-1}^{(s)})}{|\Sigma_{nsp_\prec(i),nsp_\prec(i)}(\Phi_{i-1}^{(s)})|}$$

where here we emphasize that $\Sigma_{nsp_\prec(i),nsp_\prec(i)}$ is a function of $\Phi_\mathcal{G}$ as given by (6).

3.3 General Monte Carlo Computation

If $\mathbf{Y}$ follows a Gaussian $N(0,\Sigma)$ where Σ is given a $\mathcal{G}$-$IW(\delta,\mathbf{U})$ prior, then from a sample $\mathcal{D}=\{\mathbf{Y}^{(1)},\ldots,\mathbf{Y}^{(n)}\}$ with sufficient statistic $\mathbf{S}=\sum_{d=1}^n(\mathbf{Y}^{(d)})(\mathbf{Y}^{(d)})^\mathsf{T}$, the posterior distribution for Σ given $\mathbf{S}$ will be a $\mathcal{G}$-$IW(\delta+n,\mathbf{U}+\mathbf{S})$. In order to obtain samples from the posterior or to compute its functionals, one can adapt the algorithm for computing normalizing constants. We describe an importance sampler for computing functionals, followed by a Gibbs sampling algorithm that also provides samples from the posterior.

Algorithm SAMPLEGIW-1
Input: a $m \times m$ matrix $\mathbf{U}$, scalar δ, bi-directed graph $\mathcal{G}$, an ordering $\prec$

1. Let Σ be a $m \times m$ matrix
2. Define functions $sp_{\prec}(\cdot)$, $nsp_{\prec}(\cdot)$ according to $\mathcal{G}$ and ordering $\prec$
3. Sample σ_{11} from $IG(\delta/2, u_{11}/2)$
4. For $i = 2, 3, \ldots, m$
5. Sample $\gamma_i \sim IG((\delta + i - 1 + \#nsp_{\prec}(i))/2, (u_{ii.\{i-1\},\{i-1\}} + \mathcal{U}_i)/2)$
6. Sample $\mathcal{B}_{i,sp_{\prec}(i)} \sim N(\mathbf{K}_i \mathbf{m}_i, \gamma_i \mathbf{K}_i)$
7. Set $\mathcal{B}_{i,nsp_{\prec}(i)} = -\mathcal{B}_{i,sp_{\prec}(i)} \Sigma_{sp_{\prec}(i),nsp_{\prec}(i)} \Sigma^{-1}_{nsp_{\prec}(i),nsp_{\prec}(i)}$
8. Set $\Sigma^{\mathsf{T}}_{\{i-1\},i} = \Sigma_{i,\{i-1\}} = \mathcal{B}_i \Sigma_{\{i-1\},\{i-1\}}$
9. Set $\sigma_{ii} = \gamma_i + \mathcal{B}_i \Sigma_{i,\{i-1\}}$
10. Set $w = \prod_{i=1}^{m} f'_i(\Phi_{i-1}) / |\Sigma_{nsp_{\prec}(i),nsp_{\prec}(i)}|$
11. Return (w, Σ).

Figure 4: A procedure for generating an importance sample Σ and importance weight w for computing functionals of a $\mathcal{G}$-Inverse Wishart distribution. Variables $\{\mathbf{M}_i, \mathbf{m}_i, \mathbf{K}_i, \mathcal{U}_i\}$ and function $f'_i(\Phi_{i-1})$ are defined in Section 3.2.2.

3.3.1 THE IMPORTANCE SAMPLER

One way of computing functionals of the $\mathcal{G}$-*IW* distribution, that is, functions of the type

$$g(\delta, \mathbf{U}; \mathcal{G}) \equiv \int_{M^+(\mathcal{G})} g(\Sigma) p(\Sigma \mid \delta, \mathbf{U}, \mathcal{G}) \, d\Sigma$$

is through the numerical average

$$g(\delta, \mathbf{U}; \mathcal{G}) \approx \frac{\sum_{s=1}^{M} w_s g(\Sigma^{(s)})}{\sum_{s=1}^{M} w_s},$$

where weights $\{w_1, w_2, \ldots, w_M\}$ and samples $\{\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(M)}\}$ are generated by an importance sampler. The procedure for computing normalizing constants can be readily adapted for this task using $p_{I(\delta,\mathbf{U})}(\cdot)$ as the importance distribution and the corresponding weights from the remainder factors. The sampling algorithm is shown in Figure 4.

3.3.2 THE GIBBS SAMPLER

While the importance sampler can be useful to compute functionals of the distribution, we will need a Markov chain Monte Carlo procedure to sample from the posterior. In the Gibbs sampling

Algorithm SAMPLEGIW-2
Input: a $m \times m$ matrix $\mathbf{U}$, scalar δ, bi-directed graph $\mathcal{G}$, a $m \times m$ matrix Σ^{start}

1. Let Σ be a copy of Σ^{start}
2. Define functions $sp(\cdot)$, $nsp(\cdot)$ according to $\mathcal{G}$
3. For $i = 1, 2, 3, \ldots, m$
4. Sample $\gamma_i \sim IG((\delta + (m-1) + \#nsp(i))/2, (u_{ii.\{\setminus i\},\{\setminus i\}} + \mathcal{U}_{\setminus i})/2)$
5. Sample $\mathcal{B}_{i,sp(i)}$ from a $N(\mathbf{K}_{\setminus i}\mathbf{m}_{\setminus i}, \gamma_i \mathbf{K}_{\setminus i})$
6. Set $\mathcal{B}_{i,nsp(i)} = -\mathcal{B}_{i,sp(i)}\Sigma_{sp(i),nsp(i)}\Sigma^{-1}_{nsp(i),nsp(i)}$
7. Set $\Sigma^{\mathsf{T}}_{\{\setminus i\},i} = \Sigma_{i,\{\setminus i\}} = \mathcal{B}_i\Sigma_{\{\setminus i\},\{\setminus i\}}$
8. Set $\sigma_{ii} = \gamma_i + \mathcal{B}_i\Sigma_{i,\{\setminus i\}}$
9. Return Σ.

Figure 5: A procedure for generating a sampled Σ within a Gibbs sampling procedure.

procedure, we sample the whole i-th row of Σ, for each $1 \leq i \leq m$, by conditioning on the remaining independent entries of the covariance matrix as obtained on the previous Markov chain iteration.

The conditional densities required by the Gibbs sampler can be derived from (12), which for a particular ordering $\prec$ implies

$$p(\Sigma; \delta, \mathbf{U}, \mathcal{G}) \propto p_g(\gamma_1) \prod_{i=2}^{m} p_b(\mathcal{B}_{i,sp_\prec(i)}|\gamma_i, \Phi_{i-1}) p_g(\gamma_i|\Phi_{i-1}) f_i(\Phi_{i-1}).$$

By an abuse of notation, we used Σ in the left-hand side and the Bartlett parameters in the righ-hand side.

The conditional density of $\{\mathcal{B}_{m,sp_\prec(m)}, \gamma_m\}$ given all other parameters is therefore

$$p(\mathcal{B}_{m,sp_\prec(m)}, \gamma_m | \Phi_\mathcal{G} \setminus \{\mathcal{B}_{m,sp_\prec(m)}, \gamma_m\}) = p_b(\mathcal{B}_{m,sp_\prec(m)}|\gamma_m, \Phi_{m-1}) p_g(\gamma_m|\Phi_{m-1})$$

from which we can reconstruct a new sample of the m-th row/column of Σ after sampling $\{\mathcal{B}_{m,sp_\prec(m)}, \gamma_m\}$. Sampling other rows can be done by redefining a new order where the corresponding target variable is the last one.

More precisely: let $\{\setminus i\}$ denote the set $\{1, 2, \ldots, i-1, i+1, \ldots, m\}$. The Gibbs algorithm is analogous to the previous algorithms. Instead of $sp_\prec(i)$ and $nsp_\prec(i)$, we refer to the original $sp(i)$ and $nsp(i)$. Matrices $\Sigma_{\{\setminus i\},\{\setminus i\}}$ and $\mathbf{U}_{\{\setminus i\},\{\setminus i\}}$ are defined by deleting the respective i-th row and i-th columns. Row vector $\Sigma_{i,\{\setminus i\}}$ and scalar $u_{ii.\{\setminus i\}}$ are defined accordingly, as well as any other vector and matrix originally required in the marginal likelihood/importance sampling procedure. The algorithm is described in Figure 5. The procedure can be interpreted as calling a modification of the importance sampler with a dynamic ordering $\prec_i$ which, at every step, moves Y_i to the end of the global ordering $\prec$.

3.4 Remarks

The importance sampler suffers from the usual shortcomings in high-dimensional problems, where a few very large weights dominate the procedure (MacKay, 1998). This can result in unstable estimates of functionals of the posterior and the normalizing constant.

The stability of the importance sampler is not a simple function of the number of variables in the domain. For large but sparse graphs, the number of parameters might be small. For large but fairly dense graphs, the importance distribution might be a good match to the actual distribution since there are few constraints. In Section 7, we performe some experiments to evaluate the sampler.

When used to compute functionals, the Gibbs sampler is more computationally demanding considering the cost per step, but we expect it to be more robust in high-dimensional problems. In problems that require repeated calculations of functionals (such as the variational optimization procedure of Section 4.3), it might be interesting to run a few preliminary comparisons between the estimates of the two samplers, and choose the (cheaper) importance sampler if the estimates are reasonably close.

Naïvely, the Gibbs sampler costs $O(m^4)$ per iteration, since for each step we have to invert the matrix $\Sigma_{nsp\{\backslash i\},nsp\{\backslash i\}}$, which is of size $O(m)$ for sparse graphs. However, this inversion can cost much less than $O(m^3)$ if sparse matrix inversion methods are used. Still, the importance sampler can be even more optimized by using the methods of Section 6.

4. Gaussian Directed Mixed Graph Models

As discussed in Section 2, Gaussian directed mixed graph models are parameterized by the set with parameters $\Theta = \{\mathbf{V}, \mathbf{B}, \alpha\}$. Our prior takes the form $p(\Theta) = p(\mathbf{B})p(\alpha)p(\mathbf{V})$. We assign priors for the parameters of directed edges (non-zero entries of matrix **B**) in a standard way: each parameter b_{ij} is given a Gaussian $N(c_{ij}^B, s_{ij}^B)$ prior, where all parameters are marginally independent in the prior, that is, $p(\mathbf{B}) = \prod_{ij} p(b_{ij})$. The prior for intercept parameters α is analogous, with α_i being a Gaussian $N(c_i^\alpha, s_i^\alpha)$.

Recall from Equation (1) that the implied covariance of the model is given by the matrix $\Sigma(\Theta) = (\mathbf{I}-\mathbf{B})^{-1}\mathbf{V}(\mathbf{I}-\mathbf{B})^{-\mathsf{T}}$. Similarly, we have the implied mean vector $\mu(\Theta) \equiv (\mathbf{I}-\mathbf{B})^{-1}\alpha$. The likelihood function for data set $\mathcal{D} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}\}$ is defined as

$$\begin{aligned}\mathcal{L}(\Theta;\mathcal{D}) &= |\Sigma(\Theta)|^{-n/2}\prod_{d=1}^{n}\exp\left(-\tfrac{1}{2}(\mathbf{Y}^{(d)}-\mu(\Theta))^{\mathsf{T}}\Sigma(\Theta)^{-1}(\mathbf{Y}^{(d)}-\mu(\Theta))\right)\\ &= \left\{|(\mathbf{I}-\mathbf{B})^{-1}||\mathbf{V}||(\mathbf{I}-\mathbf{B})^{-\mathsf{T}}|\right\}^{-n/2}\left\{\exp\left(-\tfrac{1}{2}tr(\mathbf{V}^{-1}(\mathbf{I}-\mathbf{B})\mathbf{S}(\mathbf{I}-\mathbf{B})^{\mathsf{T}})\right)\right\},\end{aligned}$$

where now $\mathbf{S} \equiv \sum_{d=1}^{n}(\mathbf{Y}^{(d)}-\mu(\Theta))(\mathbf{Y}^{(d)}-\mu(\Theta))^{\mathsf{T}}$.

Given a prior $\mathcal{G}$-$IW(\delta, \mathbf{U})$ for **V**, it immediately follows that the posterior distribution of **V** given the data and other parameters is

$$\mathbf{V} \mid \{\mathbf{B}, \alpha, \mathcal{D}\} \sim \mathcal{G}\text{-}IW(\delta+n, \mathbf{U}+(\mathbf{I}-\mathbf{B})\mathbf{S}(\mathbf{I}-\mathbf{B})^{\mathsf{T}}).$$

Therefore it can be sampled using the results from the previous section. Notice this holds even if the directed mixed graph $\mathcal{G}$ is cyclic.

Sampling α_i given $\{\mathcal{D}, \Theta\backslash\{\alpha_i\}\}$ can also be done easily for both cyclic and acyclic models: the posterior is given by a normal $N(c_i^{\alpha'}/s_i^{\alpha'}, 1/s_i^{\alpha'})$ where

$$
\begin{aligned}
s_i^{\alpha'} &\equiv \frac{1}{s_i^{\alpha}} + n(\mathbf{V}^{-1})_{ii}, \\
c_i^{\alpha'} &\equiv \frac{c_i^{\alpha}}{s_i^{\alpha}} - n \sum_{t=1,t\neq i}^{m} (\mathbf{V}^{-1})_{it}\alpha_t + \sum_{d=1}^{n}\sum_{t=1}^{m} (\mathbf{V}^{-1})_{it}\left(Y_t^{(d)} - \sum_{p_t} b_{tp_t} Y_{p_t}^{(d)}\right),
\end{aligned}
$$

with p_t being an index running over the parents of Y_t in $\mathcal{G}$.

However, sampling the non-zero entries of $\mathbf{B}$ results in two different cases depending whether $\mathcal{G}$ is cyclic or not. We deal with them separately.

4.1 Sampling from the Posterior: Acyclic Case

The acyclic case is simplified by the fact that $\mathbf{I}-\mathbf{B}$ can be rearranged in a way it becomes lower triangular, with each diagonal element being 1. This implies the identity $|(\mathbf{I}-\mathbf{B})^{-1}||\mathbf{V}||(\mathbf{I}-\mathbf{B})^{-\mathsf{T}}| = |\mathbf{V}|$, with the resulting log-likelihood being a quadratic function of the non-zero elements of $\mathbf{B}$. Since the prior for coefficient b_{ij} is Gaussian, its posterior given the data and all other parameters will be the Gaussian $N(c_{ij}^{b'}/s_{ij}^{b'}, 1/s_{ij}^{b'})$ where

$$
\begin{aligned}
s_{ij}^{b'} &\equiv \frac{1}{s_{ij}^{b}} + (\mathbf{V}^{-1})_{ii}\sum_{d=1}^{n}(Y_j^{(d)})^2, \\
c_{ij}^{b'} &\equiv \frac{c_{ij}^{b}}{s_{ij}^{b}} + \sum_{d=1}^{n} Y_j^{(d)} \sum_{t=1}^{m} (\mathbf{V}^{-1})_{it}\left(Y_t^{(d)} - \sum_{p_t,(t,p_t)\neq(i,j)} b_{tp_t} Y_{p_t}^{(d)} - \alpha_t\right).
\end{aligned}
\tag{16}
$$

As before, p_t runs over the indices of the parents of Y_t in $\mathcal{G}$. Notice that in the innermost summation we exclude $b_{ij}Y_j^{(d)}$. We can then sample b_{ij} accordingly.

It is important to notice that, in practice, better mixing behavior can be obtained by sampling the coefficients (and intercepts) jointly. The joint distribution is Gaussian and can be obtained in a way similar to the above derivation. The derivation of the componentwise conditionals is nevertheless useful in the algorithm for cyclic networks.

4.2 Sampling from the Posterior: Cyclic Case

Cyclic directed graph models have an interpretation in terms of causal systems in equilibrium. The simultaneous presence of directed paths $Y_i \to \cdots \to Y_j$ and $Y_j \to \cdots \to Y_i$ can be used to parameterize instantaneous causal effects in a feedback loop (Spirtes, 1995). This model appears also in the structural equation modeling literature (Bollen, 1989). In terms of cyclic graphs as families of conditional independence constraints, methods for reading off constraints in linear systems also exist (Spirtes et al., 2000).

The computational difficulty in the cyclic case is that the determinant $|\mathbf{I}-\mathbf{B}|$ is no longer a constant, but a multilinear function of coefficients $\{b_{ij}\}$. Because b_{ij} will appear outside the exponential term, its posterior will no longer be Gaussian.

From the definition of the implied covariance matrix $\Sigma(\Theta)$, it follows that $|\Sigma(\Theta)|^{-n/2} = (|\mathbf{I}-\mathbf{B}||\mathbf{V}|^{-1}|\mathbf{I}-\mathbf{B}|)^{n/2}$. As a function of coefficient b_{ij},

$$
|\mathbf{I}-\mathbf{B}| = (-1)^{i+j+1}C_{ij}b_{ij} + \sum_{k=1,k\neq j}^{k=m} (-1)^{i+k+1}C_{ik}b_{ik},
$$

where C_{ij} is the determinant of respective co-factor of $\mathbf{I}-\mathbf{B}$, $b_{ik} \equiv 0$ if there is no edge $Y_i \leftarrow Y_k$, and $b_{ii} \equiv -1$. The resulting density function of b_{ij} given $\mathcal{D}$ and $\Theta \backslash \{b_{ij}\}$ is

$$p(b_{ij}|\Theta \backslash \{b_{ij}\}, \mathcal{D}) \propto |b_{ij} - \kappa_{ij}|^n \exp\left\{ -\frac{(b_{ij} - c_{ij}^{b'}/s_{ij}^{b'})^2}{2s_{ij}^{b'}} \right\},$$

where

$$\kappa_{ij} \equiv C_{ij}^{-1} \sum_{k=1, k\neq j}^{k=m} (-1)^{k-j+1} C_{ik} b_{ik}$$

and $\{c_{ij}^{b'}, s_{ij}^{b'}\}$ are defined as in Equation (16). Standard algorithms such as Metropolis-Hastings can be applied to sample from this posterior within a Gibbs procedure.

4.3 Marginal Likelihood: A Variational Monte Carlo Approach

While model selection of bi-directed graphs can be performed using a simple Monte Carlo procedure as seen in the previous Section, the same is not true in the full Gaussian DMG case. Approaches such as nested sampling (Skilling, 2006) can in principle be adapted to deal with the full case. For problems where there are many possible candidates to be evaluated, such a computationally demanding sampling procedure might be undesirable (at least for an initial ranking of graphical structures). As an alternative, we describe an approximation procedure for the marginal likelihood $p(\mathcal{D}|\mathcal{G})$ by combining variational bounds (Jordan et al., 1998) with the $\mathcal{G}$-Inverse Wishart samplers, and therefore avoiding a Markov chain over the joint model of coefficients and error covariances. This is described for acyclic DMGs only.

