Bayesian Inference for Discrete Mixed Graph Models: Normit Networks, Observable Independencies and Infinite Mixtures

Ricardo Silva
Gatsby Computational Neuroscience Unit
University College London
rbas@gatsby.ucl.ac.uk

Zoubin Ghahramani
Department of Engineering
University of Cambridge
zoubin@eng.cam.ac.uk

Abstract

Directed mixed graphs are graphical representations that include directed and bi-directed edges. Such a class is motivated by dependencies that arise when hidden common causes are marginalized out of a distribution. In previous work, we introduced an efficient Monte Carlo algorithm for sampling from Gaussian mixed graph models. An analogous model for discrete distributions is likely to be doubly-intractable, in the sense that even a single Markov Chain Monte Carlo step might have a computational cost that scales exponentially with the number of variables. Instead, we built upon our results on Gaussian distributions to describe algorithms and priors for discrete binary and ordinal modeling. The models we describe are based on link functions, where a multivariate Gaussian distribution encoded by a mixed graph is projected into a discrete space. In order to account for flexible discrete distributions, we embed this model within a Dirichlet process mixture of Gaussians.

1 CONTRIBUTION

Directed mixed graphs (DMGs) are graphs with directed and bi-directed edges. They are motivated by considering marginal dependencies that are obtained out of a directed acyclic graph (DAG) when some variables are marginalized. DMGs generalize the class of conditional independencies that can be represented by a DAG. Any two vertices in a DMG might be connected by more than one edge. An example of such a graph is depicted in Figure 1.

Acyclic DMGs (ADMGs) do not have directed cycles, i.e., no sequence $Y \rightarrow \cdots \rightarrow Y$ with directed edges only. Richardson (2003) describes several properties of acyclic DMGs, including a generalization of d-separation, called m-separation, which is a sound and complete procedure for reading independencies off an ADMG. Richardson and Spirtes (2002) provide a detailed account of a particular type of mixed graphs (ancestral graphs) which can encode the same conditional independencies represented by any ADMG.

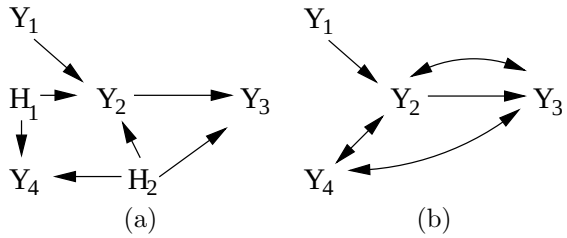

Figure 1: Marginalizing variables H_1 and H_2 out of the DAG in (a) results in the dependency structure represented by the DMG in (b). Notice that other choices of DMGs can also represent the same conditional independencies (Richardson and Spirtes, 2002).

There are standard approaches for parameterizing a Gaussian distribution according to a DMG (Richardson and Spirtes, 2002; Bollen, 1989). Maximum likelihood estimators for Gaussian ancestral graph models can be obtained by an iterative procedure (Drton and Richardson, 2004). Sampling algorithms for Bayesian inference are given by Silva and Ghahramani (2006)

Modeling discrete distributions according to mixed graphs is still an open problem. A parameterization and two maximum likelihood estimation algorithms for multivariate Bernoulli *bi-directed* graph models (i.e., no directed edges allowed) are given by Drton and Richardson (2005). The only constraints imposed by such a parameterization are the marginal independence constraints entailed by the graph: namely, if two vertices are not adjacent, then they are marginally independent. The fact these are the only constraints in

the model implies several desirable statistical properties, such as allowing for consistent asymptotic approximations using the BIC score (Drton and Richardson, 2005; Richardson and Spirtes, 2002).

However, such algorithms are required to solve, for each step in an iterative procedure, a constrained optimization problem within a polytope of possibly exponentially many faces (in the number of variables), even for sparse graphs such as a bi-directed chain. We expect that doing Bayesian inference for such type of discrete models is going to be as hard as Bayesian inference for Markov random fields (Murray et al., 2006), which is doubly-intractable.

Although the original problem (explicitly defined by independence constraints only) of Bayesian modeling of discrete DMG models still deserves treatment, in this paper we will focus on a different parameterization of such models. Observed variables are modeled as discretizations of underlying continuous latent variables. While such a setup implies constraints that may be hard to characterize even asymptotically, we will show that there are practical applications and MCMC algorithms for such a class. Moreover, by allowing the underlying continuous variables to follow a nonparametric distribution, we account for a flexible class of contingency tables. Although the model can be seen as most natural for binary and ordinal data, by using an infinite mixture basis, it should be possible in principle to model any contingency table, as argued by Kottas et al. (2005).

