Fronk, Giudici:

Markov Chain Monte Carlo Model Selection for DAG Models

Sonderforschungsbereich 386, Paper 221 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Markov Chain Monte Carlo Model Selection for DAG Models

Eva–Maria Fronk

Department of Statistics
Ludwig–Maximilians–University
Munich
Germany
fronk@stat.uni-muenchen.de

Paolo Giudici

Department of Economics and
Quantitative Methods
University of Pavia
Italy
giudici@unipv.it

Abstract

We present two methodologies for Bayesian model choice and averaging in Gaussian directed acyclic graphs (dags). In both cases model determination is carried out by implementing a reversible jump Markov Chain Monte Carlo sampler. The dimension–changing move involves adding or dropping a (directed) edge from the graph. The first methodology extends the results in Giudici and Green (1999), by excluding all non–moralized dags and searching in the space of their essential graphs. The second methodology employs the results in Geiger and Heckerman (1999) and searches directly in the space of all dags. To achieve this aim we rely on the concept of adjacency matrices, which provides a relatively inexpensive check for acyclicity. The performance of our procedure is illustrated by means of two simulated datasets.

Keywords: Adjacency matrix; Bayesian model selection; Gaussian dag models; inverse Wishart distribution; reversible jump Markov Chain Monte Carlo.

1 Introduction

Model selection by reversible jump (rj) MCMC has been recently developed by Giudici and Green (1999) for pure continous variables and by Giudici, Green, and Tarantola (1999) for the pure discrete case. Both approaches consider only undirected decomposable models (udg) which allow a factorization by cliques and separators and thereby also local computations.
Factorization is possible also dealing with Gaussian directed acyclic graphs (dags) and using a Normal–Wishart distribution as a prior, as shown for instance in Geiger

and Heckerman (1999). They propose a method for the construction of the prior distribution in dag models which allows a simple derivation of the marginal likelihood for every model. A problem which arises in the directed case is the possible Markov equivalence of different dags. Andersson, Madigan, and Perlman (1997a) have shown that any class of equivalent dags can be represented by a single chain graph, the so–called essential graph. Another result of the authors (Andersson, Madigan, and Perlman, 1997b) says that the undirected decomposable graphs are equivalent to the essential graphs of all moralized dags, which means all dags without immoralities.
We shall present two reversible jump algorithms for model selection for directed acyclic graphs. The first one considers the equivalence classes by excluding all non–moralized dags and searching in the space of their essential graphs. This algorithm corresponds essentially to the above mentioned algorithm of Giudici and Green. It has been extended to allow for a mean parameter different from zero. Our long-term objective is to develop an algorithm that moves in the space of the essential graphs of *all* dags. This would decrease the huge search space enormously. But this calls for further discussions and careful graph–theoretical considerations and it therefore requires further research.
The second algorithm, which makes use of the results in Geiger and Heckerman (1999), searches directly in the space of all dags, without accounting for the equivalence classes. The representation of a graph in this algorithm relies on the concept of adjacency matrices which is well known in graph theory. This representation also provides a relatively inexpensive check for acyclicity. The algorithm is incorporated into the software package *BayesX*, which is available for public use under http://www.stat.uni-muenchen.de/∼lang/ (see also Lang and Brezger [**?**]).
We compare the results obtained from application of these two methods by simulated datasets. In order to have another criterion for the comparison of the algorithms we also perform exact calculations for a trivial simulated example with three variables and compare them with the corresponding results obtained from the simulations.

2 Gaussian UDG Models

In this section we extend the (undirected) Bayesian graphical Gaussian model proposed by Giudici and Green (1999) by allowing for the presence of a mean parameter $\boldsymbol{\mu}$. Let $\boldsymbol{X} = (X_1, \dots X_k)'$ be a vector of $p \geq 3$ random variables, such that

$$(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} > 0$ a positive definite matrix. We assume that for a given undirected graph g:

$$\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, g \;\; \sim \;\; \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\begin{aligned} \boldsymbol{\mu} \mid \boldsymbol{\Sigma}, g &\sim \mathrm{N}_p(\boldsymbol{\nu}, \frac{1}{\alpha_\mu}\boldsymbol{\Sigma}), \\ \boldsymbol{\Sigma} \mid g &\sim HIW(\alpha, \boldsymbol{\Phi}), \end{aligned}$$

where $HIW(\cdot)$ indicates a hyper inverse Wishart distribution. We remark that the above is, in the terminology of Giudici and Green (1999), a non–hierarchical model; one may want to take $(\alpha, \boldsymbol{\Phi}) \sim \pi_1(\cdot)$ and $(\boldsymbol{\nu}, \alpha_\mu) \sim \pi_2(\cdot)$.
Finally, supposing that there exist G possible decomposable undirected models, which, in the absence of subject–matter information, have all the same probability, we get a discrete uniform distribution for g:

$$p(g) = 1/G.$$

Given a complete sample $\boldsymbol{X}$, the joint distribution of all random quantities results in