We adopt the following approximation in our variational approach, accounting also for possible latent variables $\mathbf{X}$:

$$p(\mathbf{V}, \mathbf{B}, \alpha, \mathbf{X}|\mathcal{D}) \approx q(\mathbf{V}) q(\mathbf{B}, \alpha) \prod_{d=1}^{n} q(\mathbf{X}^{(d)}) \equiv q(\mathbf{V}) q(\mathbf{B}, \alpha) q(\mathbf{X})$$

with $q(\mathbf{B}, \alpha)$ being a multivariate Gaussian density of the non-zero elements of $\mathbf{B}$ and α. Function $q(\mathbf{X}^{(d)})$ is also a Gaussian density, and function $q(\mathbf{V})$ is a $\mathcal{G}$-Inverse Wishart density.

From Jensen's inequality, we obtain the following lower-bound (Beal, 2003, p. 47):

$$\begin{aligned} \ln p(\mathcal{D}|\mathcal{G}) &= \ln \int p(\mathbf{Y}, \mathbf{X}|\mathbf{V}, \mathbf{B}, \alpha) p(\mathbf{V}, \mathbf{B}, \alpha)\, d\mathbf{X}\, d\mathbf{B}\, d\mathbf{V}\, d\alpha \\ &\geq \langle \ln p(\mathbf{Y}, \mathbf{X}|\mathbf{V}, \mathbf{B}, \alpha) \rangle_{q(\mathbf{V})q(\mathbf{B},\alpha)q(\mathbf{X})} \\ &\quad + \langle \ln p(\mathbf{V})/q(\mathbf{V}) \rangle_{q(\mathbf{V})} \\ &\quad + \langle \ln p(\mathbf{B}, \alpha)/q(\mathbf{B}, \alpha) \rangle_{q(\mathbf{B},\alpha)} - \langle \ln q(\mathbf{X}) \rangle_{q(\mathbf{X})} \end{aligned} \tag{17}$$

where this lower bound can be optimized with respect to functions $q(\mathbf{V})$, $q(\mathbf{B})$, $q(\mathbf{X})$. This can be done by iterative coordinate ascent, maximizing the bound with respect to a single $q(\cdot)$ function at a time.

The update of $q(\mathbf{V})$ is given by

$$q^{new}(\mathbf{V}) = p_{\mathcal{G}\text{-}IW}(\delta + d, \mathbf{U} + \left\langle (\mathbf{I}-\mathbf{B})\mathbf{S}(\mathbf{I}-\mathbf{B})^{\mathsf{T}} \right\rangle_{q(\mathbf{X})q(\mathbf{B},\alpha)})$$

where $p_{\mathcal{G}\text{-}IW}(\cdot)$ is the density function for a $\mathcal{G}$-Inverse Wishart, and $\mathbf{S}$ is the empirical second moment matrix summed over the completed data set $(\mathbf{X},\mathbf{Y})$ (hence the expectation over $q(\mathbf{X})$) centered at $\mu(\Theta)$.

The updates for $q(\mathbf{B},\alpha)$ and $q(\mathbf{X})$ are tedious but straightforward derivations, and described in Appendix B. The relevant fact about these updates is that they are functions of $\left\langle \mathbf{V}^{-1} \right\rangle_{q(\mathbf{V})}$. Fortunately, we pay a relatively small cost to obtain these inverses using the Monte Carlo sampler of Figure 4: from the Bartlett parameters, define a lower triangular $m \times m$ matrix $\mathcal{B}$ (by placing on the ith line the row vector $\mathcal{B}_i$, followed by zeroes) and a diagonal matrix Γ from the respective vector of γ_i's. The matrix $\mathbf{V}^{-1}$ can be computed from $(\mathbf{I}-\mathcal{B})^{\mathsf{T}}\Gamma^{-1}(\mathbf{I}-\mathcal{B})$, and the relevant expectation computed according to the importance sampling procedure. For problems of moderate dimensionality,[7] the importance sampler might not be recommended, but the Gibbs sampler can be used.

At the last iteration of the variational maximization, the (importance or posterior) samples from $q(\mathbf{V})$ can then be used to compute the required averages in (17), obtaining a bound on the marginal log-likelihood of the model. Notice that the expectation $\langle \ln p(\mathbf{V})/q(\mathbf{V}) \rangle_{q(\mathbf{V})}$ contains the entropy of $q(\mathbf{V})$, which will require the computation of $\mathcal{G}$-inverse Wishart normalizing constants.

For large problems, the cost of this approximation might still be prohibitive. An option is to partially parameterize $\mathbf{V}$ in terms of ancillary latents and another submatrix distributed as a $\mathcal{G}$-inverse Wishart, but details on how to best do this partition are left as future work (this approximation will be worse but less computationally expensive if ancillary latents are independent of the coefficient parameters in the variational density function $q(\cdot)$). Laplace approximations might be an alternative, which have been successfully applied to undirected non-decomposable models (Roverato, 2002).

We emphasize that the results present in this section are alternatives that did not exist before in previous approaches for learning mixed graph structures through variational methods (e.g., Silva and Scheines, 2006). It is true that the variational approximation for marginal likelihoods will tend to underfit the data, that is, generate models simpler than the true model in simulations. Despite the bias introduced by the method, this is less of a problem for large data sets (Beal and Ghahramani, 2006) and the method has been shown to be useful in model selection applications (Silva and Scheines, 2006), being consistently better than standard scores such as BIC when hidden variables are present (Beal and Ghahramani, 2006). An application in prediction using the variational posterior instead of MCMC samples is discussed by Silva and Ghahramani (2006). It is relevant to explore other approaches for marginal likelihood evaluation of DMG models using alternative methods such as annealed importance sampling (Neal, 2001) and nested sampling (Skilling, 2006), but it is unrealistic to expect that such methods can be used to evaluate a large number of candidate models. A pre-selection by approximations such as variational methods might be essential.

5. Discrete Models: The Probit Case

Constructing a discrete mixed graph parameterization is not as easy as in the Gaussian case. Advances in this area are described by Drton and Richardson (2008a), where a complete parameterization of binary bi-directed graph models is given. In our Bayesian context, inference with the mixed graph discrete models of Drton and Richardson would not to be any computationally easier than the case for Markov random fields, which has been labeled as *doubly-intractable* (Murray et al., 2006).

7. We observed a high ratio of the highest importance weight divided by the median weight in problems with dimensionality as low as 15 nodes. However, notice that in practice the error covariance matrix $\mathbf{V}$ has a block diagonal structure, and only the size of the largest block is relevant. This is explained in more detail in Section 6.

Instead, in this paper we will focus on a class of discrete models that has been widely used in practice: the probit model (Bartholomew and Knott, 1999). This model is essentially a projection of a Gaussian distribution into a discrete space. It also allows us to build on the machinery developed in the previous sections. We will describe the parameterization of the model for acyclic DMGs, and then proceed to describe algorithms for sampling from the posterior distribution.

5.1 Parameterizing Models of Observable Independencies

A probit model for the conditional probability of discrete variable Y_i given a set of variables $\{Y_{i1},...,Y_{ik}\}$ can be described by the two following relationships:

$$\begin{aligned} Y_i^\star &= \alpha_i + b_{i1}Y_{i1} + b_{i2}Y_{i2} + \cdots + b_{ik}Y_{ik} + \varepsilon_i \\ \mathcal{P}(Y_i = v_l^i \mid Y_i^\star) &= 1(\tau_{l-1}^i \leq Y_i^\star < \tau_l^i) \end{aligned} \tag{18}$$

where $\mathcal{P}(\cdot)$ is the probability mass function of a given random variable, as given by the context, and $1(\cdot)$ is the indicator function. Y_i assumes values in $\{v_1^i, v_1^i, \ldots, v_{\kappa(i)}^i\}$. Thresholds $\{\tau_0^i = -\infty < \tau_1^i < \tau_2^i < \cdots < \tau_{\kappa(i)}^i = \infty\}$ are used to define the mapping from continuous $Y_i^\star$ to discrete Y_i. This model has a sensible interpretation for ordinal and binary values as the discretization of some *underlying latent variable* (UV) $Y_i^\star$. Such a UV is a conditionally Gaussian random variable, which follows by assuming normality of the error term ε_i. This formulation, however, is not appropriate for general discrete variables, which are out of the scope of this paper. Albert and Chib (1993) describe alternative Bayesian treatments of discrete distributions not discussed here.

Given this binary/ordinal regression formulation, the natural step is how to define a graphical model accordingly. As a matter of fact, the common practice does not strictly follow the probit regression model. Consider the following example: for a given graph $\mathcal{G}$, a respective graphical representation of a probit model can be built by first replicating $\mathcal{G}$ as a graph $\mathcal{G}^\star$, where each vertex Y_i is relabeled as $Y_i^\star$. Those vertices represent continuous underlying latent variables (UVs). To each vertex $Y_i^\star$ in $\mathcal{G}^\star$, we then add a single child Y_i. We call this the *Type-I UV model*. Although there are arguments for this approach (see, for instance, the arguments by Webb and Forster (2006) concerning stability to ordinal encoding), this is a violation of the original modeling assumption as embodied by $\mathcal{G}$: if the given graph is a statement of conditional independence constraints, it is expected that such independencies will be present in the actual model. The Type-I formulation does not fulfill this basic premise: by construction there are no conditional independence constraints among the set of variables $\mathbf{Y}$ (the marginal independencies are preserved, though). This is illustrated by Figure 6(b), where the conditional independence of Y_1 and Y_3 given Y_2 disappears.

An alternative is illustrated in Figure 6(c). Starting from the original graph $\mathcal{G}$ (as in Figure 6(a)), the probit graph model $\mathcal{G}^\star$ shown in the Figure is built from $\mathcal{G}$ by the following algorithm:

1. add to empty graph $\mathcal{G}^\star$ the vertices $\mathbf{Y}$ of $\mathcal{G}$, and for each $Y_i \in \mathbf{Y}$, add a respective UV $Y_i^\star$ and the edge $Y_i^\star \to Y_i$;
2. for each $Y_i \to Y_j$ in $\mathcal{G}$, add edge $Y_i \to Y_j^\star$ to $\mathcal{G}^\star$;
3. for each $Y_i \leftrightarrow Y_j$ in $\mathcal{G}$, add edge $Y_i^\star \leftrightarrow Y_j^\star$ to $\mathcal{G}^\star$;

We call this the *Type-II UV model*, which has the following property (the proof is in Appendix C):

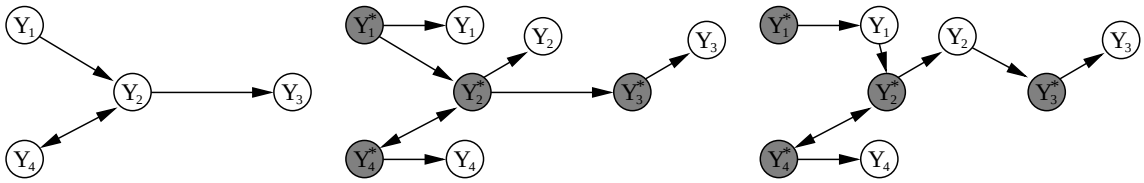


Figure 6: The model in (a) has at least two main representations as a probit network. In (b), the original structure is given to the underlying variables, with observed variables being children of their respective latents. In (c), the underlying variable inherits the parents of the original variable and the underlying latents of the spouses.

Theorem 4 *Suppose $\mathcal{G}$ is acyclic with vertex set $\mathbf{Y}$. Y_i and Y_j are m-separated given $\mathbf{Z} \subseteq \mathbf{Y} \backslash \{Y_i, Y_j\}$ in $\mathcal{G}$ if and only if Y_i and Y_j are m-separated given $\mathbf{Z}$ in $\mathcal{G}^\star$.*

The parameterization of the Type-II UV model follows from the definition of probit regression: the conditional distribution Y_i given its parents in $\{Y_{i1}, ..., Y_{ik}\}$ in $\mathcal{G}$ is given as in Equation (18), while the error terms $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$ follow the multivariate Gaussian $N(0, \mathbf{V})$. The entry corresponding to the covariance of ε_i and ε_j is assumed to be zero if there is no bi-directed edge $Y_i \leftrightarrow Y_j$ in $\mathcal{G}$.

In what follows, we discuss algorithms for Type-II models. The approach here described can be easily adapted to cover Type-I models. We say that Type-II models are models of *observable independencies*, since independencies hold even after marginalizing all UVs.

5.2 Algorithm

As before, we provide a Gibbs sampling scheme to sample parameters $\Theta = \{\alpha, \mathbf{B}, \mathbf{V}, \mathcal{T}\}$ from the posterior distribution given data set $\mathcal{D} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}\}$. The set $\mathcal{T} = \{\mathcal{T}_i\}$ is the set of threshold parameters, $\mathcal{T}_i = \{\tau_0^i = -\infty < \tau_1^i < \tau_2^i < \cdots < \tau_{\kappa(i)}^i = \infty\}$ for each random variable Y_i with $\kappa(i)$ different values. We will not discuss priors and algorithms for sampling $\mathcal{T}$ given the other parameters: this can be done by standard approaches (e.g., Albert and Chib, 1993).[8]

For the purposes of the Gibbs procedure, we augment the data set with the underlying variables $\mathcal{D}^\star = \{\mathbf{Y}^{\star(1)}, \mathbf{Y}^{\star(2)}, \ldots, \mathbf{Y}^{\star(n)}\}$ at each sampling step.

From the set of structural equations

$$\mathbf{Y}^{\star(d)} = \alpha + \mathbf{B}\mathbf{Y}^{(d)} + \varepsilon$$

it follows that the conditional distribution of $\mathbf{Y}^{\star(d)}$ given the $\mathcal{D} \cup \Theta$ is a truncated Gaussian with mean $\alpha + \mathbf{B}\mathbf{Y}^{(d)}$ and covariance matrix $\mathbf{V}$. The truncation levels are given by the thresholds and observed data $\mathbf{Y}^{(d)}$: for each $Y_i^{(d)} = v_l^i$, the range for $Y_i^{\star(d)}$ becomes $[\tau_{l-1}^i, \tau_l^i)$. Sampling from a truncated Gaussian is a standard procedure. We used the algorithm of Kotecha and Djuric (1999) in our implementation.

To sample $\mathbf{V}$ from its conditional, we will rely on the following result.

8. In Section 7, we perform experiments with binary data only. In this case, the thresholds are set to fixed values: $\{\tau_0^i = -\infty, \tau_1^i = 0, \tau_2^i = \infty\}$ for all $0 \leq i \leq m$.

Proposition 5 *Let $\mathcal{G}$ be an acyclic DMG, and $(\alpha, \mathbf{B}, \mathbf{V}, \mathcal{T})$ be the respective set of parameters that defines the probit model. For a fixed $(\alpha, \mathbf{B}, \mathcal{T})$, there is a bijective function $f_{\mathbf{B}\alpha\mathcal{T}}(\cdot)$ mapping $\mathbf{Y}^\star$ to ε. This is not true in general if $\mathcal{G}$ is cyclic.*

Proof: If the graph is acyclic, this follows directly by recursively solving the model equations, starting from those corresponding to $Y_j^\star$ vertices with no parents. This results in $\varepsilon = \mathbf{Y}^\star - \alpha - \mathbf{B}\mathbf{Y}$, as expected.

For cyclic graphs, the following model provides a counter-example. Let the graph be $Y_1^\star \to Y_1 \to Y_2^\star \to Y_2 \to Y_1^\star$. Let the model be $Y_1^\star = Y_2 + \varepsilon_1, Y_2^\star = Y_1 + \varepsilon_2$, that is, $b_{12} = b_{21} = 1$ and $\alpha = 0$. Let the variables be binary, with a threshold at zero ($Y_i = 1$ if and only if $Y_i^\star \geq 0$). Then the two instantiations $(Y_1^\star = -0.8, Y_2^\star = 0)$, $(Y_1^\star = 0.2, Y_2^\star = 1)$ imply the same pair $(\varepsilon_1 = -0.8, \varepsilon_2 = 0)$. □

The negative result for discrete models with cycles is the reason why such models are out of the scope of the paper.

Let $\mathcal{D}_\varepsilon^\star = \{\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}\}$, where $\varepsilon^{(d)} = f_{\mathbf{B}\alpha\mathcal{T}}(\mathbf{y}^{(d)\star})$. Due to this bijection (and the determinism mapping $\mathbf{Y}^\star$ to $\mathbf{Y}$), the density $p(\mathbf{V} \mid \Theta \backslash \mathbf{V}, \mathcal{D}, \mathcal{D}^\star) = p(\mathbf{V} \mid \Theta \backslash \mathbf{V}, \mathcal{D}^\star) = p(\mathbf{V} \mid \Theta \backslash \mathbf{V}, \mathbf{y}^{(1)\star}, \ldots, \mathbf{y}^{(d)\star})$ is equivalent to

$$\begin{aligned} p(\mathbf{V} \mid \Theta \backslash \mathbf{V}, \mathcal{D}^\star) &= p(\mathbf{V} \mid \alpha, \mathbf{B}, \mathcal{T}, \mathcal{D}^\star, \mathcal{D}_\varepsilon^\star) \\ &= p(\mathbf{V} \mid \alpha, \mathbf{B}, \mathcal{T}, \mathcal{D}_\varepsilon^\star) \\ &\propto p(\mathbf{V} \mid \alpha, \mathbf{B}, \mathcal{T}) p(\mathcal{D}_\varepsilon^\star \mid \alpha, \mathbf{B}, \mathcal{T}, \mathbf{V}) \\ &\propto p(\mathbf{V}) \textstyle\prod_{d=1}^n p(\varepsilon^{(d)} \mid \mathbf{V}). \end{aligned}$$

For the given data set $\mathcal{D} \cup \mathcal{D}^\star$, define $\mathbf{S}^\star$ as the sum of $(\mathbf{Y}^{\star(d)} - \alpha - \mathbf{B}\mathbf{Y}^{(d)})(\mathbf{Y}^{\star(d)} - \alpha - \mathbf{B}\mathbf{Y}^{(d)})^\mathsf{T}$ over all $d \in \{1, 2, \ldots, n\}$. Since $p(\varepsilon \mid \mathbf{V})$ is normal with zero mean and covariance matrix $\mathbf{V}$, the posterior for $\mathbf{V}$ given all other parameters and variables is

$$\mathbf{V} \mid \{\Theta \backslash \mathbf{V}, \mathcal{D}, \mathcal{D}^\star\} \sim \mathcal{G}\text{-}IW(\delta + n, \mathbf{U} + \mathbf{S}^\star).$$

Sampling $\mathbf{B}$ and α is analogous to the Gaussian case, except that we have to consider that the left-hand side of the structural equations now refer to $\mathbf{Y}^\star$. We give the explicit conditional for α_i, with the conditional for b_{ij} being similarly adapted from Section 4. The posterior for α_i is given by a normal $N((s_i')^{-1} m_i', s_i')$ where

$$\begin{aligned} s_i^{\alpha'} &= \frac{1}{s_i^\alpha} + n(\mathbf{V}^{-1})_{ii}, \\ c_i^{\alpha'} &= \frac{c_i^\alpha}{s_i^\alpha} - n \sum_{t=1, t\neq i}^{m} (\mathbf{V}^{-1})_{it} \alpha_t + \sum_{d=1}^{n} \sum_{t=1}^{m} (\mathbf{V}^{-1})_{it} \left(Y_t^{\star(d)} - \sum_{p_t} b_{t p_t} Y_{p_t}^{(d)} \right). \end{aligned}$$

5.3 A Note on Identifiability

The scale of the underlying latent variables in the probit model is arbitrary. As such, it has been often suggested that such latents should have constant (e.g., unity) variance (Pitt et al., 2006). There are two usual arguments for fixing the variance: improving the interpretability of the model, and improving the mixing of the Markov chain. The interpretability argument is not particularly appealing within the Bayesian setting with proper priors, such as the one proposed in this paper: the posterior distribution of the parameters is well-defined by the prior uncertainty and the data.