The paper is organized as follows: Section 2 is a description of normit link functions for discrete models and how they can be extended to parameterize DMG models. We also discuss the choice of normit representations and Dirichlet process mixtures of such models. An MCMC algorithm for Bayesian inference is given in Section 3. Section 4 presents experiments.

2 MODELS FOR NORMIT NETWORKS

Following (Bollen, 1989; Richardson and Spirtes, 2002) and many others, a Gaussian DMG is parameterized as follows. Let $\mathcal{G}$ be a DMG with vertices (variables) $\mathbf{Y}$. For each variable Y_j with parents $Y_{(1_j)}, ..., Y_{(k_j)}$ in $\mathcal{G}$, we provide a "structural equation"

$$Y_j = \mu_j + b_{j(1_j)}Y_{(1_j)} + \cdots + b_{j(k_j)}Y_{(k_j)} + \epsilon_j \qquad (1)$$

where ϵ_j is a Gaussian random variable with zero mean. We will make use of the following notation: q designates the number of variables (vertices) in our model; $\mathbf{B}$ is a $q \times q$ matrix corresponding to the linear coefficients defined above, where b_{ji} is different from zero only if Y_i is a parent of Y_j; $\mathbf{V}$ is the $q \times q$ covariance matrix of "error terms" ϵ.

$\mathbf{V}$ will in general be a non-diagonal matrix, where, for any pair $\{Y_i, Y_j\}$, we have that $v_{ij} \neq 0$ only if edge $Y_i \leftrightarrow Y_j$ exists in $\mathcal{G}$. In DMG terminology, if Y_i is connected to Y_j by a bi-directed edge, then Y_i is a *spouse* of Y_j (Richardson, 2003).

In our setup, we allow for up to two edges connecting any pair of vertices (one directed, and one bi-directed), and we assume the graph is acyclic, i.e., an ADMG.

We will first focus on discrete distributions for binary variables based on the *normit* link function (also known as *probit*). Namely, for a binary variable Y_j

$$P(Y_j = 1|Y_{j1}, \ldots, Y_{jk}) = P(Y_j^* > 0) \qquad (2)$$

where $Y_j^* = \mu_j + \mathbf{B_j}^T\mathbf{Y} + \epsilon_j$ is called the *underlying latent variable* for Y_j: our observed variable is a discretization of some hidden continuous variable with a Gaussian distribution. The name normit follows by the fact that $P(Y_j = 1|Y_{j1}, \ldots, Y_{jk})$ is given by the cumulative distribution function of the normal Y_j^*.

An extension of this model for ordinal variables is natural (but less so for multilevel discrete variables). It also indirectly provides a way of modeling joint distributions of continuous and discrete variables. We refer to Bartholomew and Knott (1999) for variations on underlying latent variable models. In Section 4, we explain how to adapt it to ordinal models and perform experiments accordingly.

2.1 Going nonparametric

As pointed out by Kottas et al. (2005), the normit model can be too restrictive. For instance, in a collaborative filtering application where one wants to measure the agreement on assessments by two customers, an underlying Gaussian model cannot represent raters that agree strongly on extreme scores (one-star and five-stars movies, for instance), while simultaneouly being weakly associated on the intermediate scores.

In order to model flexible contingency tables, Kottas et al. (2005) propose using Dirichlet process mixtures (Neal, 2000) of normit models. The setup is as follows. Let G be a random measure that is distributed as a Dirichlet process (DP) with parameters α_0 and G_0,

$$G \sim DP(\alpha_0, G_0) \qquad (3)$$

Here, G_0 is some base probability measure and α_0 is a smoothness factor over G_0 (with the limit $\alpha_0 \to 0$ corresponding to the original normit model).

Let $\{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(n)}\}$ be the random parameters associated with the respective data points

$\{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}\}$. The DP mixture model states

$$\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(n)} | G \overset{iid}{\sim} G(\cdot) \tag{4}$$

In the case studied by Kottas et al. (2005), each $\Theta^{(i)}$ is the mean and covariance matrix of an unconstrained Gaussian[1], representing the distribution of the underlying latents $\mathbf{Y}^*$.

$$\mathbf{Y}^{*(i)} | \Theta^{(i)} \sim N(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)}) \tag{5}$$

Pairs $\{\mathbf{Y}^{*(i)}, \mathbf{Y}^{*(j)}\}$, $i \neq j$, are independent given Θ.

Kottas et al. (2005) describe a Markov chain Monte Carlo algorithm. Latents $\mathbf{Y}^*$ are sampled within a Gibbs sampling scheme, which in this case reduces to sampling from a univariate truncated Gaussian. Other algorithms for Dirichlet process mixtures that could be adapted to this problem are given by Neal (2000).

Introducing independence constraints entailed by DMG models into each $\Sigma^{(i)}$, however, brings additional challenges, since they imply non-standard distributions for the parameters (Silva and Ghahramani, 2006). Moreover, there are two possible approaches one can take when representing such constraints, as discussed next.