$$\begin{aligned} p(g, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X}) &= p(g)\, p(\boldsymbol{\Sigma}|g)\, p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, g)\, p(\boldsymbol{X}|\boldsymbol{\Sigma}, \boldsymbol{\mu}, g) \\ &\propto \frac{\prod_{C\in\mathcal{C}} \det(\boldsymbol{\Sigma}^C)^{-(\alpha+2|C|)/2} \exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^C(\boldsymbol{\Sigma}^C)^{-1})\}}{\prod_{S\in\mathcal{S}} \det(\boldsymbol{\Sigma}^S)^{-(\alpha+2|S|)/2} \exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^S(\boldsymbol{\Sigma}^S)^{-1})\}} \\ &\quad \times \det(\boldsymbol{\Sigma})^{-1/2} \exp\{-\frac{\alpha_\mu}{2}(\boldsymbol{\mu}-\boldsymbol{\nu})'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu})\} \\ &\quad \times \frac{\prod_{C\in\mathcal{C}} \det(\boldsymbol{\Sigma}^C)^{-n/2} \exp\{-\frac{1}{2}\mathrm{tr}(S_C(\boldsymbol{\Sigma}^C)^{-1})\}}{\prod_{S\in\mathcal{S}} \det(\boldsymbol{\Sigma}^S)^{-n/2} \exp\{-\frac{1}{2}\mathrm{tr}(S_S(\boldsymbol{\Sigma}^S)^{-1})\}} \\ &\quad \times \frac{1}{G}, \end{aligned}$$

where C and S denote a clique respectively a separator, while $\mathcal{C}$ and $\mathcal{S}$ represent the sets of cliques and separators. Furthermore $\boldsymbol{\Sigma}^C$ is the part of the covariance matrix corresponding to the clique C, analogously $\boldsymbol{\Sigma}^S$ the one corresponding to the separator S.

3 Gaussian DAG Models

A Gaussian dag model d can be represented as a regression model for each variable X_i, $i = 1, \ldots, p$, given the parents of X_i, denoted by $X_{pa(i)}$,

$$X_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}, d \sim \mathrm{N}(\beta_{i0} + \sum_{x_l \in pa(x_i)} \beta_{il} x_l, \sigma^2_{i|pa(i)})$$

and the joint distribution of all variables $\mathbf{X} = (X_1, \ldots, X_p)'$ is then given by

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{p} p(x_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}),$$

where $\boldsymbol{\beta}_{i|pa(i)}$ is the $|pa(i)|+1$–dimensional vector of the intercept β_{i0} and the $|pa(i)|$ regression coefficients of X_i. Furthermore, $\sigma^2_{i|pa(i)}$ is the partial variance of X_i given

its parents $\boldsymbol{x}_{pa(i)}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{1|pa(1)}, \ldots, \boldsymbol{\beta}'_{p|pa(p)})'$ denote the vector of the $\boldsymbol{\beta}_{i|pa(i)}$'s and accordingly $\boldsymbol{\sigma}^2 = (\sigma^2_{1|pa(1)}, \ldots, \sigma^2_{p|pa(p)})'$ the vector of the conditional variances $\sigma^2_{i|pa(i)}$.
As the dag model is Gaussian, when d is complete the joint distribution of $\boldsymbol{X}$ is a p–dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Therefore one could work equivalently with the $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ parametrization. See for instance Geiger and Heckerman (1994) or Schachter and Kenley (1989). We prefer to work with the former; however, when comparing results with the undirected ones we shall indeed make use of this reparametrisation. We remark the well–known connection $\sigma^2_{i|pa(i)} = \sigma^2_{ii} - \boldsymbol{\Sigma}_{i,pa(i)}\boldsymbol{\Sigma}^{-1}_{pa(i)}\boldsymbol{\Sigma}_{pa(i),i}$ between the partitioned covariance matrices of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{i,pa(i)}$, and the partial variances of the i–th regression model. Let in addition n denote the number of observations.

The vector $\boldsymbol{\beta}_{i|pa(i)}$ is assumed to be normally distributed with mean $\boldsymbol{b}_{i|pa(i)}$ and covariance matrix $\frac{1}{\alpha_i}\sigma^2_{i|pa(i)}\boldsymbol{I}$, where α_i is a known scaling factor. For the sake of simplicity, we shall assume $\alpha_i = \alpha$, which gives

$$\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)}, d \sim \mathrm{N}_{|pa(i)|+1}\left(\boldsymbol{b}_{i|pa(i)},\ \frac{1}{\alpha}\sigma^2_{i|pa(i)}\boldsymbol{I}\right).$$

This implies that the coefficients of a regression model are assumed to be mutually independent. For the partial variance $\sigma^2_{i|pa(i)}$ we use an inverse gamma distribution with parameters $\delta_{i|pa(i)}$ and $\lambda_{i|pa(i)}$:

$$\sigma^2_{i|pa(i)} \mid d \sim \mathrm{IG}\left(\delta_{i|pa(i)},\ \lambda_{i|pa(i)}\right).$$