The goal of improving the mixing of the chain might be important: if some parameters can assume arbitrary values and still allow for the same model over the observables, then fixing such parameters may help sampling by eliminating largely flat regions from the posterior (which will happen for large data sets and broad priors). In practice, however, scaling UVs might not be advantageous. In some cases it might increase the computational cost of each sampling step, while sampling from the non-scaled model might work just fine. Many MCMC algorithms work well on highly unidentifiable models such as multilayer perceptrons (Neal, 1996). In our experiments, we do not use any scaling.

5.4 Remarks

It is clear that the given approach can be generalized to other generalized linear models by changing the link function that maps underlying latent variables (UVs) to observables. For instance, a model containing discrete and continuous variables can be constructed by using the identity link function instead of probit for the continuous variables. Notice that the continuous variables will not necessarily be marginally Gaussian if some of its parents are discrete. Other link functions will have different parameters besides thresholds, such as in multivalued ("polychotomous") discrete distributions. A Bayesian account of Gaussian copula models is given by Pitt et al. (2006), to which a DMG-based family could in principle be defined. For continuous, marginally non-Gaussian, variables joined by a Gaussian copula, it is possible that all link functions are invertible. In this case, it is easier in principle to define cyclic models through Type-I UV models (e.g., Figure 6(b)) while preserving the observable independencies.

It is important to point out that Type-II probit models with Markov equivalent graphs will not, in general, be likelihood equivalent. A simple example is given by the two-node graphs $Y_1 \rightarrow Y_2$ and $Y_1 \leftrightarrow Y_2$: if Y_1 is binary, then the marginal for Y_2 in the first case is equivalent to having an underlying latent variable that follows a mixture of two Gaussians. While some of these issues can be solved by adopting a mixture of Gaussians marginal independence model to account for bi-directed edges (Silva and Ghahramani, 2009), details need to be worked out. When the goal of model selection is to find causal structures (Spirtes et al., 2000), the usual guarantees of search methods based on Markov equivalence classes do not hold. However, it remains to be seen whether the parametric constraints implied by the Type-II formulation will allow for other consistent approaches for causal discovery, as shown in the case of non-linearities with additive noise (Hoyer et al., 2008).

6. Scaling Up: Factorizations and Perfect Sequences

Each Monte Carlo sampling step for the given mixed graph models is theoretically tractable, but not necessarily practical when the dimensionality m of the data is high. By using clever factorizations of the graph and ordering of the variables, it is possible to sometimes scale to high-dimensional problems. In this section, we describe approaches to minimize the run-time of the marginal likelihood computation for bi-directed graphs, which is also important for computing variational bounds for DMG models. We start, however, with a discussion on factorizations of the posterior density for coefficient parameters $\mathbf{B}$. The context is the Gibbs sampler for acyclic models.

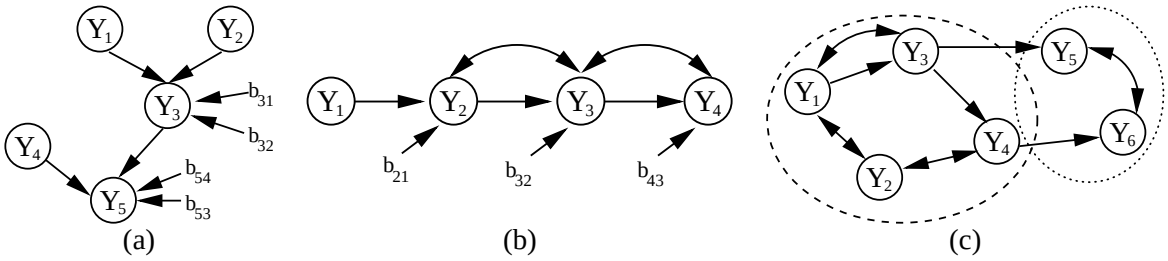


Figure 7: The coefficients b_{31} and b_{32}, represented as nodes in (a), become dependent after conditioning on $\mathbf{Y}$. However, they are still independent of b_{54}. This a general property of DAG models. In DMG models, a sequence of bi-directed edges will connect extra coefficients. In graph (b), coefficients b_{21}, b_{32} and b_{43} will all be dependent given $\mathbf{Y}$. Coefficients into nodes in different districts will still be independent. The graph in (c) has districts $\{Y_1, Y_2, Y_3, Y_4\}$ and $\{Y_5, Y_6\}$.

6.1 Factorizations

Our prior for coefficients $\{b_{ij}\}$ is fully factorized. In directed acyclic graphs, this is particularly advantageous: coefficients corresponding to edges into different nodes are independent in the posterior.[9] One can then jointly sample a whole set of $\{b_{ij}\}$ coefficients with same i index, with no concern for the other coefficients. Figure 7(a) illustrates this factorization. This means that, in Equation (16), the summation over t does not go over all variables, but only for $t = i$. This also follows from the fact that $(\mathbf{V})^{-1}_{it} = 0$ unless $i = t$, since $\mathbf{V}$ is diagonal.

In ADMGs, however, this is not true anymore. For any pair of vertices linked by a path of bi-directed edges, for example, $Y_i \leftrightarrow Y_{i+1} \leftrightarrow \cdots \leftrightarrow Y_t$, one will have in general that $(\mathbf{V})^{-1}_{it} \neq 0$. This can be shown by using the graphical properties of the model when conditioning on some arbitrary datapoint $\mathbf{Y}$:

Proposition 6 *Let $\mathcal{G}$ be an acyclic DMG with vertex set $\mathbf{Y}$, and $\mathcal{G}'$ the DMG obtained by augmenting $\mathcal{G}$ with a vertex for each parameter b_{ij} and a respective edge $b_{ij} \rightarrow Y_i$. Then if there is a bi-directed path $Y_i \leftrightarrow \cdots \leftrightarrow Y_t$ in $\mathcal{G}$, $\{b_{ij}, b_{tv}\}$ are not m-separated given $\mathbf{Y}$ in $\mathcal{G}'$.*

Proof: The joint model for $\{\mathbf{Y}, \mathbf{B}\}$ with independent priors on the non-zero entries of $\mathbf{B}$ is Markov with respect to $\mathcal{G}'$. The sequence of bi-directed edges between Y_i and Y_t implies a path between b_{ij} and b_{tv} where every vertex but the endpoints is a collider in this path. Since every collider is in $\mathbf{Y}$, this path is active. □

This Proposition is illustrated by Figure 7(b). The practical implication is as follows: m-connection means that there is no further graphical property that would entail $(\mathbf{V})^{-1}_{it} = 0$ (i.e., only particular cancellations on the expression of the inverse, unlikely to happen in practice, would happen to generate such zeroes).

9. Sampling in Gaussian DAG models is still necessary if the model includes latent variables (Dunson et al., 2005).

Consider the maximal sets of vertices in an ADMG such that each pair of elements in this set is connected by a path of bi-directed edges. Following Richardson (2003), we call this a *district*.[10] It follows that is not possible in general to factorize the posterior of $\mathbf{B}$ beyond the set of districts of $\mathcal{G}$. Figure 7(c) illustrates a factorization. Fortunately, for many DMG models with both directed and bi-directed edges found in practical applications (e.g., Bollen, 1989), the maximum district size tends to be considerably smaller than the dimensionality of the problem.

6.2 Perfect Sequences

It is still important to speed up marginal likelihood (or variational bound) computations for models with districts of moderate size, particularly if many models are to be evaluated.

Without loss of generality, assume our graph $\mathcal{G}$ is a bi-directed graph with a single district, since the problem can be trivially separated into the disjoint bi-directed components. We will consider the case where the bi-directed graph is sparse: otherwise there is little to be gained by exploring the graphical structure. In that case, we will assume that the largest number of spouses of any node in $\mathcal{G}$ is bounded by a constant κ that is independent of the total number of nodes, m. The goal is to derive algorithms that are of lower complexity in m than the original algorithms.

The bottleneck of our procedure is the computation of the $\Sigma^{-1}_{nsp_{\prec}(i),nsp_{\prec}(i)}$ matrices, required in the mapping between independent and dependent Bartlett parameters (Equation 7), as well as computing the determinants $|\Sigma_{nsp_{\prec}(i),nsp_{\prec}(i)}|$. Since in sparse districts $nsp_{\prec}(i)$ grows linearly with m, the cost of a naïve algorithm for a single sampling step is $O(m^3)$ per node. Iterating over all nodes implies a cost of $O(m^4)$ for a Monte Carlo sweep. Therefore, our goal is to find a procedure by which such mappings can be computed in less than $O(m^3)$ time. The general framework is reusing previous inverses and determinants instead of performing full matrix inversion and determinant calculation for each Y_i. The difficulty on applying low-rank updates when we traverse the covariance matrix according to $\prec$ is that the sets of non-spouses $nsp_{\prec}(i)$ and $nsp_{\prec}(i+1)$ might differ arbitrarily. We want sensible orderings where such sets vary slowly and allow for efficient low-rank updates, if any.

The foundation of many scaling-up procedures for graphical models is the graph decomposition by clique separators (Tarjan, 1985), usually defined for undirected graphs. The definition for bi-directed graphs is analogous. Such a decomposition identifies overlapping *prime subgraphs* $\{\mathcal{G}_{P(1)}, \mathcal{G}_{P(2)}, \ldots, \mathcal{G}_{P(k)}\}$ of the original graph $\mathcal{G}$. A prime graph is a graph that cannot be partitioned into a triple $(\mathbf{Y}', \mathbf{S}, \mathbf{Y}'')$ of non-empty sets such that $\mathbf{S}$ is a complete separator (i.e., $\mathbf{S}$ is a clique and removing $\mathbf{S}$ disconnects the graph). Notice that a clique is also a prime subgraph.

The prime components of a graph can be ordered in a *perfect sequence* $\{\mathbf{Y}_{P(1)}, \ldots, \mathbf{Y}_{P(k)}\}$ of subsets of $\mathbf{Y}$ (Roverato, 2002; Lauritzen, 1996). Define $\mathbf{H}_j \equiv \mathbf{Y}_{P(1)} \cup \cdots \cup \mathbf{Y}_{P(j)}$ as the *history* of the perfect sequence up to the j-th subgraph. Let $\mathbf{R}_j \equiv \mathbf{Y}_{P(j)} \backslash \mathbf{H}_{j-1}$ be the *residual* of this history (with $\mathbf{R}_1 \equiv \mathbf{Y}_{P(1)}$), and $\mathbf{S}_j \equiv \mathbf{H}_{j-1} \cap \mathbf{Y}_{P(j)}$ the separator. In a perfect sequence, the triple $(\mathbf{H}_{j-1} \backslash \mathbf{S}_j, \mathbf{S}_j, \mathbf{R}_j)$ forms a decomposition of the subgraph of $\mathcal{G}$ induced by the vertex set $\mathbf{H}_j$.

Surprisingly, although bi-directed and undirected graph models have very different Markov properties (in undirected models, conditioning removes dependencies; in bi-directed models, it adds dependencies), perfect prime graph sequences prove to be also useful, but in an entirely different

10. Kang and Tian (2005) call such structures *c-components* and reserve the word "district" to refer to the function mapping a vertex to its respective c-component, as originally introduced by Richardson (2003). We choose to overload the word and call "district" both the structure and the mapping.

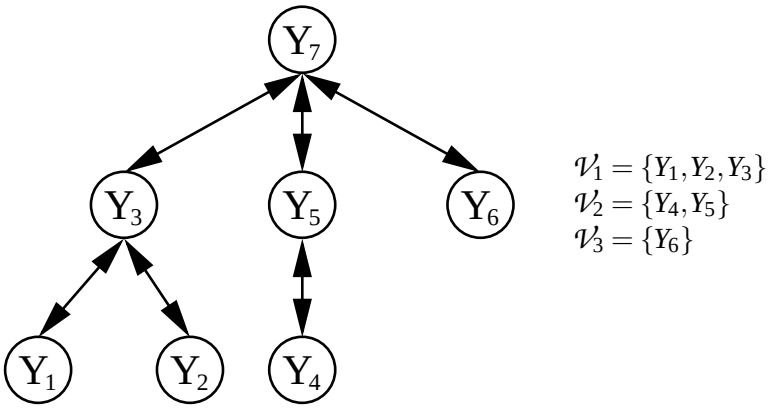


Figure 8: On the left, we have a bi-directed graph of 7 vertices arranged and ordered such that nodes are numbered by a depth-first numbering starting from "root" Y_7, with $\{Y_1, Y_2, Y_4, Y_6\}$ being leaves. Vertices $\{Y_1, Y_2, \ldots, Y_6\}$ can be partitioned as the union $\cup_{t=1}^{3} \mathcal{V}_t$, as illustrated on the right.

way. The next subsection describes the use of prime graph decompositions in a particularly interesting class of bi-directed graphs: the decomposable case. The general case is treated in the sequel.

6.2.1 DECOMPOSABLE MODELS

In a recursively decomposable graph, all prime subgraphs are cliques. We will assume that any perfect sequence in this case contains all and only the (maximal) cliques of the graph. The resulting decomposition can be interpreted as a hypergraph where nodes are the maximal cliques of the original graph, and edges correspond to the separators. In the statistics literature, a decomposable model is defined as a model that is Markov with respect to a recursively decomposable undirected graph (Lauritzen, 1996). Its widespread presence on applications of Markov random fields is due to nice computational properties, with tree-structured distributions being a particular case. Our definition of bi-directed decomposable models is analogous: a model Markov with respect to a recursively decomposable bi-directed graph.

Given the residual sequence $\{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_k\}$ obtained through a perfect sequence of maximal cliques of $\mathcal{G}$, we define a *perfect ordering* $\prec$ by numbering nodes in $\mathbf{R}_t$ before nodes in $\mathbf{R}_1, \ldots, \mathbf{R}_{t-1}$, $1 < t \leq k$ and ordering nodes according to this numbering.[11] Any ordering that satisfies this restriction is a perfect ordering. Such an ordering has the following property.

Theorem 7 *Let $\mathcal{G}$ be a recursively decomposable bi-directed graph such that the indexing of its vertices $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$ follows a perfect ordering $\prec$. Then for each $1 < i \leq m$, the set $\{Y_1, Y_2, \ldots, Y_{i-1}\}$ can be partitioned as $\cup_{t=1}^{K(i)} \mathcal{V}_t$ such that:*

1. *each $\mathcal{V}_t$ induces a connected subgraph of $\mathcal{G}$, and for each $Y_t \in \mathcal{V}_t$ and $Y_{t'} \in \mathcal{V}_{t'}$, $t \neq t'$, Y_t is not adjacent to $Y_{t'}$ in $\mathcal{G}$;*

11. Lauritzen (1996) describes other uses of perfect sequences in undirected graphs. Notice that the notion of perfect numbering described by Lauritzen (1996) is not equivalent to our notion of perfect ordering, which is always derived from a perfect sequence of prime graphs.

2. for each $\{Y_p, Y_q\} \subseteq \mathcal{V}_t$, if Y_p is a spouse of Y_i, and Y_q is a non-spouse of Y_i, then $p > q$;

The proof is in Appendix C. This result is easier to visualize in trees. One can take as a perfect ordering some depth-first ordering for a given choice of root. Then for each vertex Y_i, the set $\{Y_1, Y_2, \ldots, Y_{i-1}\}$ is partitioned according to the different branches "rooted" at Y_i. The starting point of each branch is a spouse of Y_i, and all other vertices are non-spouses of Y_i. The ordering result then follows directly from the definition of depth-first traversal, as illustrated in Figure 8.