2.2 Models of observable independencies

A common approach for modeling with structured normit models is illustrated as follows. Consider the graph in Figure 2(a). A possible normit model for it, with underlying latents $\mathbf{Y}^*$ explicitly represented as vertices, is given graphically in Figure 2(b). Notice that while the model in (a) encodes that Y_1 and Y_3 are independent given Y_2, no conditional independencies among observed variables exist in (b).

In general, for a given graph $\mathcal{G}$, a respective graphical representation of a normit model can be built by first replicating $\mathcal{G}$ as a graph $\mathcal{G}^*$ with underlying latent variables (UVs) in place of each respective original vertex. To each vertex Y^* in $\mathcal{G}^*$, we then add a single child Y. 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), for some applications one might still want to encode the original conditional independencies directly.

This alternative is illustrated in Figure 2(c). Starting from the original graph $\mathcal{G}$ (as in Figure 2(a)), the normit graph model $\mathcal{G}^*$ shown in the figure is built from $\mathcal{G}$ by the following algorithm:

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

We call this the *Type-II UV model*, which has the following property:

Theorem 1 *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 $\mathcal{G}^*$.*

Proof: We first show that there is a one-to-one mapping between paths in $\mathcal{G}$ and paths 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 $Y_s^* \leftrightarrow Y_t^*$ in $\mathcal{G}^*$ and each $Y_s \leftrightarrow Y_t$ in $\mathcal{G}$. All directed edges in $\mathcal{G}^*$ are of two types: $Y_s \rightarrow Y_t^*$, with $s \neq t$, or $Y_s^* \rightarrow Y_s$. Therefore, one can show that any path P^* in $\mathcal{G}^*$ corresponds to an unique path P in $\mathcal{G}$ obtained by relabeling each Y^* as Y, and by collapsing any $Y \rightarrow Y$ edges that might result from this relabeling into a single vertex Y.

A collider in a path is any vertex within a head-to-head collision in the path, i.e., 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 mapping between paths P and P^* is such that, Y_t is a collider in P if and only if Y_t is in P^* and is a collider, or Y_t^* is in P^* and is a collider. Since by construction any Y_t^* will have the same $\mathbf{Y}$-descendants in $\mathcal{G}^*$ as Y_t has in $\mathcal{G}$, and $\mathbf{Z} \subset \mathbf{Y}$, the result follows. □

In this paper, we will focus on algorithms for Type-II models only. 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.

2.3 DP mixtures and DMG independencies

If a single latent measure G is used to generate all parameters, then by integrating out G all conditional independencies disappear. In this sense, in the mixture model formulation, our normit models of "observable independencies" encode independencies only to the extent where we condition on parameters and the random measure G.

[1]The thresholds that are used in ordinal variables with more then two levels are not random in their case, since the extra flexibility given by an infinite mixture of Gaussians allows for fixed thresholds.

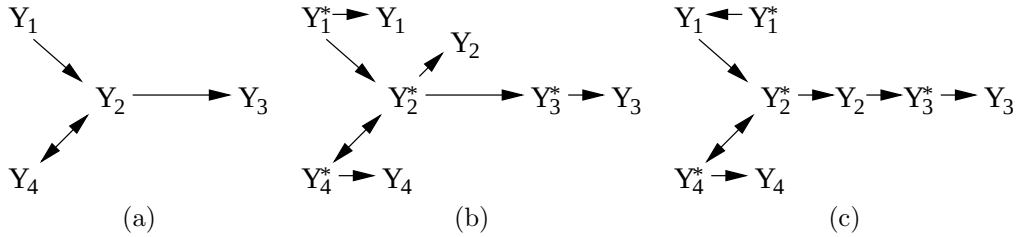

Figure 2: The model in (a) has at least two main representations as a normit network. In (b), the original structure is given to the underlying variables, with observed variables being children of their respective latents. In (c), the underlying variable inherits the parents of the original variable and the underlying latents of the spouses.

Within large families of priors, losing independencies when integrating out parameters is unavoidable, whether we are dealing with mixtures or single Gaussian models. To see this, let a *district* of a DMG $\mathcal{G}$ be a maximal subset of vertices such that any two elements in this set is connected by a path of bi-directed edges. A vertex not connected to any bi-directed edge forms a district of size 1. Notice that districts form a partition of the set of vertices.

Recall that in a Gaussian DMG model, b_{jt} is the parameter corresponding to edge $Y_t \to Y_j$.