Finally, supposing that there exist D possible dags, which, in the absence of subject–matter information, have all the same probability, we get a discrete uniform distribution for d:

$$p(d) = 1/D.$$

Taking advantage of the well–known factorization property of the joint distribution in (1) and the "global parameter independence" in (2) and (3) (for a detailed description see Geiger and Heckerman, 1999) i.e.:

$$p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) = \prod_{i=1}^{p} p(x_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}), \tag{1}$$

$$p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2, d) = \prod_{i=1}^{p} p(\boldsymbol{\beta}_{i|pa(i)} \mid \boldsymbol{\sigma}^2_{i|pa(i)}), \tag{2}$$

$$p(\boldsymbol{\sigma}^2 \mid d) = \prod_{i=1}^{p} p(\sigma^2_{i|pa(i)}), \tag{3}$$

we get for the joint distribution:

$$
\begin{aligned}
& p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) \\
&= p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\sigma}^2 \mid d)\, p(d) \\
&= \prod_{i=1}^{p} p(\boldsymbol{x}_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}) \prod_{i=1}^{p} p(\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)}) \\
&\quad \prod_{i=1}^{p} p(\sigma^2_{i|pa(i)})\, p(d) \\
&= \prod_{i=1}^{p} (2\pi\sigma^2_{i|pa(i)})^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2_{i|pa(i)}} \sum_{l=1}^{n} (x_{li} - \beta_{i0} - \boldsymbol{\beta}_{i|pa(i)} \boldsymbol{x}_{pa(i)})^2 \right\} \\
&\quad \times \prod_{i=1}^{p} (2\pi \frac{1}{\alpha} \sigma^2_{i|pa(i)})^{-\frac{1}{2}} \exp\left\{ -\frac{\alpha}{2\sigma^2_{i|pa(i)}} (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)})' (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)}) \right\} \\
&\quad \times \prod_{i=1}^{p} \frac{\lambda_{i|pa(i)}^{\delta_{i|pa(i)}}}{\Gamma(\delta_{i|pa(i)})} (\sigma^2_{i|pa(i)})^{\delta_{i|pa(i)}-1} \exp(-\lambda_{i|pa(i)}/\sigma^2_{i|pa(i)}) \\
&\quad \times \frac{1}{D}. \qquad (4)
\end{aligned}
$$

4 On the Representation of DAGs

This section deals with the problem of representing a directed acyclic graph and on how to test for acyclicity exploiting the concept of adjacency matrices.

Definition: Let $\mathcal{G} = (V, E)$ be a graph with $|V| = p$. The adjacency matrix of $\mathcal{G}$ is defined as the $(p \times p)$–matrix A, $[A]_{ij} = a_{ij}$, with

$$
a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E. \end{cases}
$$

All three types of graphs (undirected, directed and chain graph) can be uniquely represented by the corresponding adjacency matrix. The following corollary allows to develop an algorithm to test for acyclicity in a directed graph.

Corollary: The (i, j)–th entry $a_{ij}^{(l)}$ of the l–th power of A, $A^l = A^{l-1}A$, is equal to the number of directed paths of length l from i to j and $a_{ij}^{(l)} = \sum_{k=1}^{p} a_{ik}^{(l-1)} a_{kj}$.

Since a cycle is defined as a path from a vertex i to itself, an entry different from zero of the i–th diagonal element of A^l corresponds to a cycle of length l containing the vertex i. As a cycle in $\mathcal{G} = (V, E)$ has maximal length $|V| = p$, all diagonal elements $a_{ii}^{(l)}$ have to be zero for $l = 3, \ldots, \min(p, |E|)$ and $i = 1, \ldots, p$ to ensure acyclicity.

Making use of the fact that for a cycle of length l there are at least l diagonal elements of A^l different from zero, only the first $p-l+1$ diagonal elements have to be checked. Our proposed algorithm uses this fact as the numbers of executions of the inner loop decreases with every execution of the outer one. The algorithm starts with the variable $no\_cycle = TRUE$ and returns $no\_cycle = FALSE$ if a cycle is discovered.

Algorithm 4.1: Test for acyclicity

1. Initialize variable no_cycle with $no\_cycle = TRUE$
2. For all powers $l = 1, \ldots, \min(p, |E|)$ DO

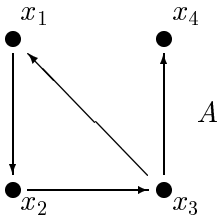
   For $i = 1, \ldots, p - l + 1$ DO

   i. Calculate the i–th row of A^l by $a_{ij}^{(l)} = \sum_{k=1}^{p} a_{ik}^{(l-1)} a_{kj}$, $j = 1, \ldots, p$

   ii. If $a_{ii} \neq 0$, RETURN $no\_cycle = FALSE$

   RETURN $no\_cycle = TRUE$

In order to examplify the algorithm, consider the graph in Figure 1, its adjacency matrix A and the first two powers of A, A^2 and A^3. In the latter, the cycle of the three vertices 1, 2 and 3 is indicated by the first three diagonal elements of A^3 which are different from zero, namely one.



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad A^3 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 1: A directed graph containing a cycle of length three, the corresponding adjacency matrix A and the first two powers of A, A^2 and A^3.

5 Reversible Jump MCMC for Learning in DAGs

In the following, we describe a reversible jump MCMC algorithm to estimate the posterior probability $\pi(d, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mid \boldsymbol{x})$ and, therefore, the required marginals, such as $\pi(d \mid \boldsymbol{x})$, to be employed for structural learning. First a very brief version of this

algorithm that summarizes the main steps is given below. Then the different steps are developed and presented in detail. For a general introduction to MCMC see, for instance, Brooks (1998). The reversible jump algorithm was proposed and described by Greeen (1995); it allows MCMC to sample simultaneously from parameter spaces of different dimensions.

Model Selection for Gaussian DAGs by RJMCMC
The state space of the Markov chain is made up by the vector of unknowns $(d, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$. We shall consider a random scan between the following $p+1$ moves:

1. Updating the dag d by adding, switching or deleting a directed edge, remaining always in the class of directed acyclic graphs. When adding or deleting an edge this move involves a change in dimensionality of the parameter space.