Let Σ be the covariance matrix of a bi-directed decomposable model with graph $\mathcal{G}$, where Σ follows a $\mathcal{G}$-inverse Wishart distribution. Let $\prec$ be a perfect ordering for $\mathcal{G}$. By the construction of Bartlett's decomposition, mapping between parameters is given by

$$\Sigma_{sp_\prec(i),nsp_\prec(i)}\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)},$$

the computational bottleneck being the inversion. Notice this corresponds to the multiple regression coefficients of $sp_\prec(i)$ on $nsp_\prec(i)$. But according to Theorem 7, using a perfect ordering implies that within each $\mathcal{V}_s$ for a fixed Y_i, all preceding non-spouses of Y_i are ordered before the preceding spouses. Elements $\{Y_p, Y_q\}$ in different $\mathcal{V}_s$ are marginally independent given $\{Y_1, \ldots, Y_{i-1}\} \backslash \{Y_p, Y_q\}$. This implies that the regression coefficient of spouse Y_p on non-spouse Y_q will be zero if Y_p and Y_q are on different components $\mathcal{V}_s$, and will be identical to the previously computed $\mathcal{B}_{p,q}$ if they are in the same component. Splitting the set $\{Y_1, Y_2, \ldots Y_{i-1}\}$ into preceding spouses $\mathbf{Y}_{sp_\prec(i)}$ and non-spouses $\mathbf{Y}_{nsp_\prec(i)}$, we have

$$\begin{aligned}
\mathbf{Y}_{sp_\prec(i)} &= \mathcal{B}_{sp_\prec(i),sp_\prec(i)}\mathbf{Y}_{sp_\prec(i)} + \mathcal{B}_{sp_\prec(i),nsp_\prec(i)}\mathbf{Y}_{nsp_\prec(i)} + \varepsilon_{sp_\prec(i)} \Rightarrow \\
\mathbf{Y}_{sp_\prec(i)} &= (\mathbf{I} - \mathcal{B}_{sp_\prec(i),sp_\prec(i)})^{-1}(\mathcal{B}_{sp_\prec(i),nsp_\prec(i)}\mathbf{Y}_{nsp_\prec(i)} + \varepsilon_{sp_\prec(i)})
\end{aligned}$$

where each ε_j is an independent Gaussian with variance γ_j, and each element (p,q) in $\mathcal{B}_{sp_\prec(i),nsp_\prec(i)}$ corresponds to the known (i.e., previously computed) regression coefficient of the spouse Y_p on the non-spouse Y_q. Matrix $\mathcal{B}_{sp_\prec(i),sp_\prec(i)}$ is defined analogously. Hence, the regression coefficients of $\mathbf{Y}_{sp_\prec(i)}$ on $\mathbf{Y}_{nsp_\prec(i)}$ are given by

$$\Sigma_{sp_\prec(i),nsp_\prec(i)}\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)} = (\mathbf{I} - \mathcal{B}_{sp_\prec(i),sp_\prec(i)})^{-1}\mathcal{B}_{sp_\prec(i),nsp_\prec(i)}. \tag{19}$$

No inversion of $\Sigma_{nsp_\prec(i),nsp_\prec(i)}$ is ever necessary. Moreover, the determinant $|\Sigma_{nsp_\prec(i),nsp_\prec(i)}|$ is given by $\prod_{\{q \text{ s.t. } Y_q \in nsp_\prec(i)\}}\gamma_q$, since all non-spouses precede the spouses (which means their marginal covariance matrix is given by the previously computed Bartlett parameters).

Hence, calculating $\mathcal{B}_{i,nsp_\prec(i)}$ for all $1 \leq i \leq m$ according to a perfect ordering has as a bottleneck the inversion (of a triangular matrix) and multiplication in Equation (19), with a cost of $O(\kappa^2 + m\kappa^2)$, κ being the maximum number of spouses for any given node. The cost of the remaining operations for the i-th stage in the importance sampler is $O(\kappa^3)$. As a function of m, the cost of the parameter sampling step falls from $O(m^3)$ to $O(m)$. The cost of computing the weights is dominated by the computation of $\mathbf{K}_i$ from Equation (13), which is $O(\kappa^3 + \kappa m^2) = O(m^2)$. Figure 9 illustrates the derivation of the new ordering in a tree-structured model.

6.2.2 NON-DECOMPOSABLE MODELS

In a non-decomposable model, some prime graphs $\mathbf{Y}_{P(t)}$ will no longer be cliques. In what follows, we once again assume that $\prec$ is a perfect ordering. Unlike in the decomposable case, the product

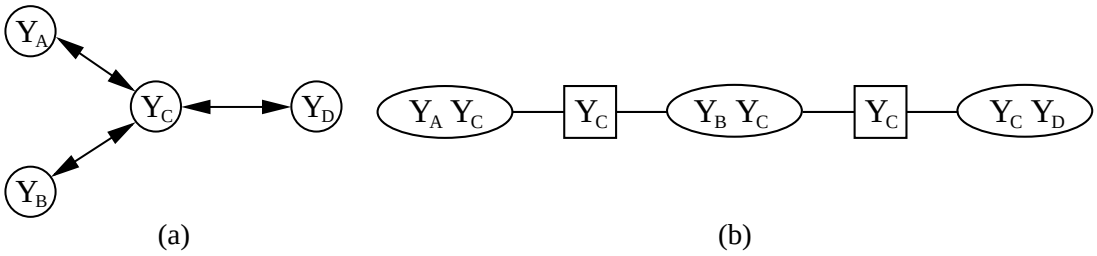


Figure 9: The tree-structured (i.e., cycle-free) bi-directed graph in (a) has as maximal cliques the adjacent pairs. Such cliques can be ordered in a perfect sequence as shown in (b), where rectangles indicate the separators. Notice that $\mathbf{R}_1 = \{Y_A, Y_C\}, \mathbf{R}_2 = \{Y_B\}, \mathbf{R}_3 = \{Y_D\}$. One possible perfect ordering is $\{Y_D, Y_B, Y_C, Y_A\}$.

$\Sigma_{sp_\prec(i),nsp_\prec(i)}\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}$ does not simplify in general. Instead we will focus only on fast methods to compute $\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}$.

As we shall see, the function of the perfect sequence is now to provide a sensible choice of which inverse submatrices $\{\Sigma^{-1}_{\mathbf{W},\mathbf{W}}\}$, $\mathbf{W} \subseteq \mathbf{Y}$, to cache and reuse when computing $\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}$. The same can be done to compute determinants $|\Sigma_{nsp_\prec(i),nsp_\prec(i)}|$.

A simple way of reusing the results from the previous section is by triangulating the non-decomposable graph $\mathcal{G}$, transforming it into a decomposable one, $\mathcal{G}'$, which is then used to generate the perfect sequence. We need to distinguish between the "true" spouses of a node Y_i in $\mathcal{G}$ and the artificial spouses in $\mathcal{G}'$ that result from the extra edges added.

Let $nsp_{\prec\mathcal{G}'}(i)$ be the non-spouses of Y_i in $\mathcal{G}'$ that precede it according to $\prec$: by construction, these are also non-spouses of Y_i in $\mathcal{G}$. Let $sp_{\Delta\prec\mathcal{G}'}(i)$ be the spouses of Y_i in $\mathcal{G}'$ that are *not* spouses of Y_i in $\mathcal{G}$. That is, the set of preceding non-spouses of Y_i in $\mathcal{G}$ is given by $nsp_\prec(i) = nsp_{\prec\mathcal{G}'}(i) \cup sp_{\Delta\prec\mathcal{G}'}(i)$.

Recall that the inverse of a partitioned matrix can be given by the following identity:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}. \quad (20)$$

In order to compute $\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}$, we consider its partitioned version

$$\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)} = \begin{pmatrix} \Sigma_{nsp_{\prec\mathcal{G}'}(i),nsp_{\prec\mathcal{G}'}(i)} & \Sigma_{nsp_{\prec\mathcal{G}'}(i),sp_{\Delta\prec\mathcal{G}'}(i)} \\ \Sigma_{sp_{\Delta\prec\mathcal{G}'}(i),nsp_{\prec\mathcal{G}'}(i)} & \Sigma_{sp_{\Delta\prec\mathcal{G}'}(i),sp_{\Delta\prec\mathcal{G}'}(i)} \end{pmatrix}^{-1}. \quad (21)$$

Let κ_{nsp} be the maximum number of non-spouses among all Y_i within any prime subgraph induced by $\mathbf{Y}_{P(t)}$. By using relation (20), where we assume for now that we know $\mathbf{A}^{-1} \equiv \Sigma^{-1}_{nsp_{\prec\mathcal{G}'}(i),nsp_{\prec\mathcal{G}'}(i)}$, the cost of computing (21) is $O(m^2\kappa_{nsp}) + O(\kappa^3_{nsp}) = O(m^2\kappa_{nsp})$ (the cost of computing $\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is $O(m^2\kappa_{nsp}) + O(\kappa^2_{nsp}) = O(m^2\kappa_{nsp})$, while the cost of inverting it is $O(\kappa^3_{nsp})$). Treating κ_{nsp} as a constant, this reduces the complexity of sampling the i-th row of Σ from $O(m^3)$ to $O(m^2)$. A similar procedure applies to the computation of the determinant $|\Sigma_{nsp_\prec(i),nsp_\prec(i)}|$, using in this case the relationship (26).

The advantage of using the perfect sequence is to allow for the computation of all $\mathbf{A}^{-1} \equiv \Sigma^{-1}_{nsp_{\prec\mathcal{G}'}(i),nsp_{\prec\mathcal{G}'}(i)}$ at a total cost, across all nodes, of $O(m^3)$: each set $nsp_{\prec\mathcal{G}'}(i)$ is guaranteed to be equal to $\{Y_1, Y_2, \ldots, Y_{lns}\}$ where Y_{lns} is the last non-spouse of Y_i in $\mathcal{G}'$ that antecedes Y_i. This follows from the result in the previous section, since all non-spouses of a node in a decomposable graph precede its spouses. Therefore, if we store the inverse covariance matrices for $\{Y_1, Y_2, \ldots, Y_i\}$, $1 \leq i \leq m$, we can obtain the required matrices $\mathbf{A}^{-1}$. This requires the storage of $O(m)$ matrices, and each matrix can be obtained by the previous one by a low-rank update (20) with a $O(m^2)$ cost.

Arbitrary orderings do not guarantee such an incremental pattern and, hence, no efficient low-rank updates. Notice that depending on the problem, many of such inverse matrices can be dynamically removed from memory if they are not used by any node placed after a particular position.

6.3 Remarks

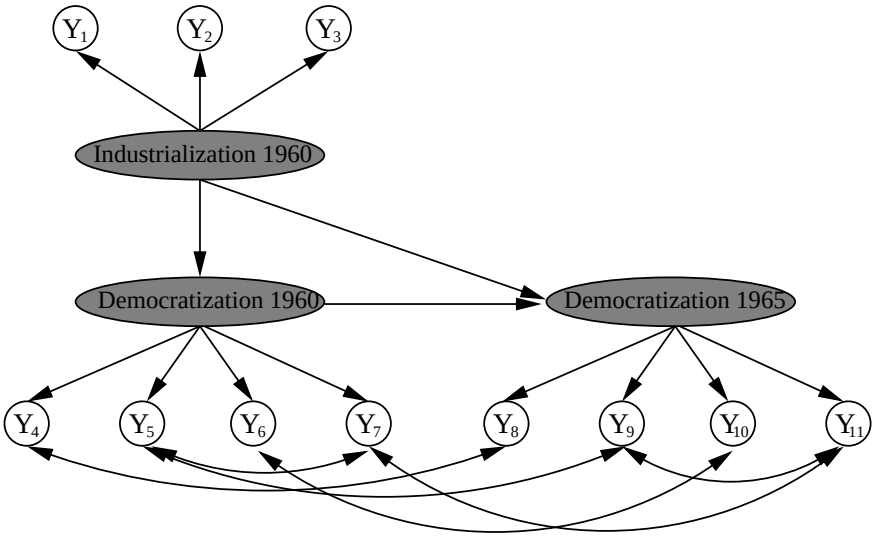
In Gaussian undirected models, the problem of covariance matrix sampling can also be reduced to sampling within each prime graph at the cost of $O(|\mathcal{P}|^4)$, $|\mathcal{P}|$ being the size of the largest prime component (Atay-Kayis and Massam, 2005). Since both κ and κ_{nsp} are $O(|\mathcal{P}|)$, our procedure costs $O(m^2|\mathcal{P}|^2 + |\mathcal{P}|^4)$ per prime graph, plus a cost of $O(m^2)$ per node to compute the importance weights. Considering a number of $m/|\mathcal{P}|$ prime graphs and $|\mathcal{P}| < m$, the total cost is $O(m^3|\mathcal{P}|)$, down from $O(m^4)$. For undirected models, the corresponding cost by sampling step using the perfect ordering decomposition is $O(m|\mathcal{P}|^3)$. The higher-order dependency on m in bi-directed models is to be expected, since the Markov blanket of any node Y_t in a connected bi-directed graph is $\mathbf{V}\backslash\{Y_t\}$. It is clear that inference with a given bi-directed graph model will never scale at the same rate of a undirected model with the same adjacencies, but this does not justify adopting an undirected representation if it is ill-suited to the problem at hand. One has also to consider that in problems with directed and bi-directed edges, the actual maximum district size might be much smaller than the number of variables. For large problems, however, further approximation schemes will be necessary. Drton and Richardson (2008b) describe some reduction techniques for transforming bi-directed edges into directed edges such that the resulting Gaussian model remains the same. As future work, such methods could be adapted to the $\mathcal{G}$-inverse Wishart sampling procedures and combined with the ordering techniques developed here into a single framework. It will also be interesting to develop similar schemes for the Gibbs sampler.

7. Experiments

We now evaluate the advantages of the Gaussian and probit models in Bayesian inference on real problems.

7.1 Industrialization and Democratization Study

Bollen (1989) describes a structural equation model of political and democratization factors within nations. "Democratization" and "industrialization" levels are abstract notions, but nevertheless of clearly observable impact. They are tied to empirical observations through different sets of *indicators*. For instance, an indicator of industrialization level is the gross national product. Hence, democratization and industrialization levels are here defined as scalar latent variables never observed directly, while the observed data is composed of indicators. In this model, there is a total of three indicators of industrialization, and four indicators of democratization. Democratization is



1. Gross national product (GNP) 1960
2. Energy consumption per capita 1960
3. Percentage of labor force in industry 1960
4. Freedom of press 1960
5. Freedom of opposition 1960
6. Fairness of elections 1960
7. Elective nature of legislative body 1960
8. Freedom of press 1965
9. Freedom of opposition 1965
10. Fairness of elections 1965
11. Elective nature of legislative body 1965

Figure 10: A directed mixed graph representing dependencies between 11 observed political and economical indicators and three latent concepts (shaded nodes) (Dunson et al., 2005; Bollen, 1989).

measured in a longitudinal study, where data was collected in two years (1960 and 1965). The indicators of democratization are pooled expert opinions summarized in an ordinal number scaled from 1 to 10. Following Bollen, we will treat the model as multivariate Gaussian, which provides an excellent fit (a p-value greater than 0.3 using a chi-square test) for a sample of 75 countries.

The corresponding mixed graph is depicted in Figure 10, along with a description of all indicators. The graph is taken from Bollen (1989). Other hidden common causes affect the democratization indicators over time, but the nature of such hidden variables is irrelevant to the problem at hand: that is, the bi-directed edges are motivated by unmeasured causes of variability in the observed indicators that exist over time. For instance, the records of freedom of press in 1960 (Y_4) and 1965 (Y_8) co-vary due to other unmeasured factors not accounted by democratization factors.

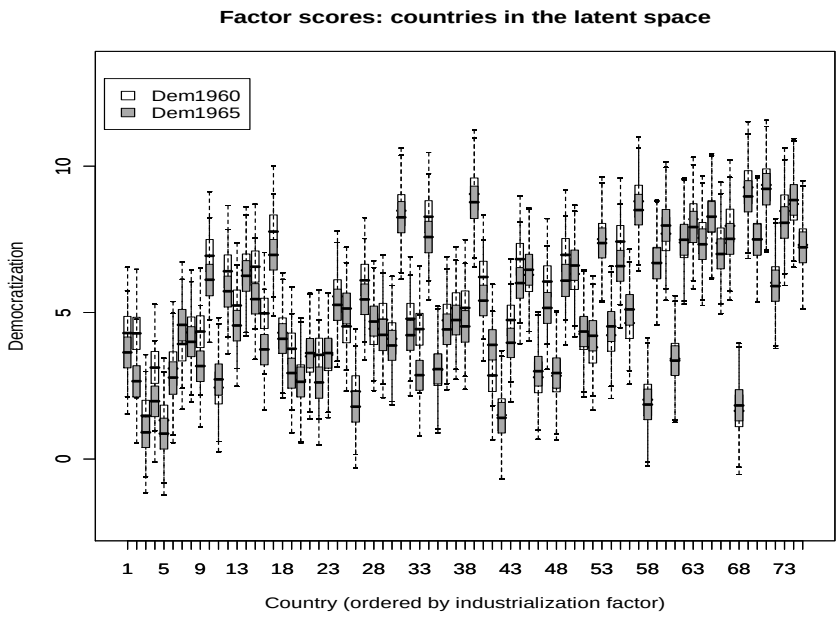


Figure 11: An embedding of 75 countries in a two-dimensional latent space: democratization level in 1960 and 1965. Boxplots of the Bayesian posterior distribution of the projection in the two dimensions are depicted in the vertical axis. Countries are arranged in the horizontal axis by the increasing order of their posterior expected industrialization level. Figure adapted from Dunson et al. (2005).

An example of Bayesian inference application is shown in Figure 11. Boxplots of the posterior values of *Democratization Level 1960* and *Democratization Level 1965* are generated. Dunson et al. (2005) use this information to, for instance, find clusters of countries in the latent space. An example of a cluster is the one formed by the bottom 16 countries in the industrialization level ranking: the growing trend of democratization levels after the first 16 countries is interrupted. This type of analysis might provide new insights to a polical scientist, for example, by revealing particular characteristics for such a group of nations.

7.1.1 EVALUATING THE MCMC ALGORITHM FOR DIFFERENT MODELS

In our analysis, we fix to unity the coefficients corresponding to the edges *Industrialization 1960* $\rightarrow Y_1$, *Democratization 1960* $\rightarrow Y_4$ and *Democratization 1965* $\rightarrow Y_8$, since the scale and sign of the latent variables is arbitrary. The intercept terms of the equations for Y_1, Y_4 and Y_8 are set to zero, since the mean of the latents is also arbitrary. The resulting model is identifiable.

We apply the Gibbs sampling procedure to three different models. The Gaussian DMG model as described in this paper, and two modified DAG models. The first DAG model is the one described by Dunson et al. (2005), where each bi-directed edge is substituted by an "ancillary" latent (as mentioned in Section 2.3). For instance, the pathway corresponding to $Y_4 \leftrightarrow Y_8$ is substituted by the chain $Y_4 \leftarrow D_{48} \rightarrow Y_8$, where D_{48} is unobserved. Dunson et al. further assume that all covariances due to such ancillary latents are positive. As such, the coefficients from D_{ij} into $\{Y_i, Y_j\}$ are set

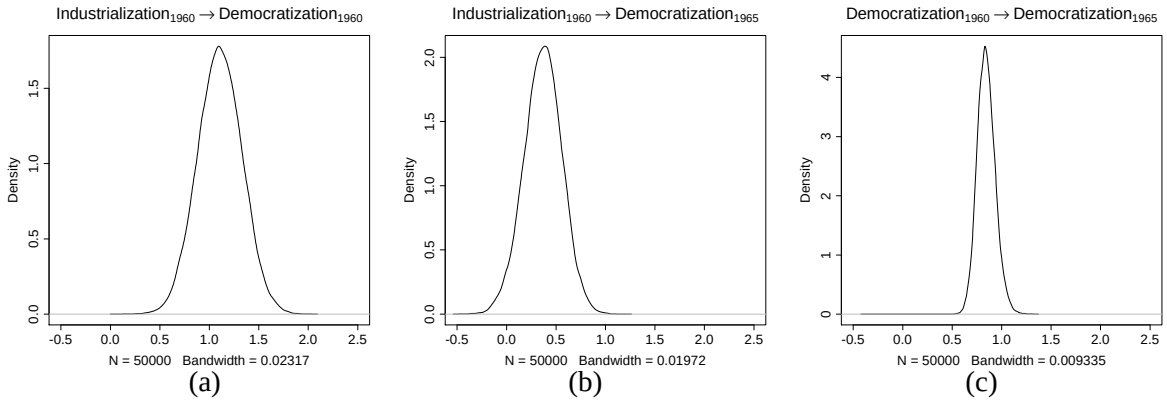


Figure 12: Posterior distribution of parameters associated with the respective edges in the industrialization/democratization domain. Smoothed posterior obtained using the output of our Gibbs sampler and the DENSITY function of R 2.6.0.