Proposition 1 *Suppose $\mathcal{G}$ is acyclic with vertex set $\mathbf{Y}$. In Let $\mathcal{G}'$ be the DMG obtained by augmenting $\mathcal{G}$ with a vertex for each parameter b_{jt} and a respective edge $b_{jt} \to Y_j$. Then if Y_j and Y_k belong to the same district, $\{b_{jt}, b_{kv}\}$ are not m-separated given $\mathbf{Y}$ in $\mathcal{G}'$.*

Proof: The sequence of bi-directed edges between Y_j and Y_k implies a path between b_{jt} and b_{kv} where every vertex but the endpoints is a collider in this path. Since every collider is in $\mathbf{Y}$, this path is active. □

The implication is that, even if the prior for $\mathbf{B}$ is fully factorized, in general b_{jt} and b_{kv} will not be independent in the posterior[2]. Therefore, the predictive distribution for any two variables with parents and in the same district cannot be factorized in the same way a model with fixed parameters can. This is illustrated in Figure 3. A way of avoiding the rise of such dependencies, if desired, is to exclude them a priori, i.e., by incorporating a hyperprior that generates factorized distributions. Concerning mixture models, independencies between vertices in different districts can be directly preserved by using a different random measure G for each district (a type of "factorial Dirichlet process"). A treatment of such priors is out of scope of this paper and suggested as future work. The DMG graphs here discussed encode independencies given parameters and mixture components.

[2]This assumes the posterior distribution will be *faithful* to this graph (Spirtes et al., 2000).

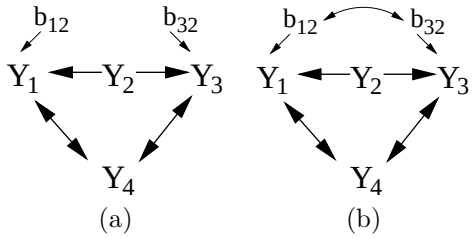

Figure 3: After conditioning on some datapoint, the prior Gaussian model in (a) assumes the posterior dependency structure as in (b), since Y_1, Y_3 and Y_4 are in the same district.

3 DIRICHLET PROCESS MIXTURES OF NORMIT ADMGS

Before introducing DP mixtures of Type-II models for binary variables, we will formally describe the mixture of Gaussians model. The normit case is an adaptation of it with the same family of priors, as explained in Section 3.2. The description of the Gaussian case is also useful to extend the results of Silva and Ghahramani (2006) to the continuous nonparametric case.

3.1 Priors and the Gaussian mixture

Starting from the parametric formulation of Gaussian models in (1) as a set of linear equations

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{Y} + \epsilon \tag{6}$$

we have that, for a particular data point $\mathbf{Y}^{(d)}$ and parameter set $\Theta^{(d)} = \{\boldsymbol{\mu}^{(d)}, \mathbf{B}^{(d)}, \mathbf{V}^{(d)}\}$, the conditional distribution $p(\mathbf{Y}^{(d)}|\Theta^{(d)})$ is Gaussian with mean.

$$\mathbf{m}(\Theta^{(d)}) = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu} \tag{7}$$

and covariance matrix

$$\Sigma(\Theta^{(d)}) = (\mathbf{I} - \mathbf{B}^{(d)})^{-1}\mathbf{V}^{(d)}(\mathbf{I} - \mathbf{B}^{(d)})^{-T} \tag{8}$$

where $\mathbf{M}^{-T}$ is the transpose of the inverse of $\mathbf{M}$. This follows from Equation (6). The shape of this conditional suggests priors based on normal and inverse Wishart distributions.

In the original parametric formulation of Silva and Ghahramani (2006) for multivariate Gaussian data, each non-zero b_{ji} in $\mathbf{B}$ is independent a priori of all other parameters, with prior distribution given by

$$b_{ji} \sim N(\kappa_{ji}, \sigma_{ji}) \tag{9}$$

A prior for μ_j is chosen in an analogous way:

$$\mu_j \sim N(\lambda_j, \upsilon_j) \tag{10}$$

The error covariance matrix $\mathbf{V}$ is given a non-standard prior based on the inverse Wishart with parameters $(\delta, \mathbf{U})$, δ being the "degrees of freedom" and $\mathbf{U}$ any positive definite matrix. This modified inverse Wishart has a different support, which is restricted to a particular cone of positive definite matrices $M^+(\mathcal{G})$. With respect to a fixed DMG $\mathcal{G}$, this space is composed of matrices with a constant value of zero in entries corresponding to pairs of vertices that are not connected by a bi-directed edge. This distribution is called *G-Inverse Wishart*, as described by Silva and Ghahramani (2006). We denote it as follows

$$\mathbf{V} \sim \textit{G-IW}_{\mathcal{G}}(\delta, \mathbf{U}) \tag{11}$$

Let G be a random measure given by a Dirichlet process $G \sim DP(\alpha_0, G_0)$ for the parameter set $\Theta = \{\boldsymbol{\mu}, \mathbf{B}, \mathbf{V}\}$, where the base measure G_0 is given by the product of priors (9), (10) and (11). Following Neal (2000), we will represent a Dirichlet process mixture by a countably infinite mixture of parameters $\{\phi_1, \phi_2, \phi_3, \dots\}$ where