2. Updating $\boldsymbol{\beta}_{i|pa(i)}$ and $\sigma^2_{i|pa(i)}$ of the i–th regression model, $i = 1, \ldots, p$.

Updating of d: At first it has to be decided which kind of step (a birth, a death or a switch step) has to be performed. For this purpose, an edge (i, j) is randomly chosen. If it is already contained in the actual graph d ($a_{ij} = 1$), the deletion of this edge will be proposed (death step). If there already exists an edge from j to i ($a_{ji} = 1$), the direction of this edge will be changed in the proposed new graph d' (switch step). The third possibility is that there is no edge from i to j and vice versa ($a_{ij} = 0 \wedge a_{ji} = 0$). In this case adding of (i, j) will be proposed (birth step). If the step is not allowed because of a cycle in d', detected by algorithm 4.1, another pair (i, j) is randomly drawn. This is repeated until an allowed change of an edge (i, j) is chosen. We now describe in detail each of these steps (birth, death and switch), which are the only possible ones.
Birth step: After having proposed to add the edge (i, j) we check if the resulting new graph d' is acyclic. Only if this is the case the birth step is continued. Suppose this is the case and d' contains no cycles. By adding the edge (i, j) the parenthood structure of d' changes, as the variable j has one more parent than in d, namely i. It follows that one more regression coefficient arises, namely β'_{ji}.
The proposal distribution of this new coefficient, $q(\cdot)$, is chosen as Gaussian with zero mean and variance η^2. Making use of the available graph factorizations the posterior ratio $\mathcal{R}$ is thus given by:

$$\mathcal{R} \;=\; \frac{p(x_j \mid \boldsymbol{x}_{pa'(j)}, \boldsymbol{\beta}_{j|pa'(j)}, \sigma^2_{j|pa'(j)})\, p(\boldsymbol{\beta}_{j|pa'(j)} \mid \sigma^2_{j|pa'(j)})}{p(x_j \mid \boldsymbol{x}_{pa(j)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{j|pa(j)})\, p(\boldsymbol{\beta}_{j|pa(j)} \mid \sigma^2_{j|pa(j)})}.$$

The Jacobian matrix $\mathcal{J}$ of the mapping $g : \boldsymbol{\beta}_{i|pa(i)} \longmapsto (\boldsymbol{\beta}_{i|pa(i)}, \beta'_{ji}) = \boldsymbol{\beta}'_{i|pa(i)}$ is equal to one. Consider now the proposal ratio $\frac{r(d'|d)}{r(d|d')q(u)}$ where $r(d' \mid d)$ is the

probability to propose the new edge (i,j) in d', providing that there will be no cycles produced by this. Looking only at this kind of moves it is obvious that $r(d' \mid d) = r(d \mid d') = \frac{1}{n(n-1)}$, so they cancel out in the proposal ratio. If the additional edge (i,j) is proposed, the acceptance probability $\mathcal{A}_{\mathcal{B}}$ of the new dag d' is given by

$$\mathcal{A}_{\mathcal{B}} = \min\left\{1; \frac{p(x_j \mid \boldsymbol{x}_{pa'(j)}, \boldsymbol{\beta}_{j|pa'(j)}, \sigma^2_{j|pa'(j)})\; p(\boldsymbol{\beta}_{j|pa'(j)} \mid \sigma^2_{j|pa'(j)})}{q(\beta'_{ji})\; p(x_j \mid \boldsymbol{x}_{pa(j)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{j|pa(j)})\; p(\boldsymbol{\beta}_{j|pa(j)} \mid \sigma^2_{j|pa(j)})}\right\}.$$

Death step: As the deletion of an edge (i,j) cannot induce a cycle, acyclicity has not to be checked in this case. Again the number of parents of j changes, in fact it decreases because in the proposed dag d' the variable i is not a parent of j anymore. The corresponding regression coefficient β_{ij} vanishes, the dimension of the vector $\boldsymbol{\beta}_{i|pa(i)}$ decreases by one. The acceptance probability $\mathcal{A}_{\mathcal{D}}$ is given as the the reciprocal of the corresponding birth step from dag d' to d, which is

$$\mathcal{A}_{\mathcal{D}} = \min\left\{1; \frac{q(\beta_{ji})\; p(x_j \mid \boldsymbol{x}_{pa'(j)}, \boldsymbol{\beta}_{j|pa'(j)}, \sigma^2_{j|pa'(j)})\; p(\boldsymbol{\beta}_{j|pa'(j)} \mid \sigma^2_{j|pa'(j)})}{p(x_j \mid \boldsymbol{x}_{pa(j)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{j|pa(j)})\; p(\boldsymbol{\beta}_{j|pa(j)} \mid \sigma^2_{j|pa(j)})}\right\}.$$