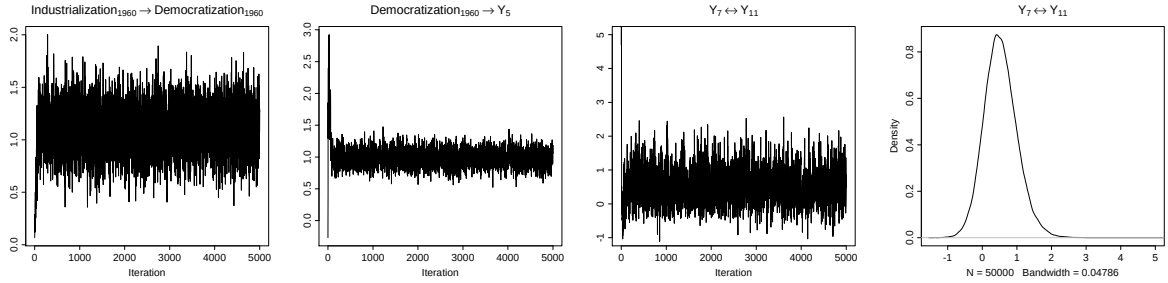


Figure 13: The first three plots show the initial 5,000 iterations of a run of the Gibbs sampling algorithm for the DMG model for three different parameters associated with edges in the graph. The last plot depicts the posterior distribution the error covariance associated with the edge $Y_7 \leftrightarrow Y_{11}$ (smoothed with the kernel density estimator from the statistical software R).

to unity, with the variance of D_{ij} corresponding to the residual covariance of $\{Y_i, Y_j\}$ given their parents. Means of ancillary latents are fixed at zero.

However, even for covariance matrices with positive covariances, this parameterization is not complete. This result is evident from the fact that the variances of Y_i and Y_j will both be larger than their covariance, which is not true of covariance matrices in general. For this particular problem, however, this extra restriction provides no measurable difference in terms of fitness. It does serve as a reminder, however, that "intuitive" parameterizations might hide undesirable constraints.

The second DAG model is an extension of the DAG model suggested by Dunson et al., the only difference being that the coefficients corresponding to edges $D_{ij} \rightarrow Y_i, i < j$, are free to vary (instead of being fixed to 1). In general, there are Gaussian DMG models that cannot be parameterized this way (Richardson and Spirtes, 2002). Notice also that because of chains such as *Democratization 1960* $\rightarrow Y_4 \leftrightarrow Y_8 \leftarrow$ *Democratization 1965*, the set of independence constraints in this graph can only be represented by a DAG if we include the ancillary latents D_{ij}. That is, there is no DAG with

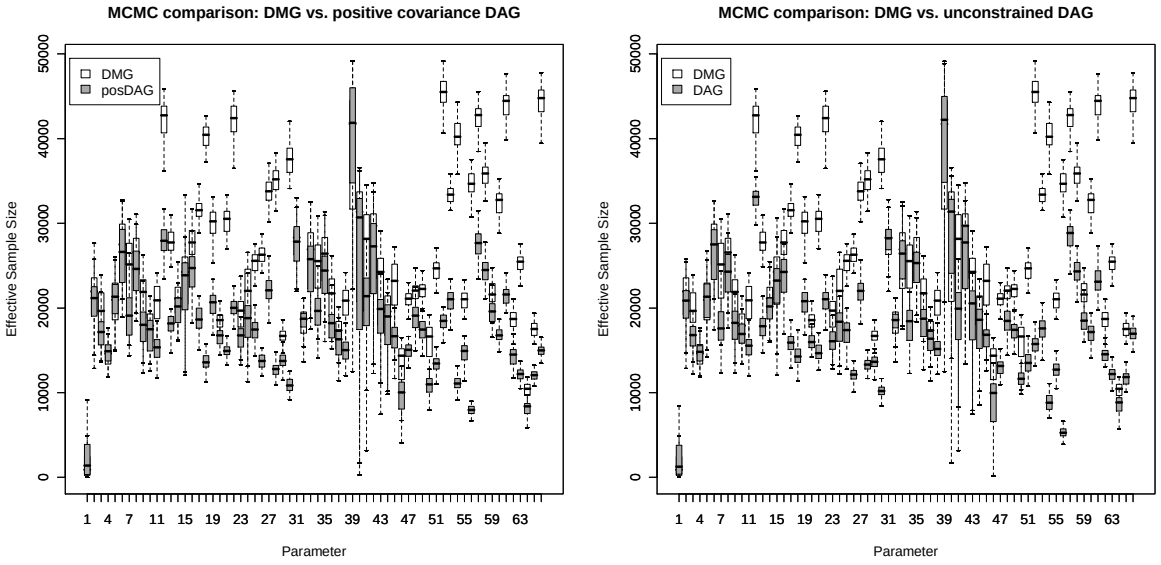


Figure 14: Comparison of the effective sample size of the MCMC algorithm applied to the three models, DMG, DAG with positive covariances (posDAG) and general DAG, as explained in the main text. The horizontal axis is the boxplot for each independent entry of the observed covariance matrix, 66 in total. The boxplots are obtained from 80 independent chains initialized randomly, where each chain runs for 50,000 iterations.

exactly the same set of independence constraints as the given DMG, unless ancillary latent variables are added.

We study the behavior of the MCMC algorithm for these three models.[12] It turns out that the mixing properties of the chain are considerably affected by the choice of model. Recall that, in the Gibbs sampling algorithm for the DMG model, a whole row of the error covariance matrix is sampled jointly conditioning on the other parameters. For the DAG models all entries of the error covariance matrix are independent and can be sampled jointly, but this requires *conditioning* on the ancillary latents, which do not exist in the DMG model and have to be sampled only in the DAG case.

For the majority of the covariance entries, the MCMC procedure mixed quite well, as illustrated in Figure 13. Notice how about 12% of the sampled DMG error covariances for $Y_7 \leftrightarrow Y_{11}$ were under zero, which could raise suspicion over the assumption of positive covariances. Autocorrelation is

12. A few technical notes: we used the priors suggested in Dunson et al. (2005), except that we changed the confidence in the prior of the covariance of the error terms $\mathbf{V}$ to be smaller (in order to minimize the influence of the priors in the models, since in this particular problem the DMG and DAG models are nearly likelihood equivalent but not posterior distribution equivalent – the priors belong to different families). We used 1 degree of freedom in our $\mathcal{G}$-Inverse Wishart, with the matrix parameter being the expected value of Dunson et al.'s prior. For the DAG models, we also used the $\mathcal{G}$-inverse Wishart prior for the error terms, but where all error terms are independent. For the DAG model with a free coefficient per ancillary latent, we assigned a standard Gaussian prior to such coefficients. The chains were initialized randomly by sampling standard Gaussians for the coefficients and latent variables. Error covariance matrices were initialized to diagonal matrices with diagonal entries sampled uniformly in $[1,2]$. Coefficient parameters were sampled jointly given the error covariance matrix and latent variables. Latent variables were also sampled jointly, given the parameters.

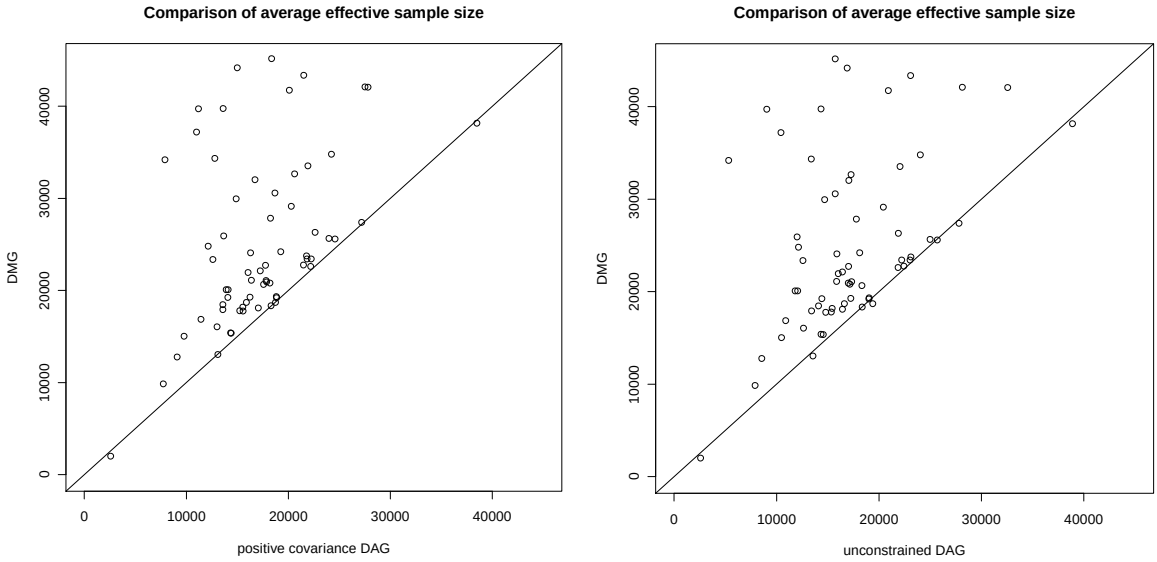


Figure 15: Comparison of the effective sample size of the MCMC algorithm applied to the three models. Here we plot the average effective sample sizes over 80 trials of 50,000 samples for each of the 66 entries of the covariance matrix. Points over the line indicate parameters where the DMG approach performed better.

essentially zero for most parameters at a lag of 50. The degree of autocorrelation, however, varied significantly between the DMG model and each DAG model. The chains for the DMG model mixed considerably better. To summarize such behavior, we calculated the effective sample size of the samples obtained from several chains. The parameters of interest in this comparison are the independent entries in the 11×11 dimensional observed covariance matrix. This is a total of 66 parameters. The effective sample size statistics were obtained by 80 independent chains of 50,000 samples each, for the three models. For each chain and each parameter, we compute the desired statistic using the EFFECTIVESIZE function implemented in the R package CODA, freely available in the Internet.

Results are summarized by boxplots in Figure 14. Parameters are ordered in the x-axis following the upper triangular covariance matrix, scanning it in the order $\{\sigma_{Y_1Y_1}, \sigma_{Y_1Y_2}, \ldots, \sigma_{Y_1Y_{11}}, \sigma_{Y_2Y_2}, \ldots, \sigma_{Y_{11}Y_{11}}\}$. White boxplots correspond to the distribution of effective sample size statistics with the DMG model across the 80 independent chains. Gray boxplots correspond to the two DAG variants. There is no significant difference between the behaviour of the Gibbs sampling procedure for the two DAG models. The procedure with the DMG model is clearly better behaved. As a summary statistic, the average effective sample size over 80 trials was steadly larger in the DMG outcome than in the positive DAG outcome (61 out of 66 parameters) and unconstrained DAG (59 out of 66). The comparison of averages is illustrated by Figure 15.

By caching the sufficient statistics of the data and factorizing the sampling procedure according to the districts of the graph, the running time for generating 50,000 samples out of the DMG model was of 34 seconds in a dual core Pentium IV 2.0 GHz. Depending on the connectivity of the bi-directed components of the graph and on the implementation of matrix inversion, sampling from the

DAG model might be faster than sampling from the DMG. In this particular study, sampling from the DAG models was substantially slower, an approximate average of 60 seconds for both variants. This can be explained by the fact that sampling latent variables is very expensive, especially considering that in the given DAG models all ancillary latents become dependent when conditioning on the data. To summarize, the DMG approach allowed for a complete parameterization with significantly better mixing properties, while still resulting in a faster MCMC procedure.

7.2 Structure Learning Applications

When trying to find a point estimate of graphical structures (i.e., returning a single graph that explains the data well), simple approaches such as testing for marginal independencies are reasonable learning algorithms under the Gaussian assumption. The Bayesian approach, however, allows one to compute odds and distributions over graphs and graph statistics, for example, the joint probability of small substructures (Friedman and Koller, 2003). Moreover, it is not clear how the independence test procedure controls for the predictive ability of the model, which is not a straightforward function of the edges that are selected due to the quantitative aspects of the dependencies.

We evaluate our Bayesian model selection contribution, focusing on the Monte Carlo sampler for bi-directed models. Jones et al. (2005) propose the following priors for graphs:

$$\mathcal{P}(\mathcal{G}|\beta) = \beta^{|E|}(1-\beta)^{0.5m(m-1)-|E|}$$

where β is a hyperparameter, $|E|$ is the number of edges in $\mathcal{G}$, and m is the number of nodes. As suggested by Jones et al., we choose $\beta = 0.5/(m-1)$, which puts more mass on graphs with $O(m)$ edges than the uniform prior.

We start with a brief synthetic study to compare the approach against a simple but effective approach based on the BIC approximation.[13] An experiment with gene expression data closes this subsection.

7.2.1 SYNTHETIC STUDIES

As a sanity check for the procedure, we generate synthetic 10-dimensional Gaussian data from models that are Markov with respect to a bi-directed graph. One hundred data sets of 50 datapoints each are generated, each coming from a different model.[14] We initially find a structure by marginal independence tests using the Fisher's Z statistic at a 0.05 level. From this starting point, we perform two searches: one using the BIC score, and the other using the marginal likelihood with a $\mathcal{G}$-*IW* prior.[15] Given the best model for each procedure, we evaluate the predictive log-likelihood on a test set of 2,000 points which are independently sampled for each of the 100 models.

13. The BIC approach is an asymptotically consistent score for selecting the maximum a posteriori Gaussian bi-directed graph model (Richardson and Spirtes, 2002).
14. The details of the simulated data are as follows: we start with DAG with no edges, with observed nodes $\{Y_1, Y_2, \ldots, Y_{10}\}$ and hidden nodes $\{X_1, X_2, X_3, X_4\}$. Each individual edge $X_i \to Y_j$ is added with probability 0.35, and no other edges are allowed. We reject graphs with fewer than 10 edges. All coefficient parameters are sampled from a standard Gaussian, and variances from an uniform distribution in $[0,1]$. The model over $\mathbf{Y}$ corresponds to a bi-directed graph, where the edge $Y_i \leftrightarrow Y_j$ exists if and only if Y_i and Y_j have a common latent parent X_k in the DAG. We then store 50 samples for the $\mathbf{Y}$ variables in a data set. The procedure is repeated 100 times with different parameters and graphical structures each time. The average number of edges in the resulting simulation was of 18.4 edges per graph.
15. In both cases, we center the data at the empirical mean of the training set and assume the data to have been generated from a zero-mean Gaussian. The $\mathcal{G}$-Inverse Wishart is an empirical prior: a diagonal matrix with the training variance

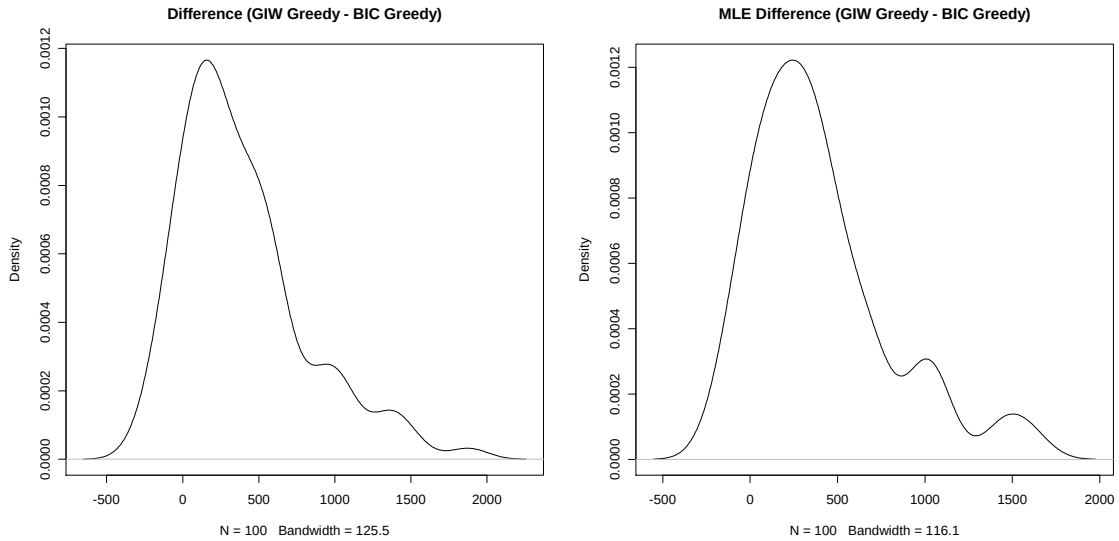


Figure 16: The differente in predictive log-likelihood with models learned with the $\mathcal{G}$-*IW* prior and the best BIC models found by greedy search. Although the difference per point is small, it reflects a persistent advantage of the full Bayesian approach. Figure (a) shows the estimated density of the distribution of differences when predicting points using the Bayesian predictive log-likelihood. Since the BIC search method does not atempt to maximize the finite sample posterior distribution, we provide Figure (b) for completeness: in this case, the predictive log-likelihood for the BIC model was calculated using the maximum likelihood estimator. The difference hardly changes, and the fully Bayesian model still wins (density estimates produced by the DENSITY$(\cdot)$ function of R 2.6.0.).