$$\phi_k = \{\boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{V}_k\} \tag{12}$$

and each data point $\mathbf{Y}^{(d)}$ follows a normal with mean $\mathbf{m}(\phi_{c^{(d)}})$ and covariance $\Sigma(\phi_{c^{(d)}})$, $\{\mathbf{m}, \Sigma\}$ being analogous to (7, 8). That is, there is a latent indicator $c^{(d)}$ for each data point d such

$$\mathbf{Y}^{(d)}|c^{(d)}, \phi \sim N(\mathbf{m}(\phi_{c^{(d)}}), \Sigma(\phi_{c^{(d)}})) \tag{13}$$

and for each data point d one has $\Theta^{(d)} = \phi_{c^{(d)}}$. Although the pool of mixture indicators is infinite, a typical posterior distribution will have many data points sharing the same value for c, meaning that typically the observed data is generated by a number of parameters much smaller than the number of data points.

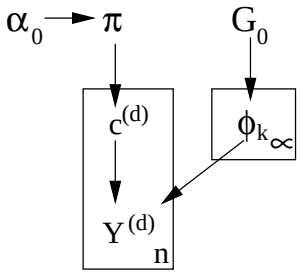

Figure 4: A DP mixture of parameters coming from a infinite pool ϕ_k with probabilities given by π.

A sampling representation of a Dirichlet process mixture is illustrated in Figure 4 as a plate model: one puts a Dirichlet prior with parameter α_0 for the values of $\mathbf{c}$. This prior is over the countable infinite space of indices (say, $1, 2, 3, \dots$) corresponding to the mixture indicators for the data points. As explained by Neal (2000), there are several advantages on working under this DP representation when designing MCMC algorithms for Bayesian inference.

3.2 Binary models and an MCMC algorithm

We now describe the DP mixture of normit models and the corresponding Gibbs sampling algorithm.

By adapting Equation (1), the equations that define the Type II model are as follows:

$$\begin{aligned} \mathbf{Y}^* &= \boldsymbol{\mu} + \mathbf{B}\mathbf{Y} + \epsilon \\ \mathbf{Y} &= 1(\mathbf{Y}^* > 0) \end{aligned}$$

where $1(\cdot)$ is the indicator function (applied elementwise, in case of a vector), equal to 1 if its argument is true, and 0 otherwise.

In this algorithm, we will be conditioning on the latent indicator variables $\mathbf{c} = \{c^{(1)}, c^{(2)}, \dots, c^{(n)}\}$. Also, let $\mathbf{Y}^N = \{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(n)}\}$ denote our dataset, and let $\mathbf{Y}^{*N} = \{\mathbf{Y}^{*(1)}, \dots, \mathbf{Y}^{*(n)}\}$ denote the hidden data for the corresponding UVs.

Within our DP process, $\mathbf{Y}^{*(d)}$ given $\{c^{(d)}, \phi_{c^{(d)}}, \mathbf{Y}^{(d)}\}$ is a truncated Gaussian, with parameters $(\boldsymbol{\mu}_{c^{(d)}} + \mathbf{B}_{c^{(d)}}\mathbf{Y}, \mathbf{V}_{c^{(d)}})$ and support given according to the values of $\mathbf{Y}$ ($Y_j^* > 0$ if and only if $Y_j = 1$). Sampling from this distribution using Gibbs sampling is a simple procedure (Kottas et al., 2005).

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

Proposition 2 *Let $\mathcal{G}$ be an acyclic DMG, and $(\boldsymbol{\mu}, \mathbf{B}, \mathbf{V})$ be the respective set of parameters that define a normit model. For a fixed $(\boldsymbol{\mu}, \mathbf{B})$, there is a bijective function $f_{\mathbf{B}\boldsymbol{\mu}}(\cdot)$ mapping $\mathbf{Y}^*$ 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^* vertices with no parents.

For cyclic graphs, the following model provides a counter-example. Let the graph be $Y_1^* \to Y_1 \to Y_2^* \to Y_2 \to Y_1^*$. Let the model be $Y_1^* = Y_2 + \epsilon_1, Y_2^* = Y_1 + \epsilon_2$, i.e., $b_{12} = b_{21} = 1$ and $\boldsymbol{\mu} = 0$. Then the two instantations $(Y_1^* = -0.8, Y_2^* = 0)$, $(Y_1^* = 0.2, Y_2^* = 1)$ imply the same pair $(\epsilon_1 = -0.8, \epsilon_2 = 0)$. □