Switch step: Note that this step, where only the direction of an existing edge is changed, is of special importance to move from a dag to an equivalent one. Like in the birth step first of all the acyclicity of the proposed new graph has to be verified. A switch step implies no changes in dimension as the original dag d and the proposed one d' differ only in the direction of an edge. By switching the edge (j,i) into (i,j) the number of parents of i is changing as those of j. While i looses j as parent and, therefore, the corresponding regression coefficient β_{ij} vanishes from $\boldsymbol{\beta}_{i|pa(i)}$, j gets i as a new parent and $\boldsymbol{\beta}_{j|pa(j)}$ increases by β'_{ji}. To achieve a high acceptance rate we propose to assign new values for **all** parameters of the two regression models for i and j proposed. This is emphasized by using twice the prime symbol in the notation, which denotes that a new value has been proposed, but not yet accepted: in the index for the new structure of the parents and for the parameter vectors $\boldsymbol{\beta}_{i|pa(i)}$, $\boldsymbol{\beta}_{j|pa(j)}$ as well as for the partial variances $\sigma^2_{i|pa(i)}$ and $\sigma^2_{j|pa(j)}$. This makes clear that really none of the old values of the two considered regression models is being kept. Let us first look at the regression model of the variable i. Proposals $(\boldsymbol{\beta}'_{i|pa'(i)}, \sigma'^2_{i|pa'(i)})$ are sampled from $p(\boldsymbol{\beta}'_{i|pa'(i)}, \sigma'^2_{i|pa'(i)} \mid \boldsymbol{x}_i, \boldsymbol{X}_{pa(i)})$ without considering the prior information for $\boldsymbol{\beta}'_{i|pa'(i)}$ and $\sigma'^2_{i|pa'(i)}$. Following Gelman et al. (1995) the proposal distribution $q_\sigma(.)$ of $\sigma'^2_{i|pa'(i)}$ is then given by

$$\sigma'^2_{i|pa'(i)} \mid \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)} \;\sim\; \text{Inv}\text{–}\chi^2(n - |pa'(i)|; s'^2), \tag{5}$$

$s'^2 = \frac{1}{n-|pa'(i)|}(\boldsymbol{x}_i - \boldsymbol{X}_{pa'(i)}\hat{\boldsymbol{\beta}}')'(\boldsymbol{x}_i - \boldsymbol{X}_{pa'(i)}\hat{\boldsymbol{\beta}}')$ and $\hat{\boldsymbol{\beta}}' = (\boldsymbol{X}'_{pa'(i)}\boldsymbol{X}_{pa'(i)})^{-1}\boldsymbol{X}'_{pa'(i)}\boldsymbol{x}_i$. The scaled inverse χ^2 form in (5) can also be expressed by an inverse gamma distribution with parameters $\frac{n-|pa(i)'|}{2}$ and $\frac{n-|pa(i)'|}{2}s^2$. Furthermore, the proposal distribution

of $\boldsymbol{\beta}'_{i|pa'(i)}$, $q_{\boldsymbol{\beta}}(.)$, is a normal distribution, namely

$$\boldsymbol{\beta}'_{i|pa'(i)} \mid \sigma'^2_{i|pa'(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)} \quad \sim \quad \mathrm{N}(\hat{\boldsymbol{\beta}}', \sigma'^2_{i|pa'(i)} V')$$

with $V' = (\boldsymbol{X}'_{pa'(i)} \boldsymbol{X}_{pa'(i)})^{-1}$ and again $\hat{\boldsymbol{\beta}}'$ as defined above. The proposals $\boldsymbol{\beta}'_{j|pa'(j)}$ and $\sigma'^2_{j|pa'(j)}$ for the j–th regression model are derived analogously. The proposed dag d' is then accepted with probability:

$$\begin{aligned}
\mathcal{A}_S \quad = \quad \min \Bigg\{ 1; & \frac{p(x_j \mid \boldsymbol{x}_{pa'(j)}, \boldsymbol{\beta}_{j|pa'(j)}, \sigma^2_{j|pa'(j)}) p(\boldsymbol{\beta}_{j|pa'(j)} \mid \sigma^2_{j|pa'(j)})\, p(\sigma^2_{j|pa'(j)})}{p(x_j \mid \boldsymbol{x}_{pa(j)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{j|pa(j)})\, p(\boldsymbol{\beta}_{j|pa(j)} \mid \sigma^2_{j|pa(j)})\, p(\sigma^2_{j|pa(j)})} \\
& \times \frac{p(x_i \mid \boldsymbol{x}_{pa'(i)}, \boldsymbol{\beta}_{i|pa'(i)}, \sigma^2_{i|pa'(i)})\, p(\boldsymbol{\beta}_{i|pa'(i)} \mid \sigma^2_{i|pa'(i)})\, p(\sigma^2_{i|pa'(i)})}{p(x_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)})\, p(\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)})\, p(\sigma^2_{i|pa(i)})} \\
& \times \frac{q_\beta(\boldsymbol{\beta}_{j|pa(j)} \mid \sigma^2_{j|pa(j)}, \boldsymbol{x}_j, \boldsymbol{X}_{pa(j)})\, q_\sigma(\sigma^2_{j|pa(j)} \mid \boldsymbol{x}_j, \boldsymbol{X}_{pa(j)})}{q_\beta(\boldsymbol{\beta}'_{j|pa'(j)} \mid \sigma'^2_{j|pa'(j)}, \boldsymbol{x}_j, \boldsymbol{X}_{pa'(j)})\, q_\sigma(\sigma'^2_{j|pa'(j)} \mid \boldsymbol{x}_j, \boldsymbol{X}_{pa'(j)})} \\
& \times \frac{q_\beta(\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa(i)})\, q_\sigma(\sigma^2_{i|pa(i)} \mid \boldsymbol{x}_i, \boldsymbol{X}_{pa(i)})}{q_\beta(\boldsymbol{\beta}'_{i|pa'(i)} \mid \sigma'^2_{i|pa'(i)}, \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)})\, q_\sigma(\sigma'^2_{i|pa'(i)} \mid \boldsymbol{x}_i, \boldsymbol{X}_{pa'(i)})} \Bigg\}.
\end{aligned}$$