The average difference in log-likelihood prediction[16] between the structure learned with the Bayesian prior and the BIC-induced model is depicted in Figure 16(a). This is computed by conditioning on the learned structures (fully Bayesian vs. BIC maximum a posteriori graphs) and marginalizing over the posterior of the parameters. The parameter priors are those used for the structure learning step. This might be unfair for the BIC procedure, since it is not designed to maximize the finite sample posterior: hence we also show in Figure 16(b) the results obtained when the predictions given the BIC model are obtained by using the maximum likelihood estimators of the

of each variable used as the diagonal. The number of degrees of freedom is set to 1. The search is a standard greedy procedure: we evaluate the marginal log-likelihood or BIC score for each graph that differs from the current candidate by one edge (i.e., graphs with one more or one fewer edge) and pick the one with the highest score. We stop when no improvement is possible.

16. In terms of incorrect edge additions and deletions, the procedures behave about the same: an average of one third of the edges is missed, and 7% of edges are incorrectly added (individual percentages are with respect to total number of possible mistakes in each graph). Unlike BIC, however, our procedure allows for different trade-offs by using different priors. It should also be pointed out that counting edge errors is just one possible measure. A more global quantitative score such as predictive log-likelihood takes into account, indirectly, the magnitude of the errors—although it is not a direct measure of model fitness.

parameters. The average difference in the first case is 400.07, and only slightly less for the second case (389.63). Most of the mass of the difference distribution is positive (85 out of 100 for the first case, 89 out of 100 in the second case), which passes a sign test at a 0.05 level. The difference is still relatively small, suggesting that informative prior knowledge might be necessary to produce substantially better predictions.

7.2.2 GENE EXPRESSION ANALYSIS

To illustrate the use of Bayesian model selection approaches, we analyse the gene expression data previously studied by Drton and Perlman (2007), also as Gaussian bi-directed models. As before, our goal will be to compare the predictive power of models learned by greedy search with BIC and greedy search with the Bayesian posterior.

The data consists of 13 gene expression measurements from a metabolic network. A total of 118 points is available. Using all data, the BIC-induced graph has 39 edges, while the finite sample posterior graph had 44. The same procedure used in the synthetic studies, for initializing graphs and choosing priors and centering the data, was applied in this case with two choices of degrees of freedom δ for the $\mathcal{G}$-*IW* prior: $\delta = 1$ and $\delta = 5$. Preliminary experiments where 90% of the samples are assigned to the training set showed a negligible difference between methods. We then generate 10 random splits of the data, 50% of them assigned to the training set. Predictive results using the MCMC method for evaluating the Bayesian predictions (with half a million samples) are shown in Table 1. The BIC graphs are by definition the same in the three sets of evaluation, but parameters are learned in three different ways (maximum likelihood point estimation and Bayesian averaging with two different priors). There is a steady advantage for the Bayesian approach, although a small one. Notice that using Bayesian averaging over parameters given the BIC graph improves prediction when compared to using the maximum likelihood point estimate, despite the simplistic choice of prior in this study. Notice also that the cases where the Monte Carlo method has small or no advantage over the BIC method were the ones where the maximum likelihood estimators produced their best results.

7.2.3 REMARKS

The procedure based on the sampler is doable for reasonably sized problem on the order of a few dozen variables in desktop machines. Further improvements are necessary for larger problems. One aspect that was not explored here was re-using previous computations when calculating the probability of a new candidate, in a way similar to the local updates in DAG models (Chickering, 2002). How to combine local updates with the ordering-based improved sampler of Section 6 is left as future research. Several practical variations can also be implemented, such as vetoing the inclusion of edges associated with high p-values in the respective independence tests. Such tabu lists can significantly shrink the search space.

It is important to evaluate how the Monte Carlo procedure for computing normalizing constants behaves in practice. For all practical purposes, the procedure is an importance sampler and as such is not guaranteed to work within a reasonable amount of time for problems of high dimensionality (MacKay, 1998). We can, however, exploit the nature of the problem for our benefit. Notice that the procedure depends upon a choice of ordering $\prec$ for the variables. Different orderings correspond in general to *different importance distributions*. We can play with this feature to choose an suitable ordering. Consider the following algorithm for choosing an ordering given a bi-directed graph $\mathcal{G}$:

	MLE	$\delta = 1$		$\delta = 5$	
Folder	BIC	BIC	MC	BIC	MC
1	-6578.44	-6382.45	-6308.19	-6342.82	**-6296.14**
2	-6392.67	-6284.64	**-6277.94**	-6279.54	-6285.26
3	-8194.54	-6567.89	**-6433.51**	-6553.88	-6452.15
4	-6284.00	-6265.16	-6285.77	**-6252.54**	-6258.42
5	-9428.00	-6473.93	**-6400.51**	-6483.43	-6469.45
6	-7111.45	-6573.85	-6572.74	-6528.76	**-6513.02**
7	-6411.43	-6329.53	-6317.18	-6313.05	**-6309.18**
8	-6350.44	-6319.87	**-6295.19**	-6299.53	-6297.80
9	-6374.31	-6307.13	-6308.21	**-6297.47**	-6304.25
10	-7247.82	-6584.96	-6468.51	-6528.61	**-6444.55**

Table 1: Results for the 10 random splits of the gene expression data, with 50% of the points assigned to the training set. The first column shows the predictive log-likelihood for the graph learned with the BIC criterion and parameters fit by maximum likelihood. The next two columns show predictive log-likelihood results for the graphs learned with BIC and the Monte Carlo (MC) marginal likelihood method using a $\mathcal{G}$-*IW* prior with degrees of freedom $\delta = 1$. The last two columns are the results of a prior where $\delta = 5$. Best results in bold.

1. Let $\prec$ be an empty queue.

2. Let $\mathcal{G}'$ be the graph complement of $\mathcal{G}$, that is, the graph where $\{Y_i, Y_j\}$ are neighbors if and only if they are not adjacent in $\mathcal{G}$.

3. Let $\mathcal{C}$ be an arbitrary maximum clique of $\mathcal{G}'$. Add all elements of $\mathcal{C}$ to the end of $\prec$ in any arbitrary order.

4. For each pair $\{Y_i, Y_j\}$, not intersecting $\mathcal{C}$, such that the path $Y_i \leftrightarrow Y_k \leftrightarrow Y_j$ exists in $\mathcal{G}$ and $Y_k \in \mathcal{C}$, add the edge $Y_i \leftrightarrow Y_j$ to $\mathcal{G}$.

5. Remove all elements $Y_k \in \mathcal{C}$ from $\mathcal{G}$, including any edge into Y_k.

6. Iterate Steps 2-5 until $\mathcal{G}$ is an empty graph.

The resulting queue $\prec$ is an ordering that attempts to maximize the number of variables that are marginally independent given their common predecessors. This is just one possibility to simplify the importance sampling distribution: perfect orderings and the approaches for simplifying maximum likelihood estimation described by Drton and Richardson (2008b) could be adapted to provide even better orderings, but we leave this as future work.[17]

17. In our actual implementation used in the experiments in this Section, we implemented an even simpler approach: instead of finding maximum cliques, we start to build a clique from a particular node, "greedily" adding other nodes to the clique according to the column order of the data set. Each node generates a candidate clique, and we pick an arbitrary clique of maximal size to be our new set $\mathcal{C}$.

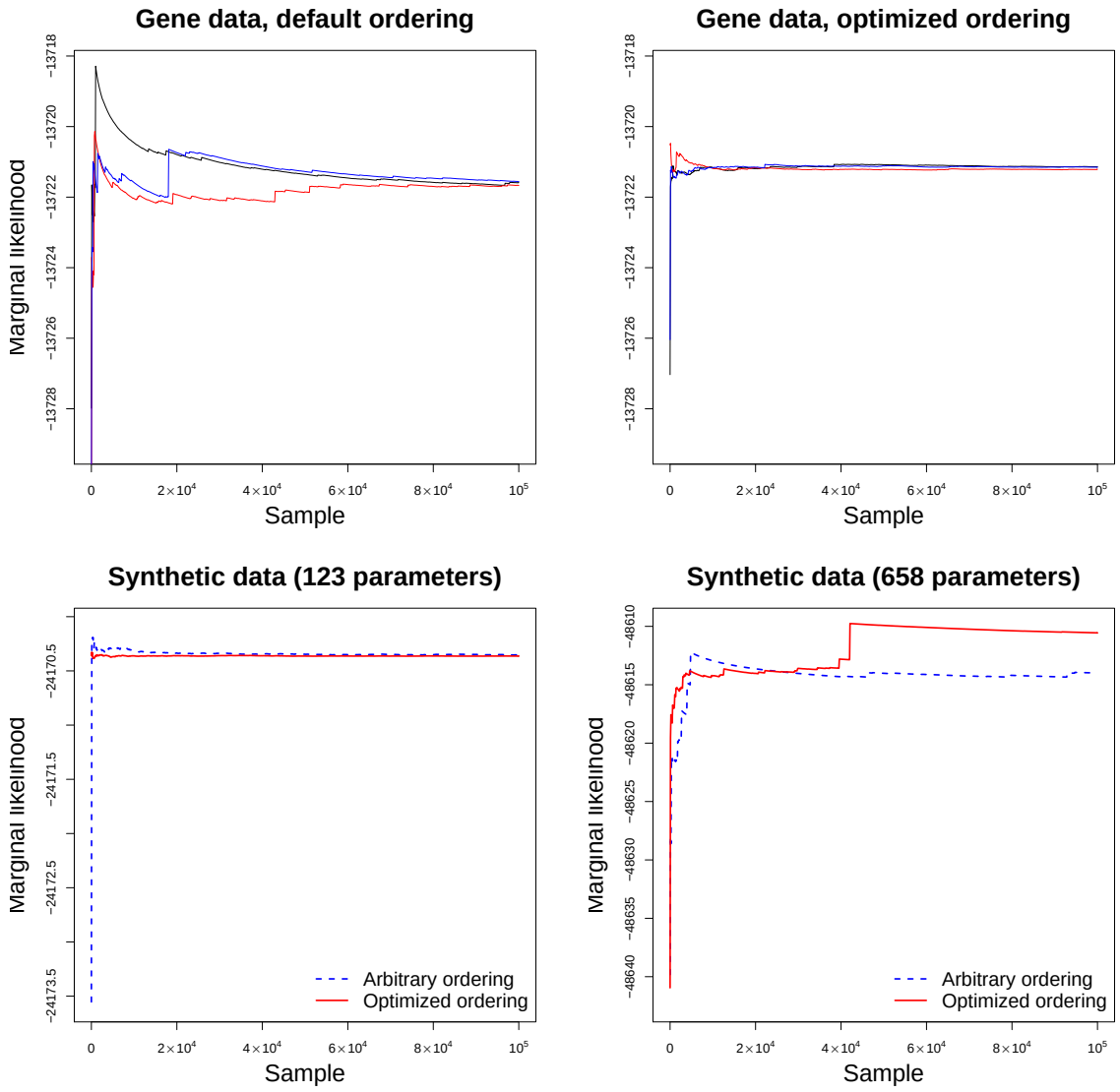


Figure 17: An evaluation on the stability of the Monte Carlo normalizing function procedure. The top row depicts the marginal likelihood estimates for the gene problem using two different distributions implied by two different orderings, as explained in the text. Experiments with synthetic data are shown in the bottom, and the bottom-right figure illustrates major differences.

Figure 17 illustrates the difference that a smart choice of ordering can make. The top left graph in Figure 17 depicts the progress of the marginal likelihood Monte Carlo estimator for the gene expression problem using the graph given by the hypothesis testing procedure. The model has 55 parameters. We obtain three estimates, each using a sample of 100,000 points, which allows us to observe how the estimates change at the initial stages. The variable ordering in this case is the

ordering present in the original database (namely, DXPS1, DXPS2, DXPS3, DXR, MCT, CMK, MECPS, HDS, HDR, IPPI1, GPPS, PPDS1 and PPDS2). The top right graph shows three runs using the optimized ordering criterion. Convergence is much faster in this case, and both samplers agree on the normalizing constant estimate.

As an illustration of the power of the procedure and its limitations, we generated a synthetic sample of 1,000 training points from a graph with 25 nodes, using the same procedure of Section 7.2.1. A run of two different samplers is shown at the bottom left of Figure 17. They are seemingly well-behaved, the ratio between the largest and median weight being at the order of one hundred in the "optimally" ordered case. In contrast, the bottom right corner of Figure 17 illustrates the method with a covariance matrix of 50 random variables and 1,000 training points. Notice this is a particularly dense graph. Much bigger jumps are observed in this case and there is no clear sign of convergence at 100,000 iterations.

While there is no foolproof criterion to evaluate the behavior of an importance sampler, the relationship between orderings provides a complementary technique: if the normalizing constant estimators vary substantially for a given set of random permutations of the variables, then the outcomes are arguably not to be trusted even if the respective estimators appear to have converged.

Concerning the choice of priors, in this Section we exploited empirical priors. The $\mathcal{G}$-Inverse Wishart matrix hyperparameter is a diagonal matrix where variance entries are the sample variances. While this adds an extra bias towards diagonal matrices, at least in our experiments we performed close to or better than other approaches—it is however not clear whether we could have done much better. It is still an open question which practical "default" hyperparameters will prove useful for the $\mathcal{G}$-*IW*. Elicitation of subjective priors in the context of structural equation models can benefit from pre-existing work on Bayesian regression, although again practical matters might be different for the $\mathcal{G}$-*IW*. Dunson et al. (2005) describe some limitations of default priors for structural equation models. A thorough evaluation of methods for eliciting subjective priors is out of the context of this work, but existing work on inverse Wishart elicitation provides a starting point (Al-Awadhi and Garthwaite, 1998). As in the case of the inverse Wishart, the $\mathcal{G}$-Inverse Wishart has a single hyperparameter for specifying degrees of freedom, a limitation which might motivate new types of priors (Brown et al., 1993).

7.3 Discrete Data Applications

We now show results on learning a discrete distribution that factorizes according to a mixed graph. Drton and Richardson (2008a) describe applications on real-world binary data modeled according to bi-directed graphs. The empirical contingency tables for two studies can be found in the corresponding technical report (Drton and Richardson, 2005). Drton and Richardson used a complete parameterization for bi-directed binary models and a maximum likelihood estimation procedure. In this section, we analyze these two data sets to illustrate the behavior of our Bayesian procedure using the probit model. Our model imposes probit constraints that are not enforced by Drton and Richardson, but it allows us to obtain Bayesian credible intervals and predictions.

The graphs used in the two studies are depicted in Figure 18. The first problem is a study on the dependence between alcoholism and depression, as shown in Figure 18(a). A data point is collected for a given pair of mono-zygotic twins. For each sibling S_i, it is recorded whether S_i is/is not alcoholic (A_i), and whether S_i suffers/does not suffer from depression (D_i). The hypothesis encoded by the graph is that alcoholism and depression do not share a common genetic cause, despite A and

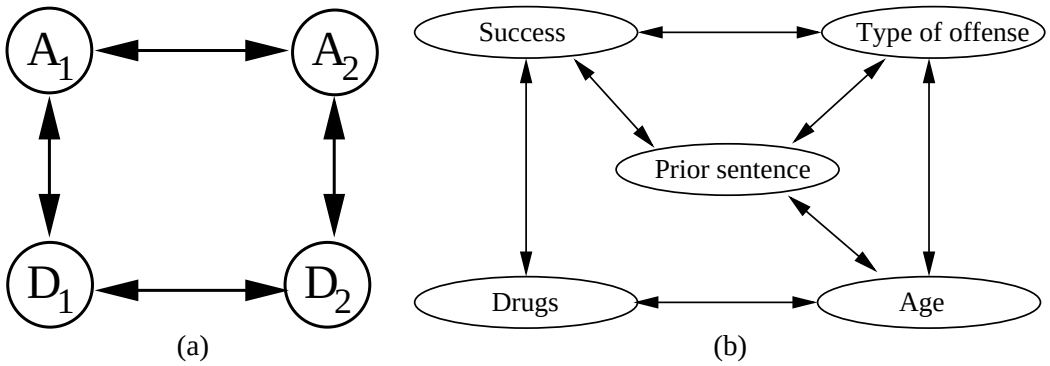


Figure 18: Two learning problems with discrete data. In (a), the graph shows dependencies concerning alcoholism (A_i) and depression (D_i) symptoms for paired twins $\{1,2\}$. In (b), a model for dependencies among features of a study on parole appeals, including the success of the parole, if the type of offense was a person offense or not, and if the offender had a dependency on drugs and was over 25 years old. All variables in these studies are binary and further details and references are provided by Drton and Richardson (2008a).

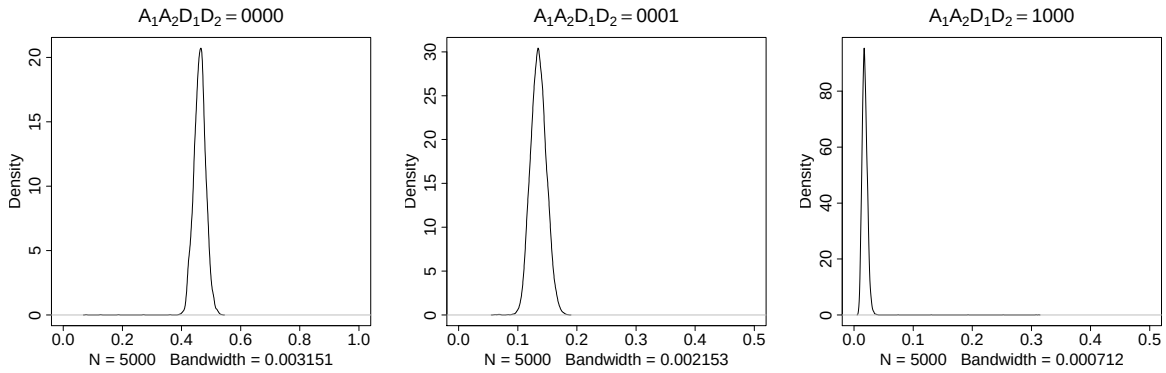


Figure 19: Posterior distribution of some of the marginal contingency table entries for the twin model.

D having some hidden (but different) genetic causes. If A and D did have genetic common causes, one would expect that the edges $A_1 \leftrightarrow D_2$ and $A_2 \leftrightarrow D_1$ would be also required. The compounded hypothesis of marginal independencies for A_i and D_j, $i \neq j$, can be tested jointly by testing a bidirected model. Notice that no reference to particular genetic hidden causes of alcoholism and depression is necessary, which again illustrates the power of modeling by marginalizing out latent variables.