Due to this bijection (and the determinism linking $\mathbf{Y}$ to $\mathbf{Y}^*$), the density $p(\mathbf{V}_k \mid \boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{Y}, \mathbf{Y}^*)$ for some $\{\mathbf{Y}, \mathbf{Y}^*\}$ coming from cluster k, is equivalent to

$$\begin{aligned} & p(\mathbf{V}_k \mid \boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{Y}^* = \mathbf{y}^*) \\ = \; & p(\mathbf{V}_k \mid \boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{Y}^* = \mathbf{y}^*, \epsilon = f_{\mathbf{B}\boldsymbol{\mu}}(\mathbf{y}^*)) \\ = \; & p(\mathbf{V}_k \mid \boldsymbol{\mu}_k, \mathbf{B}_k, \epsilon = f_{\mathbf{B}\boldsymbol{\mu}}(\mathbf{y}^*)) \\ \propto \; & p(\epsilon = f_{\mathbf{B}\boldsymbol{\mu}}(\mathbf{y}^*) \mid \boldsymbol{\mu}_k, \mathbf{B}_k, \mathbf{V}_k) p(\mathbf{V}_k \mid \boldsymbol{\mu}_k, \mathbf{B}_k) \\ = \; & p(\epsilon = f_{\mathbf{B}_k\boldsymbol{\mu}_k}(\mathbf{y}^*) \mid \mathbf{V}_k) p(\mathbf{V}_k) \end{aligned}$$

For the given dataset $\mathbf{Y}^N$, define $\mathbf{S}_k^*$ as the sum of $(\mathbf{Y}^{*(d)} - \boldsymbol{\mu}_k - \mathbf{B}_k\mathbf{Y}^{(d)})(\mathbf{Y}^{*(d)} - \boldsymbol{\mu}_k - \mathbf{B}_k\mathbf{Y}^{(d)})^T$ over all $d \in \{1, 2, \ldots, n\}$ such that $c^{(d)} = k$. Since $p(\mathbf{V}|\epsilon) \propto p(\epsilon|\mathbf{V})p(\mathbf{V})$, where $p(\epsilon \mid \mathbf{V})$ is normal with zero mean and covariance matrix $\mathbf{V}$, the posterior for $\mathbf{V}_k$ given all other parameters and variables is

$$\mathbf{V}_k \mid \epsilon^N \sim \textit{G-IW}_{\mathcal{G}}(\delta + n_k, \mathbf{U} + \mathbf{S}_k^*)$$

where n_k is the number of data points d in $\mathbf{Y}^N$ such that $c^{(d)} = k$. Silva and Ghahramani (2006) describe an algorithm for sampling from a *G-IW* distribution.

Let $(v)_{ijk}^{-1}$ be then ijth entry of $\mathbf{V}_k^{-1}$. The posterior for $b_{jik} \in \phi_k$ given the data and all other parameters, for $k \in \mathbf{c}$, is a $N(s_{jik}^{-1} m_{jik}, s_{jik})$ where

$$s_{jik} = \sigma_{ji}^{-1} + (v)_{jjk}^{-1} \left\{ \sum_d (Y_j^{(d)})^2 \mid c^{(d)} = k \right\}$$

$$m_{jik} = \frac{\kappa_{ji}}{\sigma_{ji}} + \left\{ \sum_d Y_j^{(d)} \times T_{jd}^k \mid c^{(d)} = k \right\}$$

$$T_{jd}^k = \sum_t (v)_{jtk}^{-1} \left(Y_t^{*(d)} - \sum_{p \mid tp \neq ji} b_{tpk} Y_p^{(d)} - \mu_{tk} \right)$$

where index t can be seen as running over the other vertices in the same district as Y_j (since v_{jtk}^{-1} is zero otherwise) and index p can be seen as running over the $\mathbf{Y}$-parents of Y_t^* only (since b_{tpk} is zero otherwise).

The posterior for $\mu_{jk} \in \phi_k$ is given by a normal $N((s'_{jk})^{-1} m'_{jk}, s'_{jk})$ where

$$\begin{aligned} s'_{jk} &= \upsilon_j^{-1} + n_k \times (v)_{jjk}^{-1} \\ m'_{jk} &= \frac{\lambda_j}{\upsilon_j} - n_k \sum_t (v)_{jt}^{-1} \mu_{tk} \\ &\quad + \sum_d \left\{ \sum_t (v)_{jt}^{-1} \left(Y_t^{(*d)} - \sum_p b_{tpk} Y_p^{(d)} \right) \mid c^{(d)} = k \right\} \end{aligned}$$

For a given data point d, the likelihood of $\{\mathbf{Y}^{(d)}, \mathbf{Y}^{*(d)}\}$ given $\phi_{c^{(d)}}$, which we denote by $F(\mathbf{Y}^{(d)}, \mathbf{Y}^{*(d)}, \phi_{c^{(d)}})$ is given by