Updating of $\boldsymbol{\beta}_{i|pa(i)}$: When updating the vector of regression coefficients of the i–th regression model the new vector $\boldsymbol{\beta}'_{i|pa(i)}$ can be drawn directly from its full conditional distribution which is again normal with covariance matrix

$$\begin{aligned}
\boldsymbol{\Sigma}_{i_full} \quad &= \quad \sigma^2_{i|pa(i)} (\boldsymbol{X}'_{pa(i)} \boldsymbol{X}_{pa(i)} + \alpha I)^{-1} \\
\text{and mean} \quad m_{i_full} \quad &= \quad \boldsymbol{\Sigma}_{i_full} \left(\frac{1}{\sigma^2_{i|pa(i)}} \boldsymbol{X}'_{pa(i)} \boldsymbol{x}_i + \frac{\alpha}{\sigma^2_{i|pa(i)}} \boldsymbol{b}_i \right).
\end{aligned}$$

Note that $\boldsymbol{X}_{pa(i)}$ denotes the $(n \times |pa(i)|)$–design matrix of the i–th regression model, which means that the first column contains only 1's and the other ones correspond to the observations of the respective parent.

Updating of $\sigma^2_{i|pa(i)}$: Again it is possible to draw $\sigma'^2_{i|pa(i)}$ directly from its full conditional distribution which is an inverse gamma distribution with parameters

$$\begin{aligned}
\delta_{i_full} \quad &= \quad \delta_{i|pa(i)} + 0.5(n+p) \\
\text{and} \quad \lambda_{i_full} \quad &= \quad \lambda_{i|pa(i)} + 0.5 \Big((\boldsymbol{x}_i - \boldsymbol{X}_{pa(i)} \boldsymbol{\beta}_{i|pa(i)})' (\boldsymbol{x}_i - \boldsymbol{X}_{pa(i)} \boldsymbol{\beta}_{i|pa(i)}) \\
& \qquad + (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)})' (\boldsymbol{\beta}_{i|pa(i)} - \boldsymbol{b}_{i|pa(i)}) \Big).
\end{aligned}$$

6 Simulations

To validate the algorithms we first consider two different situations with three variables being simulated. The first one describes the marginal independence $1 \perp 2$,

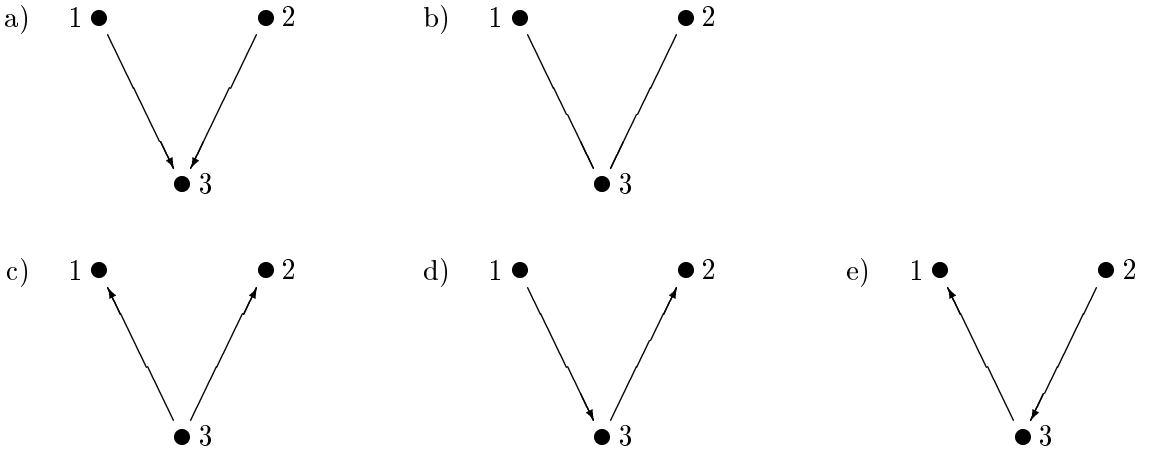


Figure 2: The situation of the marginal independence $1 \perp 2$ is given in a). The conditional independence $1 \perp 2 \mid 3$ can be expressed by the three Markov equivalent dags in c), d) , and e). Their essential graph is given in b).

the second one the conditional independence $1 \perp 2 \mid 3$. As it is shown in Figure 2 the first one can be represented by only one dag, and the second by three Markov equivalent dags. While in the former case the essential graph is the dag itself, in the latter case the essential graph is undirected and decomposable. We simulate 500 data, by specifying the covariance matrix C in the marginal independence and the concentration matrix K in the conditional independence case with

$$C = \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \qquad \text{and} \qquad K = \begin{pmatrix} 1 & 0 & -\rho \\ 0 & 1 & -\rho \\ -\rho & -\rho & 1 \end{pmatrix}$$

with ρ set to be 0.2 or 0.5. Thus, the (partial) correlation of the variables which are connected by an edge in the graphs is equal to ρ. Furthermore, we take the mean $\boldsymbol{\mu}$ different from zero, in fact $\boldsymbol{\mu} = (0, 1, 2)'$.