The second study, as shown in Figure 18(b), concerns the dependencies among several variables in an application for parole. The model implies, for instance, that the success of a parole application (*Success* node, in the Figure) is independent of the age of the offender being under 25 (*Age* node). However, if it is known that the offender had a prior sentence, these two variables become dependent (through the path *Success* $\leftrightarrow$ *Prior sentence* $\leftrightarrow$ *Age*). As reported by Drton and Richardson, their

Entry	Estimates			Entry	Estimates		
$A_1A_2D_1D_2$	$\mathrm{E}[\Theta\|\mathcal{D}]$	MLE	uMLE	$A_1A_2D_1D_2$	$\mathrm{E}[\Theta\|\mathcal{D}]$	MLE	uMLE
0000	0.461	0.461	0.482	1000	0.018	0.018	0.013
0001	0.136	0.138	0.134	1001	0.003	0.004	0.007
0010	0.157	0.159	0.154	1010	0.021	0.020	0.013
0011	0.097	0.096	0.085	1011	0.009	0.009	0.015
0100	0.032	0.032	0.025	1100	0.008	0.010	0.005
0101	0.022	0.021	0.015	1101	0.003	0.002	0.003
0110	0.007	0.008	0.012	1110	0.003	0.005	0.007
0111	0.012	0.012	0.017	1111	0.006	0.005	0.012

Figure 20: The posterior expected value of the 16 entries in the twin study table ($E[\Theta|\mathcal{D}]$). Results generated with a chain of 5,000 points. We also show the maximum likelihood estimates of Drton and Richardson (MLE) and the maximum likelihood values obtained using an unconstrained model (uMLE). Despite the probit parameterization, in this particular study there is a reasonable agreement between the Bayesian estimator and the estimator of Drton and Richardson.

binary bi-directed model passes a significance test. Drton and Richardson also attempted to learn an undirected (Markov) network structure with this data, but the outcome was a fully connected graph. This is expected, since Markov networks cannot represent marginal independencies unless the graph is disconnected, which would introduce all sorts of other independencies and possibly not fit the data well. If many marginal independencies exist in the data generating process, Markov networks might be a bad choice of representation. For problems with symmetries such as the twin study, DAGs are not a natural choice either.

7.3.1 RESULTS

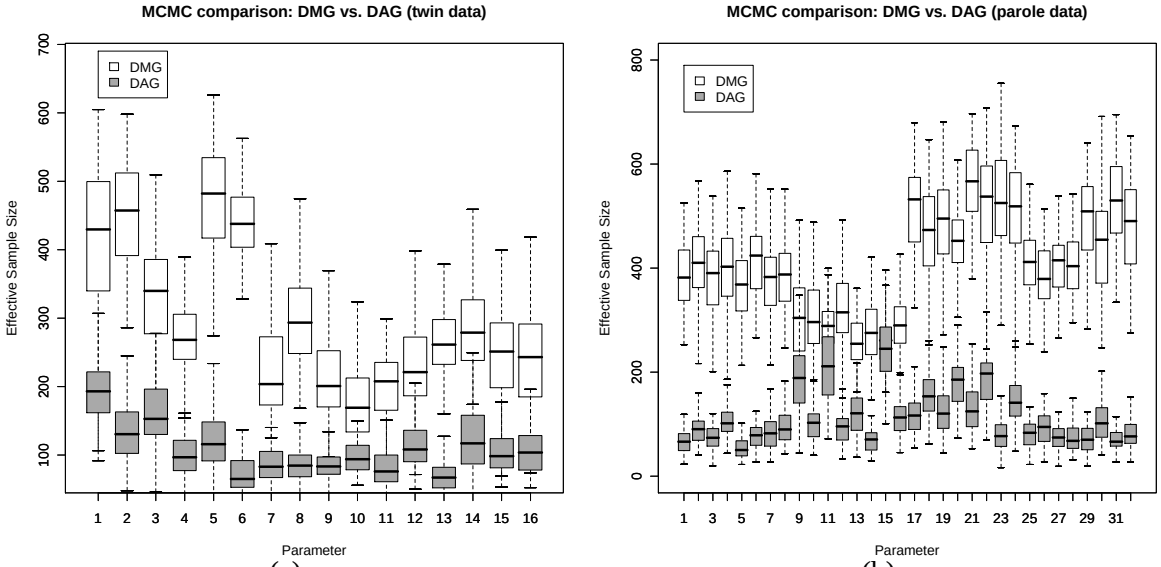
For the twin data problem, we used a simple prior for the covariance matrix of the underlying latent variables: a $\mathcal{G}$-inverse Wishart with 1 degree of freedom and a complete covariance with a value of 2 for each element in the diagonal and 1 outside the diagonals. Thresholds are fixed at zero, since we have binary data. We present the expected posterior values of the contingency table entries in Figure 20. The outcome is essentially identical to the maximum likelihood estimates of Drton and Richardson despite the probit parameterization. Moreover, with our procedure we are able to generate Bayesian confidence intervals, as illustrated in Figure 19. The results are very stable for a chain of 1,000 points.

For the parole data, we used a $\mathcal{G}$-inverse Wishart prior for the covariance matrix of underlying variables $\mathbf{Y}^\star$ with 1 degree of freedom and the identity matrix as hyperparameters. We compare the effective sample size of the Gibbs sampler for our DMG model and the DAG model obtained by using the ancillary latent parameterization of Section 7.1 for the underlying latent variable covariance matrix.[18] Boxplots for the 16 contingency table entries of the twin network and the 32 entries of the parole study are shown in Figure 21. The setup is the same as in the democratization and

18. The priors used are as follows: the ancillary representation was given a prior with mean 1 and variance 1 for the coefficients $X_{ij} \rightarrow Y_j^\star$, for $j > i$, and set constant to 1, if $i < j$. The means of the ancillary latents were fixed at

industrialization experiment, where we run 80 independent chains and plot the distribution of the effective sample sizes to measure the mixing time. We ran a shorter chain of 2,000 points, since computing the contingency table entries is expensive.

There is a substantial difference in effective sample size for the parole study. Notice that we are comparing MCMC samples for the entries in the contingency table, which in the DAG case requires integrating out not only the underlying latent variables implicit in the probit parameterization, but also the ancillary latents that account for the bi-directed edges. This hierarchy of latent variables, which does not exist in the DMG case, causes a considerable increase on autocorrelation of the chain compared to the DMG model. The standard DMG parameterization can be seen as a way of obtaining a collapsed Gibbs sampler, where the parameterization by construction reflects latent variables that were analytically marginalized.



(a) (b)

Figure 21: Comparison of effective sample sizes for the twin data (a) and parole data (b). 80 independent chains of 2,000 points were obtained using the Gibbs sampling algorithm, and the respective box-plots shown above. The Markov chain with the DMG approach easily dominates the DAG one. For the parole data, the average effective sample size for the DAG was as low as 60 points.

8. Conclusion

Directed mixed graph models are a generalization of directed graph models. Whenever a machine learning application requires directed graphs, one should first consider whether directed mixed graphs are a better choice of representation instead. DMGs represent conditional independencies of DAGs where hidden variables have been marginalized out. Given that in most applications it is

0. Variance parameters were given $(0.5, 0.5)$ inverse gamma priors, which approximately matches the priors in the DMG model.

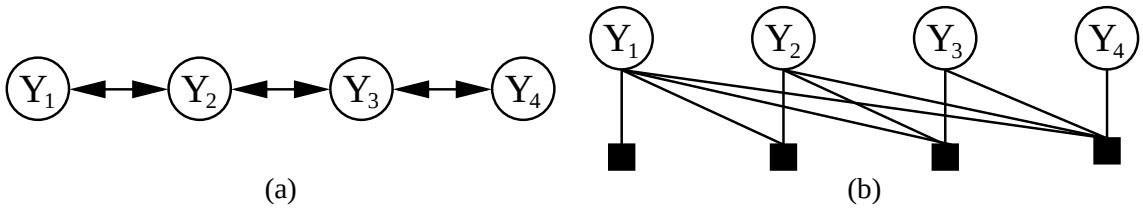


Figure 22: In (a), a simple bi-directed chain with four random variables. In (b), the respective factor graph that is obtained from a Bartlett parameterization using the ordering $\prec \equiv \{Y_1, Y_2, Y_3, Y_4\}$. In this case, the factors are $p(Y_1) \times p(Y_2|Y_1) \times p(Y_3|Y_1, Y_2) \times p(Y_4|Y_1, Y_2, Y_3)$. A different choice of ordering (e.g., the perfect ordering) could provide simpler factors on average, but the presence of a factor linked to all variables is unavoidable.

unlikely that all relevant variables are known, DMGs are a natural representation to use. In this paper, we introduced priors and inference algorithms for Bayesian learning with two popular families of mixed graph models: Gaussian and probit. We discussed some implementations and approximations to scale up algorithms. We showed examples of applications with real data, and demonstrated that Bayesian inference in Gaussian and probit DMG models using MCMC can have substantially faster mixing than in comparable DAGs.

It is part of the machine learning folklore that factor graphs can subsume directed networks. In an important sense, this is known not to be true: undirected and factor graphs only allow for monotonic independence models, where explaining away is ruled out. This excludes a vast number of realistic, non-monotonic, models. While factor graphs are perhaps the *data structures* of choice for general message-passing algorithms (e.g., Yedidia et al., 2005), they are far from being universal *modeling languages* for independencies.

What is true is that for any distribution that is Markov with respect to a DAG or DMG there is at least one corresponding factor graph model, but this is a vacuous claim of little interest: any distribution can be represented by a single-factor model involving all variables. Some will *require* a factor with all variables, even under the presence of a large number of independence constraints. For instance, a factor graph corresponding to any given bi-directed chain will necessarily include a factor node adjacent to all variable nodes, as illustrated in Figure 22. When parameterizing a distribution with many marginal independencies (e.g., a bi-directed tree), the respective factor graph would be no more than an unhelpful drawing. A better strategy for solving real-world problems is to define a family of models according to the (directed/undirected/factor) graphs of choice, and let the inference algorithm decide which re-expression of the model suits the problem. This has been traditional in graphical modeling literature (Lauritzen, 1996). The strategy adopted in this paper followed this spirit.

An alternative has been recently introduced by Huang and Frey (2008). This paper discusses graphical families of marginal independence constraints (essentially identical to bi-directed graphs, although other types of constraints might implicitly follow from the parameterization). Models are parameterized using a very different strategy. The idea is to parameterize cumulative distribution functions (CDFs) instead of densities or probability mass functions. A simple factorization criterion can be defined in the space of CDFs, but densities have to be computed by a novel message-passing

scheme. The particular application discussed by Huang and Frey (2008) could in principle be approached using the Gaussian bi-directed probit model of Section 5, but the parameterization in Huang and Frey (2008) does not need to rely on Gaussian distributions. It is not clear, however, how to efficiently perform Bayesian inference in this case and which constraints are implicitly implied by the different choices of parameterization. The different perspective given by products of CDFs is novel and promising. It should point out to new directions in mixed graph modeling.

The structural equation modeling literature also describes several pragmatic ways of specifying non-linearities in the structural equations (Lee, 2007). Less common is the specification of non-Gaussian models for the joint density of the error terms. Silva and Ghahramani (2009) introduce a flexible mixture of Gaussians approach for models of marginal independence. There is a need on how to combine this approach with flexible families of structural equations in a computationally efficient way. Also, models with non-additive error terms remain to be explored.

Current interest in estimating sparse statistical models has lead to approaches that estimate structured covariance matrices (e.g., Bickel and Levina, 2008). This development could also lead to new families of priors. In particular, different matrix decompositions have motivated different ways of specifying priors on covariance matrices. For instance, Chen and Dunson (2003) propose a modified Cholesky decomposition for the covariance matrix of random effect parameters: standard deviations are parameterized separately with a prior that puts positive mass on zero variances (effectively allowing the random effect to be neutralized). Wong et al. (2003) describe a prior for inverse correlation matrices that is uniform conditioned on the number of structural zeros. Metropolis-Hastings schemes are necessary in this case.

Shrinkage methods have also been applied to the estimation of covariance matrices. A common approach, shrinkage towards a diagonal matrix (e.g., Daniels and Kass, 1999), could be generalized towards some sparse matrix corresponding to a bi-directed graph. Although shrinkage will not generate structural zeros in the resulting matrix, allowing for sparse shrinkage matrices other than the identity matrix could be interesting in prediction problems.

Some approaches can exploit an ordering for the variables, which is natural in some domains such as time-series analysis. While the $\mathcal{G}$-Inverse Wishart is invariant to a permutation of the variables, new types of priors that exploit a natural variable ordering should be of interest, as in the original work of Brown et al. (1993) that motivated our approach.

Other directions and applications are suggested by recent papers:

- **learning measurement models:** the industrialization and democratization problem of Section 7.1 provides an example of a measurement model. In such a family of problems, observed variables are children of latent variables, and connections from latents to observables define the measurement model. Sparsity in the measurement can be exploited to allow for more general dependencies connecting latent variables. One role of the bi-directed component is to allow for extra dependencies connecting observed variables that are not accounted by the explicit latent variables in the model. Silva et al. (2006) describes a learning algorithm for mixed graph measurement models using the "ancillary" parameterization. The natural question is which alternative optimization strategies could be used and how to scale them up;

- **structural and relational learning:** in prediction problems where given an input vector $\mathbf{X}$ we have to predict an output vector $\mathbf{Y}$, the dependence structure of $\mathbf{Y}$ given $\mathbf{X}$ can also lie within the directed mixed graph family. Silva et al. (2007) introduces mixed graph models within the context of relational classification, where $\mathbf{Y}$ are labels of different data points

not independently distributed. In such a class of problems, novel kinds of parameterization are necessary since the dimensionality of the covariance matrix increases with the sample size. Structural features of the graph are used to propose different parameterizations of the dependencies, and many other alternatives are possible;

- **causal inference:** mixed graphs have been consistently used as a language for representing causal dependencies under unmeasured confounding. Zhang (2008) describes recent advances in identifying causal effects with ancestral graphs. Algorithms for learning mixed graph structures are described by Spirtes et al. (2000) and the recent advances in parameterizing such models should result in new algorithms;

Many challenges remain. For instance, more flexible models for DMG discrete models are being developed (Drton and Richardson, 2008a), but for large graphs they pose a formidable computational problem. An important question is which other less flexible, but more tractable, parameterizations could be used, and which approximation algorithms to develop. The probit family discussed here was a choice among many. The parameterization by Drton and Richardson (2008a) could be a starting point for trading-off flexibility and computational effort. And while it is true that Gaussian copula models (Pitt et al., 2006) can be adapted to generalize the approach introduced here, it remains to be seen if other copula parameterizations easily lead to DMG models.

Acknowledgments

We thank the anonymous reviewers for their suggestions, Kenneth Bollen for providing us with the industrialization and democratization data set, and Robert Gramacy for helpful discussions. An earlier version of this paper (Silva and Ghahramani, 2006) appeared in the proceedings of the Uncertainty in Artificial Intelligence conference. This work was funded by the Gatsby Charitable Foundation and a EPSRC grant #EP/D065704/1.

Appendix A. Deriving the Sampling Distribution for the Monte Carlo Computation of Normalizing Constants

We give here the details on how to derive the sampling distribution used for computing normalizing constants $I_{\mathcal{G}}(\delta, \mathbf{U})$, as described in Section 3.2.2.

Let $\mathbf{A}_i \equiv \Sigma_{sp_\prec(i),nsp_\prec(i)}\Sigma^{-1}_{nsp_\prec(i),nsp_\prec(i)}$. Recall from Equation (7) that $\mathcal{B}_{i,nsp_\prec(i)} = -\mathcal{B}_{i,sp_\prec(i)}\mathbf{A}_i$. The original density $p(\mathcal{B}_i \mid \gamma_i)$, as given by Lemma 1, is a multivariate Gaussian with the following kernel:

$$\exp\left(-\frac{1}{2\gamma_i}\begin{bmatrix} \mathcal{B}^{\mathsf{T}}_{i,sp_\prec(i)} - \mathbf{M}_{sp_\prec(i)} \\ \mathcal{B}^{\mathsf{T}}_{i,nsp_\prec(i)} - \mathbf{M}_{nsp_\prec(i)} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{U}_{ss} & \mathbf{U}_{sn} \\ \mathbf{U}_{ns} & \mathbf{U}_{nn} \end{bmatrix} \begin{bmatrix} \mathcal{B}^{\mathsf{T}}_{i,sp_\prec(i)} - \mathbf{M}_{sp_\prec(i)} \\ \mathcal{B}^{\mathsf{T}}_{i,nsp_\prec(i)} - \mathbf{M}_{nsp_\prec(i)} \end{bmatrix}\right) \tag{22}$$

where $\mathbf{U}_{\{i-1\},\{i-1\}}$ in Lemma 1 was rearranged above as the partitioned matrix in (14). The pair $\{\mathbf{M}_{sp_\prec(i)}, \mathbf{M}_{nsp_\prec(i)}\}$ corresponds to the respective partition of the mean vector $\mathbf{M}_i$. Plugging in the expression for $\mathcal{B}_{i,nsp_\prec(i)}$ in (22), we obtain the modified kernel

$$\exp\left(-\frac{1}{2\gamma_i}\begin{bmatrix}\mathcal{B}_{i,sp_\prec(i)}^\mathsf{T}-\mathbf{M}_{sp_\prec(i)}\\ -\mathbf{A}_i^\mathsf{T}\mathcal{B}_{i,sp_\prec(i)}^\mathsf{T}-\mathbf{M}_{nsp_\prec(i)}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{U}_{ss} & \mathbf{U}_{sn}\\ \mathbf{U}_{ns} & \mathbf{U}_{nn}\end{bmatrix}\begin{bmatrix}\mathcal{B}_{i,sp_\prec(i)}^\mathsf{T}-\mathbf{M}_{sp_\prec(i)}\\ -\mathbf{A}_i^\mathsf{T}\mathcal{B}_{i,sp_\prec(i)}^\mathsf{T}-\mathbf{M}_{nsp_\prec(i)}\end{bmatrix}\right) \quad (23)$$

which can be rewritten as

$$\begin{aligned} p_b(\mathcal{B}_{i,sp_\prec(i)};\mathbf{K}_i\mathbf{m}_i,\gamma_i\mathbf{K}_i) &\times (2\pi)^{\#sp_\prec(i)/2}|\gamma_i|^{\#sp_\prec(i)/2}|\mathbf{K}_i(\Phi_{i-1})|^{1/2}\\ &\times \exp\left\{-\frac{1}{2}\gamma_i^{-1}\mathcal{U}_i\right\} \end{aligned} \quad (24)$$

where $\#sp_\prec(i)$ is the size of set $sp_\prec(i)$, $p_b(\cdot;\alpha,\Sigma)$ is the density function of a multivariate Gaussian distribution with mean α and covariance Σ, $\mathbf{K}_i(\Phi_{i-1})\equiv\mathbf{K}_i$ to emphasize the contribution of previous parameters, and

$$\begin{aligned} \mathbf{m}_i &= (\mathbf{U}_{ss}-\mathbf{A}_i\mathbf{U}_{ns})\mathbf{M}_{sp_\prec(i)}+(\mathbf{U}_{sn}-\mathbf{A}_i\mathbf{U}_{nn})\mathbf{M}_{nsp_\prec(i)},\\ \mathbf{K}_i^{-1} &= \mathbf{U}_{ss}-\mathbf{A}_i\mathbf{U}_{ns}-\mathbf{U}_{sn}\mathbf{A}_i^\mathsf{T}+\mathbf{A}_i\mathbf{U}_{nn}\mathbf{A}_i^\mathsf{T},\\ \mathcal{U}_i &= \mathbf{M}_i^\mathsf{T}\mathbf{U}_{\{i-1\},\{i-1\}}\mathbf{M}_i-\mathbf{m}_i^\mathsf{T}\mathbf{K}_i\mathbf{m}_i. \end{aligned}$$

If $sp_\prec(i)=\emptyset$, it follows that $\mathcal{B}_i=\mathcal{B}_{i,nsp_\prec(i)}=0$. The kernel (23) reduces to $\exp(-0.5\mathcal{U}_i/\gamma_i)$, and $\mathcal{U}_i\equiv\mathbf{M}_i^\mathsf{T}\mathbf{U}_{\{i-1\},\{i-1\}}\mathbf{M}_i$. If $nsp_\prec(i)=\emptyset$, then the expression for the kernel does not change ($\mathcal{U}_i\equiv 0$), and Equation (24) corresponds to the original kernel in Equation (11).