- 0, if $Y_j^{(d)} \neq 1(Y_j^{*(d)})$ for some j (this event does not happen, since $\mathbf{Y}^*$ is obtained by simulation);
- $p_N(\epsilon^{(d)}; \mathbf{V})$, the density function of a normal $N(0, \mathbf{V})$ evaluated at $\epsilon^{(d)} = (\mathbf{Y}^{*(d)} - \boldsymbol{\mu}_k - \mathbf{B}_k\mathbf{Y}^{(d)})(\mathbf{Y}^{*(d)} -_k - \mathbf{B}_k\mathbf{Y}^{(d)})^T$;

This likelihood expression can be plugged into an algorithm for sampling $\mathbf{c}$. In our implementation, we use Algorithm 8 of (Neal, 2000).

4 EXPERIMENTS

We evaluate our algorithm on the task of computing the predictive log-likelihood of particular models. We analyze one synthetic and two real-world datasets containing both ordinal and binary variables.

For all experiments we set the priors empirically, for the purposes of illustration, by assuming the data follows a multivariate normal distribution. The resulting maximum likelihood estimates of regression coefficients and error covariances are used as the mean parameters. The variance parameters for edge coefficients b_{ji} are all set to 1. We set $\delta = 1$ in our G-Inverse Wishart prior. We set $\alpha_0 = 1$ as the DP smoothing parameter. The mapping between an UV Y^* and the respective observed ordinal variable Y with domain $\{y_1, y_2, \ldots, y_r\}$ is set according to a pre-defined set of thresholds $\tau = \{\tau_0, \tau_1, \ldots, \tau_{r+1}\}$ such that

$$Y = y_j \text{ if and only if } \tau_{j-1} < Y^* \leq \tau_j$$

For all variables, we set $\tau_0 = -\infty, \tau_{r+1} = +\infty$. The other thresholds are set to correspond to the ordinal values, i.e., $\tau_j = (y_j + y_{j+1})/2$. We encode ordinal values such that the least and largest values are symmetric around zero, with a gap $y_{j+1} - y_j = 1$.

4.1 SYNTHETIC STUDY

We generated a sample of 2,000 datapoints from the graphical model depicted in Figure 5(a). This model is parameterized such that each variable has 10 different values. To simplify the data generation, we treat each of the exogenous variables Y_1, Y_2, Y_3 as multinomials with random marginal probabilities, and sample accordingly. The conditional distribution of $\{Y_4, Y_5\}$ given their respective parents is defined as follows: underlying latent variables are sampled from a mixture of three equiprobable bivariate Gaussians. The mean of each Gaussian N_i is such that $E[Y_4^*]_i = b_{41}^i Y_1 + b_{42}^i Y_2 + b_{40}^i$ and $E[Y_5^*]_i = b_{53}^i Y_3 + b_{50}^i$. Coefficients b_{jk}^i are uniformly generated in the interval $[-2, -1] \cup [1, 2]$. The entries of the covariance of $\{Y_4^*, Y_5^*\}$ are generated by uniformly sampling from $[1, 3]$ until a positive definite matrix is generated. Notice that this sampling scheme does not correspond perfectly to a Type-II model because of the way $\{Y_1, Y_2, Y_3\}$ are generated.

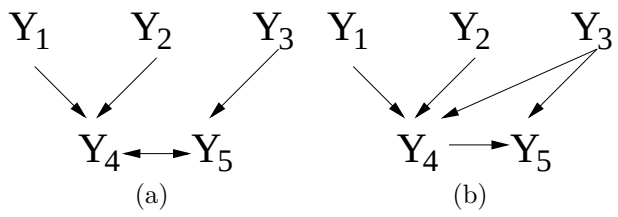

Figure 5: In (a), the DMG we used to generate our data. In (b), a corresponding minimal DAG (i.e., no DAG with fewer edges will respect the dependencies encoded in (a)).

We ran three experiments by using 1,000 points as a test set, and three training sets of 200, 400 and 1,000 data points. We compare the predictive log-likelihood of this model against a multinomial Bayesian network with the prior used in the BDeu score (Heckerman et al., 1995). This prior depends on a single parameter ω that adds a number of pseudo-counts to the data according to a uniform distribution.

One can verify that the directed acyclic graph shown in Figure 5(b) is minimal, in the sense that no edge can be removed without introducing an independence that does not exist in the true model of Figure (a). This model can badly overfit, since the conditional probability table of Y_4 given its parents has 1,000 entries for each value of Y_4. Results for the DAG multinomial model are shown below in Table 1 for a variety of prior parameters ω.

In Table 2, we show results for our model with the fixed prior and three different runs of 5,000 iterations (throwing out the first 500 samples of the chain).