To compare fairly the two approaches proposed here, consistent hyperparameters of the priors have to be chosen. In our case, we set $\boldsymbol{\mu} = (0, 0, 0)'$, $\alpha_\mu = 4$, $\alpha = 4$, and $\boldsymbol{\Phi} = \mathbf{I}$ in the reversible jump algorithm for undirected decomposable graphs. Thus, we have to take $\boldsymbol{b} = (0, 0, 0)'$, $\alpha = 4$, $\delta_{i|pa(i)} = 0.5$, and $\lambda_{i|pa(i)} = 2$ in the algorithm for directed acyclic graphs. Both algorithms run for 55 000 iterations of which the first 5000 are regarded as burn–in time. The chain is thinned out every 10–th observation.

We calculate the posterior probabilities of each model, both over the dag and the udg space. Given the simple example considered, we calculate such probabilities not only by means of the reversible jump MCMC algorithm, but also by exact computations. This allows for evaluating the accuracy of the MCMC approximations. Notice that

	$\rho = 0.2,\ 1 \perp 2$				$\rho = 0.5,\ 1 \perp 2$			
models	rj_udg	cal_udg	rj_dag	cal_dag	rj_udg	cal_udg	rj_dag	cal_dag
$1 \perp 2$	-	-	.65	.77	-	-	.75	.97
$1 \perp 2 \mid 3$	.88	.88	.19	.20	0	0	0	0
compl.	.11	.11	.16	.03	1	1	.25	.03
others	.01	.01	0	0	0	0	0	0
	$\rho = 0.2,\ 1 \perp 2 \mid 3$				$\rho = 0.5,\ 1 \perp 2 \mid 3$			
models	rj_udg	cal_udg	rj_dag	cal_dag	rj_udg	cal_udg	rj_dag	cal_dag
$1 \perp 2$	-	-	.59	.59	-	-	0	0
$1 \perp 2 \mid 3$	.97	.97	.33	.39	.94	.96	.84	.96
compl.	.03	.03	.08	.01	.06	.04	.16	.04
others	0	0	0	.01	0	0	0	0

Table 1: The table shows the probabilities of the models $1 \perp 2$, $1 \perp 2 \mid 3$, the complete model case and the sum of all other possible models. The results are obtained from rj–algorithms for undirected decomposable models (rj_udg) and for dags (rj_dag), and also from exact calculations for the former (cal_udg) and the latter (cal_dag).

to make the two algorithms comparable we sum up the probabilities of the Markov equivalent dags, which turn out to be nearly equal in all cases. Of course, there remains an incomparability due to the three possible dags corresponding to marginal independences which are not Markov equivalent to any udg. Some of the results of our analysis are presented in Table 1. It can be seen that both algorithms very well approximate the exact posterior probabilities. Of course, as the udg model space is smaller, the MCMC algorithm over the udg space performs slightly better than the algorithm over the dag space. If the underlying model is that of marginal independence, there is no udg model equivalent to this. As a result, the complete saturated model turns out to be the best model, when correlations are stronger (i.e. $\rho = 0.5$) and the second one when correlations are weaker (i.e. $\rho = 0.2$). In this latter case, the conditional independence model is the most supported. In both cases an algorithm over the dag space (which is the appropriate one to consider), either exact or MCMC based, captures well the true model.

A more complex situation with eight variables is described by the dag depicted in Figure 3. The dag in Figure 3 contains conditional as well as marginal independences and, therefore, the variables can not be sampled directly via the concentration or covariance matrix of the joint distribution, as it was the case for the previous example.

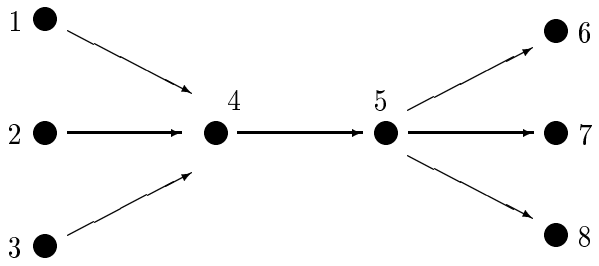


Figure 3: The complex model considered which has no further Markov equivalent dag.

Instead we have used the following recursion:

$$\begin{aligned}
X_{i1} &= f_1\epsilon_{i1} \qquad X_{i2} = f_2\epsilon_{i2} \qquad X_{i3} = f_3\epsilon_{i3} \\
X_{i4} &= X_{i1} + X_{i2} + X_{i3} + \epsilon_{i4} \\
X_{i5} &= X_{i4} + f_1\epsilon_{i5} \\
X_{i6} &= X_{i5} + f_1\epsilon_{i6} \qquad X_{i7} = X_{i5} + f_2\epsilon_{i7} \qquad X_{i8} = X_{i5} + f_3\epsilon_{i8},
\end{aligned}$$