Inserting the re-expressed kernel into the original function (11), we obtain

$$p_b(\mathcal{B}_{i,sp_\prec(i)};\mathbf{K}_i\mathbf{m}_i,\gamma_i\mathbf{K}_i)p_g\left(\gamma_i;\frac{\delta+i-1+\#nsp_\prec(i)}{2},\frac{u_{ii.\{i-1\},\{i-1\}}+\mathcal{U}_i}{2}\right)f_i(\Phi_{i-1})$$

where $p_g(\cdot;\alpha,\beta)$ is an inverse gamma density function and

$$\begin{aligned} f_i(\Phi_{i-1}) &\equiv (2\pi)^{-\frac{(i-1)-\#sp_\prec(i)}{2}}|\mathbf{K}_i(\Phi_{i-1})|^{1/2}|\mathbf{U}_{\{i-1\},\{i-1\}}|^{1/2}\\ &\times \frac{(u_{ii.\{i-1\},\{i-1\}}/2)^{(\delta+i-1)/2}}{\Gamma((\delta+i-1)/2)}\frac{\Gamma((\delta+i-1+\#nsp_\prec(i))/2)}{((u_{ii.\{i-1\},\{i-1\}}+\mathcal{U}_i)/2)^{(\delta+i-1+\#nsp_\prec(i))/2}}. \end{aligned}$$

Appendix B. Variational Updates for Gaussian Mixed Graph Models

The variational updates for the coefficient and intercept parameters are essentially identical to their joint conditional distribution given $\mathbf{V}$ and $\mathbf{X}$, where occurrences of $\mathbf{V}$ and $\mathbf{X}$ are substituted by expectations $\langle\mathbf{V}^{-1}\rangle_{q(\mathbf{V})}$ and $\langle\mathbf{X}\rangle_{q(\mathbf{X})}$, respectively. Let $\mathcal{V}_{ij}$ be the ij-th entry of $\langle\mathbf{V}^{-1}\rangle_{q(\mathbf{V})}$. The covariance matrix of $(\mathbf{B},\alpha)$ is the covariance matrix of the vector $vec(\mathbf{B},\alpha)$. Such vector is constructed using all (non-zero) coefficients and intercepts. We denote this covariance matrix by $\Sigma_{\mathbf{B},\alpha}$. For simplicity of notation, we will treat α_i as the coefficient $b_{i(m+1)}$, m being the number of variables. We will also adopt the notation $Y_{m+1}^{(d)}\equiv 1$ in the following derivations. As an abuse of notation, let $\mathbf{Y}$ also

refer to latent variables. In this case, if Y_i and Y_j refer to latent variables X_{h_i} and X_{h_j}, then define $Y_i \equiv \langle X_{h_i} \rangle_{q(\mathbf{X})}$, and $Y_i Y_j \equiv \langle X_{h_i} X_{h_j} \rangle_{q(\mathbf{X})}$.

Let b_{ij} and b_{tv} be the r-th and s-th entries of $vec(\mathbf{B}, \alpha)$, respectively. The rs-th entry of the inverse matrix $\Sigma^{-1}_{\mathbf{B}\alpha}$ is given by

$$(\Sigma^{-1}_{\mathbf{B}\alpha})_{rs} \quad = \quad \mathcal{V}_{it} \sum_{d=1}^{n} Y_j^{(d)} Y_v^{(d)} + 1(i=t)1(j=v)\frac{c^b_{ij}}{s^b_{ij}}$$

where $b_{xp_x} \equiv 0$ if no edge $Y_x \leftarrow Y_{p_x}$ exists in the graph, $1(\cdot)$ is the indicator function, and c^b_{ij}, s^b_{ij} are the given prior parameters defined in Section 4. Similarly to the factorization criterion explained in Section 6, the matrix $q(\mathbf{V})$ will in general be block-diagonal, and this summation can be highly simplified.

Define now a vector $\mathbf{c}^b$ analogous to the Gibbs sampling case, where

$$c^b_r \quad = \quad \sum_{t=1}^{m} \mathcal{V}_{it} \sum_{d=1}^{n} Y_j^{(d)} Y_t^{(d)} + \frac{c^b_{ij}}{s^b_{ij}}.$$

The variational distribution $q(\mathbf{B}, \alpha)$ is then a $N(\Sigma_{\mathbf{B},\alpha}\mathbf{c}, \Sigma_{\mathbf{B},\alpha})$. The variational distribution for the latent variables will exactly the same as the Gibbs distribution, except that references to $\mathbf{B}, \alpha, \mathbf{V}^{-1}$ are substituted by $\langle \mathbf{B} \rangle_{q(\mathbf{B},\alpha)}, \langle \alpha \rangle_{q(\mathbf{B},\alpha)}$ and $\langle \mathbf{V}^{-1} \rangle_{q(\mathbf{V})}$.

Appendix C. Proofs

Proof of Lemma 2: Arrange the columns of the Jacobian such that their order corresponds to the sequence $\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \sigma_{32}, \sigma_{33}, \ldots, \sigma_{mm}$, excluding the entries σ_{ij} that are identically zero by construction. Arrange the rows of the Jacobian such that their order corresponds to the sequence $\gamma_1, \beta_{21}, \gamma_2, \beta_{31}, \beta_{32}, \ldots, \gamma_m$, excluding the entries β_{ij} that are not in $\Phi_{\mathcal{G}}$ (i.e., exclude any β_{ij} corresponding to a pair $\{Y_i, Y_j\}$ that is not adjacent in the graph).

By the definition of Bartlett's decomposition, $\Sigma_{\{i\},\{i\}}$ and β_{st} are functionally independent for $s > i$. The same holds for $\Sigma_{\{i\},\{i\}}$ and γ_s. As such, $\partial\sigma_{ij}/\partial\beta_{st} = 0$ and $\partial\sigma_{ij}/\partial\gamma_s = 0$ for $s > i$. This implies that $J(\Phi_{\mathcal{G}})$ is a (lower) block triangular matrix of $2m-1$ blocks: for k odd, the k-th block is the singleton $\partial\sigma_{ii}/\partial\gamma_i = 1$, where $i = (k+1)/2$. For k even, the k-th block is the Jacobian $\partial\Sigma_{i,sp_{\prec(i)}}/\partial\mathcal{B}_{i,sp_{\prec(i)}}$, where $i = 1 + k/2$ and $\Sigma_{i,sp_{\prec(i)}}$ is the vector of covariances of Y_i and its preceding spouses.

From the interpretation given by Equation (8), it follows that $\mathcal{B}_{i,sp_{\prec(i)}}$ can also be defined by the regression of Y_i on $\mathbf{Z}_i$. That is

$$\mathcal{B}_{i,sp_{\prec(i)}} = \Sigma_{Y_i,\mathbf{Z}_i}\Sigma^{-1}_{\mathbf{Z}_i,\mathbf{Z}_i} \equiv \Sigma_{Y_i,\mathbf{Z}_i}R_i^{-1}. \tag{25}$$

However, $\Sigma_{Y_i,\mathbf{Z}_i} = \Sigma_{i,sp_{\prec(i)}}$, since Y_i is independent of its non-spouses. From (25) we get $\Sigma_{i,sp_{\prec(i)}} = \mathcal{B}_{i,sp_{\prec(i)}}R_i$, and as such the submatrix $\partial\Sigma_{i,sp_{\prec(i)}}/\partial\mathcal{B}_{i,sp_{\prec(i)}}$ turns out to be R_i.

Since the determinant of the block triangular Jacobian $J(\Phi_{\mathcal{G}})$ is given by the determinant of the blocks, this implies

$$|J(\Phi_{\mathcal{G}})| = \prod_{i=2}^{m} |R_i|.$$

By the matrix identity

$$\left| \begin{array}{cc} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{array} \right| = |\mathbf{A}||\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}|, \tag{26}$$

$|\Sigma_{\{i-1\},\{i-1\}}| = |\Sigma_{nsp_{\prec(i)},nsp_{\prec(i)}}||\Sigma_{sp_{\prec(i)},sp_{\prec(i)}} - \Sigma_{sp_{\prec(i)},nsp_{\prec(i)}}\Sigma^{-1}_{nsp_{\prec(i)},nsp_{\prec(i)}}\Sigma_{nsp_{\prec(i)},sp_{\prec(i)}}| \equiv |\Sigma_{nsp_{\prec(i)},nsp_{\prec(i)}}||R_i|$. Since $|\Sigma_{\{i-1\},\{i-1\}}| = \prod_{t=1}^{i-1}\gamma_t$, the second equality holds. □

Proof of Theorem 4: We first describe a mapping from each path in $\mathcal{G}$ to a path in $\mathcal{G}^\star$, and vice-versa (such mappings are not inverse functions of each other, since the number of paths in $\mathcal{G}^\star$ is larger than in $\mathcal{G}$). By construction, all bi-directed edges in $\mathcal{G}^*$ have two UVs as endpoints, with an one-to-one mapping between each $Y_s^\star \leftrightarrow Y_t^\star$ in $\mathcal{G}^\star$ and each $Y_s \leftrightarrow Y_t$ in $\mathcal{G}$. All directed edges in $\mathcal{G}^\star$ are of two types: $Y_s \rightarrow Y_t^\star$, with $s \neq t$, or $Y_s^\star \rightarrow Y_s$. Therefore, one can define an unique path P in $\mathcal{G}$ as a function of a path $P^\star$ in $\mathcal{G}^\star$, obtained by relabeling each $Y^\star$ as Y, and by collapsing any $Y \rightarrow Y$ edges that might result from this relabeling into a single vertex Y. A mapping in the opposite direction is analogous as given by the construction rule of Type-II models.

A collider in a path is any vertex within a head-to-head collision in the path, that is, any vertex Y_t where the preceding and the next vertex in the path are connected to Y_t with an edge (directed or bi-directed) into Y_t. Y_i and Y_j are m-separated by $\mathbf{Z}$ in an acyclic DMG if and only if there is no active path connecting Y_i and Y_j. Like in d-separation, a path is active if all of its colliders have some descendant in $\mathbf{Z}$, and none of its non-colliders is in $\mathbf{Z}$ (Richardson, 2003). The mappings between paths P and $P^\star$ are such that Y_t is a collider in P if and only if Y_t is in $P^\star$ and is a collider, or $Y_t^\star$ is in $P^\star$ and is a collider. Since by construction any $Y_t^\star$ will have the same $\mathbf{Y}$-descendants in $\mathcal{G}^\star$ as Y_t has in $\mathcal{G}$, and $\mathbf{Z} \subset \mathbf{Y}$, the result follows. □

Proof of Theorem 7: The first of the two claims of the theorem trivially holds, since connectivity is a transitive property and as such this partition will always exist (where $K(i) = 1$ is a possibility). We will prove the validity of the second claim by induction. Let $\{\mathbf{R}_1, \ldots, \mathbf{R}_k\}$ be the perfect sequence that generated our perfect ordering. The second claim automatically holds for all vertices in $\mathbf{R}_k$, since $\mathbf{R}_k$ is a clique.

Assume the second claim holds for the subsequence $\{\mathbf{R}_{l+1}, \mathbf{R}_{l+2}, \ldots, \mathbf{R}_k\}$. Let Y_i be an element of $\mathbf{R}_l$. Assume there is some non-spouse Y_q of Y_i in $\mathbf{R}_{l'}$, and some spouse Y_p of Y_i in $\mathbf{R}_{l''}$, such that $l < l' \leq l''$. We will assume that both Y_q and Y_p belong to the same component $\mathcal{V}_t$ and show this leads to a contradiction.

Without loss of generality, we can assume that Y_q and Y_p are adjacent: otherwise, the fact that Y_q and Y_p are in the connected set $\mathcal{V}_t$ will imply there is a path connecting Y_q and Y_p in the subgraph induced by $\{\mathbf{R}_{l+1}, \ldots, \mathbf{R}_k\}$. We can redefine $\{Y_q, Y_p\}$ to be the endpoints of the first edge in the path containing a non-spouse and a spouse of Y_i. It will still be the case that $q > p$, by the induction hypothesis.

Since $Y_p \in \mathbf{R}_{l''}$, there is a separator $\mathbf{S}_{l''}$ between $\mathbf{H}_{l''} \backslash \mathbf{S}_{l''}$ and $\mathbf{R}_{l''}$. But $Y_i \in \mathbf{H}_{l''}$, and Y_i is adjacent to Y_p, which implies $Y_i \in \mathbf{S}_{l''}$. If $l' < l''$, this will also imply that $Y_q \in \mathbf{S}_{l''}$, which is a contradiction, since $\mathbf{S}_{l''}$ is a complete set. If $l' = l''$, this implies that Y_i and Y_q are both in $\mathbf{Y}_{P(l'')}$, which is also a contradiction since $\mathbf{Y}_{P(l'')}$ is a clique. □

References

A. Al-Awadhi and P. Garthwaite. An elicitation method for multivariate normal distributions. *Communications in Statistics - Theory and Methods*, 27:1123–1142, 1998.

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

A. Atay-Kayis and H. Massam. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92:317–335, 2005.

D. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis*. Arnold Publishers, 1999.

M. Beal. Variational algorithms for approximate Bayesian inference. *PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London*, 2003.

M. Beal and Z. Ghahramani. Variational Baeysian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1:793–832, 2006.

P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36:2577–2604, 2008.

K. Bollen. *Structural Equation Models with Latent Variables*. John Wiley & Sons, 1989.

P. Brown, N. Le, and J. Zidek. Inference for a covariance matrix. *In P.R. Freeman, A.F.M. Smith (editors), Aspects of Uncertainty, a tribute to D. V. Lindley*, pages 77–92, 1993.

Z. Chen and D. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59:762–769, 2003.

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

M. Daniels and R. Kass. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94:1254–1263, 1999.

M. Drton and M. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, pages 430–449, 2007.

M. Drton and T. Richardson. A new algorithm for maximum likelihood estimation in Gaussian models for marginal independence. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, 2003.

M. Drton and T. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

M. Drton and T. Richardson. Binary models for marginal independence. *Department of Statistics, University of Washington, Tech. report 474*, 2005.

M. Drton and T. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society, Series B*, 70:287–309, 2008a.

M. Drton and T. Richardson. Graphical methods for efficient likelihood inference in Gaussian covariance models. *Journal of Machine Learning Research*, pages 893–914, 2008b.

D. Dunson, J. Palomo, and K. Bollen. Bayesian structural equation modeling. *Statistical and Applied Mathematical Sciences Institute, Technical Report #2005-5*, 2005.

N. Friedman and D. Koller. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning Journal*, 50:95–126, 2003.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems*, 2008.

J. Huang and B. Frey. Cumulative distribution networks and the derivative-sum-product algorithm. *Proceedings of 24th Conference on Uncertainty in Artificial Intelligence*, 2008.

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400, 2005.

M. Jordan. *Learning in Graphical Models*. MIT Press, 1998.

M. Jordan, Z. Ghaharamani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *In M. Jordan (Ed.), Learning in Graphical Models*, pages 105–162, 1998.

C. Kang and J. Tian. Local Markov property for models satisfying the composition axiom. *Proceedings of 21st Conference on Uncertainty in Artificial Intelligence*, 2005.

J. Kotecha and P. Djuric. Gibbs sampling approach for the generation of truncated multivariate Gaussian random variables. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1757–1760, 1999.

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

S.-Y. Lee. *Structural Equation Modeling: a Bayesian Approach*. Wiley, 2007.

D. MacKay. Introduction to Monte Carlo methods. *Learning in Graphical Models*, pages 175–204, 1998.

I. Murray, Z. Ghahramani, and D. MacKay. MCMC for doubly-intractable distributions. *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

R. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.

R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

J. Pearl. *Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93:537–554, 2006.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30:145–157, 2003.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29: 391–411, 2002.

R. Scheines, R. Hoijtink, and A. Boomsma. Bayesian estimation and testing of structural equation models. *Psychometrika*, 64:37–52, 1999.

R. Silva and Z. Ghahramani. Bayesian inference for Gaussian mixed graph models. *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

R. Silva and Z. Ghahramani. Factorial mixtures of Gaussians and the marginal independence model. *Artificial Intelligence & Statistics (AISTATS '09)*, 2009.

R. Silva and R. Scheines. Bayesian learning of measurement and structural models. *23rd International Conference on Machine Learning*, 2006.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

R. Silva, W. Chu, and Z. Ghahramani. Hidden common cause relations in relational learning. *Neural Information Processing Systems (NIPS '07)*, 2007.

J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1:833–860, 2006.

P. Spirtes. Directed cyclic graphical representations of feedback models. *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge University Press, 2000.

R. Tarjan. Decomposition by clique separators. *Discrete Mathematics*, 55:221–232, 1985.

E. Webb and J. Forster. Bayesian model determination for multivariate ordinal and binary data. *Technical report, Southampton Statistical Sciences Research Institute*, 2006.

F. Wong, C. Carter, and R. Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90:809–830, 2003.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, pages 557–585, 1921.

J. Yedidia, W. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, pages 1437–1474, 2008.