# Train	$\omega = 10$	50	100	500
200	-10.47	-10.39	-10.44	-10.84
400	-10.56	-10.39	-10.36	-10.57
1000	-10.54	-10.31	-10.22	-10.18

Table 1: Predictive log-likelihood for multinomial DAG model with BDeu prior and different prior parameters ω. The higher ω is, the closer to an uniform distribution the posterior gets.

# Train	Trial 1	Trial 2	Trial 3
200	-9.90	-9.89	-9.66
400	-9.80	-9.44	-9.80
1000	-9.35	-9.28	-9.37

Table 2: Predictive log-likelihood for ordinal ADMG model across different starting points.

It is clear that in this case the nonparametric ADMG models the distribution much better than a multinomial DAG model. Notice how a higher amount of smoothing improves results for the DAG. However, as it is typical of mixture models, there is some tendency of Gibbs sampling to get stuck around some mode of the posterior distribution. In the future we want to explore techniques on how to minimize the impact of local maxima, such as adapting the split-merge method of Jain and Neal (2004).

4.2 OTHER EMPIRICAL STUDIES

We applied our approach to model two multivariate ordinal datasets from the UCI repository. We ran a standard hill-climbing procedure to obtain a DAG structure using the BDeu score. This structure was used to evaluate DAG multinomial models. We then heuristically expand such a DAG into an ADMG by i. fitting a Gaussian model by maximum likelihood according to the DAG; ii. calculating the residual covariance between any two vertices that were not adjacent in the DAG; iii. sorting all of such pairs by the absolute value of their residual covariances; iv. adding bi-directed edges $X \leftrightarrow Y$ for all pairs on the top 50% of such a list. This is a very simple structure learning algorithm for ADMGs used strictly for the purposes of illustrating our MCMC algorithm for parameter learning. A more thorough evaluation of the advantages of such a representation is planned for the future, where fast approximations for marginal likelihoods will also have to be developed to be used in more sophisticated ADMG structure learning algorithms.

The first dataset we applied our algorithm to was the Contraceptive Method Choice (CMC). The CMC contains 1473 instances and 10 attributes. We removed

two attributes (age and number of children) which were not bounded ordinal variables. The corresponding DAG and ADMGs structures for this problem are shown in Figure 6.

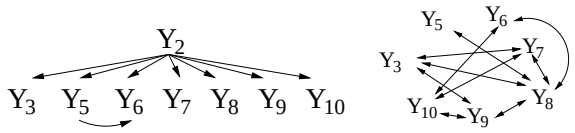

Figure 6: The DAG structure used in the CMC domain. Variable numbers correspond to those in the dataset documentation. The figure on the right depicts the extra bi-directed edges added to the DAG in order to generate our ADMG model.

For the DAG model, we set the value of ω to 10. Other values give basically the same result. For the ADMG model, we used an alpha parameter of 0.1, which gave more stable results across different runs. Using a 10-fold cross-validation setup, we got that the average difference in predictive log-likelihood between our ADMG and the multinomial DAG model was 0.16 with a standard deviation of 0.05. The ADMG model is better than the DMG, the difference being significant at the 0.05 level using a z-test. However, the small difference is an indicative that the sparse DAG model still modeled the density well enough.

We also did an experiment with the breast cancer data of the UCI repository and the same experimental setup. In this case, the difference was of -0.07 with a standard deviation of 0.11. Although the difference is not significant, we have an indication that the DP mixture model at least does not seem to add any significant bias compared to the unconstrained multinomial model. In practice, more flexible parameterizations of the joint (such as the one by Drton and Richardson (2005) for bi-directed models) might be unnecessary.

5 CONCLUSIONS

Our approach allows flexible modeling of binary and ordinal data, arguably the first direct Bayesian treatment for discrete DMG models. Combined with the methods by Albert and Chib (1993), it can in principle provide a way of modeling non-ordinal multilevel discrete data. It can also be modified to allow for data with both discrete and continuous variables.

It would be interesting to investigate how our framework could provide a practical alternative to the Bayesian Markov random fields formulation of Murray et al. (2006), by parameterizing such models through infinite mixtures of normit models (for instance, "Type-I" parametric normit log-linear models are discussed by Webb and Forster (2006)).

Finally, it is still an open problem how to perform Bayesian inference with discrete DMG models that encode conditional independency constraints only, in the spirit of the parameterization and maximum likelihood algorithms of Drton and Richardson (2005).

References

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

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

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

M. Drton and T. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. *UAI*, 2004.

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

D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

A. Kottas, P. Muller, and F. Quintana. Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14:610–625, 2005.

I. Murray, Z. Ghahramani, and D. MacKay. MCMC for doubly-intractable distributions. *UAI*, 2006.

R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

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.

R. Silva and Z. Ghahramani. Bayesian inference for Gaussian mixed graph models. *UAI*, 2006.

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

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