where $\epsilon_{ij} \sim \mathrm{N}(0,1)$, $i = 1, \ldots, n$ and $j = 1, \ldots, 4$. We now consider two versions of this model, named 2a and 2b. In 2a, the factors f_k, $k = 1, 2, 3$, are always equal to one, so the edges $1 \to 4$, $2 \to 4$, and $3 \to 4$ represent the same strength of association. The same holds for $5 \to 6$, $5 \to 7$, and $5 \to 8$. In the second version the variance of the noise term ϵ_{ij} is influenced by different factors f_k, in fact $f_1 = 0.5$, $f_2 = 1$ and $f_3 = 2$.
The search space is extremely large; disregarding equivalences and acyclicity there are $3^{\frac{n(n-1)}{2}}$ possible graphs. Therefore, exact calculations of the posterior probabilities are obviously not possible. The reversible jump algorithm on this data leads to a posterior probability of only about 4% for the best model. It is thus more reasonable to consider the conditional independence graph obtained from inspecting the adjacency matrices, averaged over the Markov chain. Here, we consider a reversible jump MCMC algorithm over the dag space with 205000 iterations, of which the first $5,000$ are burned–in. The averaged adjacency matrix is given in Table 2. For a sample size of $n = 1000$ or even $n = 200$ the edges present in the true underlying graph have a probability of presence of at least 89%, in general more than 95%. Ignoring the orientation of the edges and looking only at the skeleton graph, the true edges appear in all the most probable dags. These results mean that the true model is clearly recognized. Of course, for a smaller sample size of $n = 100$ the results become less clear.

$$\bar{A}_{2a} = \begin{pmatrix}
0 & .03 & .03 & \mathbf{1} & .03 & .04 & .07 & .05 \\
 & (.09,.20) & (.09,.26) & (.99,.74) & (.10,.13) & (.07,.10) & (.16,.19) & (.16,.09) \\
.03 & 0 & .03 & \mathbf{1} & .22 & .04 & .04 & .03 \\
(.07,.21) & & (.09,.34) & (.96,.72) & (.66,.54) & (.11,.14) & (.10,.40) & (.17,.21) \\
.03 & .03 & 0 & \mathbf{1} & .07 & .04 & .03 & .10 \\
(.08,.17) & (.11,.25) & & (.97,.73) & (.10,.10) & (.13,.37) & (.07,.12) & (.12,.11) \\
0 & 0 & 0 & 0 & \mathbf{1} & .33 & .11 & .04 \\
(.01,.18) & (.03,.13) & (.03,.27) & & (.98,.82) & (.36,.45) & (.17,.35) & (.14,.14) \\
0 & 0 & 0 & 0 & 0 & \mathbf{.89} & \mathbf{.96} & \mathbf{.99} \\
(0,.06) & (.02,.14) & (0,.03) & (.02,.18) & & (.89,.81) & (.95,.83) & (.94,.91) \\
0 & 0 & 0 & 0 & .11 & 0 & .05 & .35 \\
(0,.03) & (0,.02) & (0,.03) & (0,.04) & (.11,.19) & & (.07,.14) & (.35,.15) \\
0 & 0 & 0 & 0 & .04 & .05 & 0 & .18 \\
(0,.04) & (0,.03) & (0,.02) & (0,.04) & (.05,.17) & (.06,.10) & & (.16,.15) \\
0 & 0 & 0 & 0 & .01 & .30 & .22 & 0 \\
(0,.02) & (0,.07) & (0,.02) & (0,.02) & (.06,.09) & (.31,.12) & (.18,.15) &
\end{pmatrix}$$

$$\bar{A}_{2b} = \begin{pmatrix}
0 & .05 & .09 & \mathbf{1} & .06 & .03 & .01 & .11 \\
 & (.20,.26) & (.78,.73) & (.34,.25) & (.22,.24) & (.08,.11) & (.21,.28) & (.25,.26) \\
.02 & 0 & .06 & \mathbf{1} & .20 & .01 & .03 & .04 \\
(.06,.06) & & (.75,.75) & (.27,.22) & (.38,.42) & (.05,.09) & (.08,.26) & (.18,.23) \\
.04 & .03 & 0 & \mathbf{.99} & .04 & .02 & .02 & .04 \\
(.03,.07) & (.08,.20) & & (.25,.19) & (.04,.06) & (.04,.07) & (.04,.05) & (.06,.09) \\
0 & 0 & .01 & 0 & \mathbf{1} & .13 & .07 & .08 \\
(.08,.09) & (.13,.25) & (.75,.81) & & (.42,.46) & (.2,.47) & (.12,.25) & (.10,.17) \\
0 & 0 & 0 & 0 & 0 & \mathbf{.96} & \mathbf{.97} & \mathbf{.88} \\
(.09,.09) & (.52,.45) & (.12,.12) & (.58,.54) & & (.82,.73) & (.83,.66) & (.24,.29) \\
0 & 0 & 0 & 0 & .04 & 0 & .08 & .44 \\
(.05,.06) & (.09,.15) & (.09,.14) & (.1,.31) & (.18,.27) & & (.10,.22) & (.74,.57) \\
0 & 0 & 0 & 0 & .03 & .04 & 0 & .29 \\
(.10,.13) & (.04,.16) & (.04,.09) & (.06,.11) & (.17,.25) & (.04,.10) & & (.20,.24) \\
0 & 0 & 0 & 0 & .03 & .27 & .21 & 0 \\
(.05,.04) & (.05,.06) & (.04,.06) & (.03,.03) & (.05,.06) & (.15,.10) & (.12,.12) &
\end{pmatrix}$$

Table 2: The averaged adjacency matrices of the more complex models with equal noise (2a-top) or with varying noise (2b-bottom). The first number outside the parentheses gives the estimated probability of an edge for a sample size of 1000 observations, the following two in parentheses for sample sizes of 200 and 100 observations.

It it also striking that some additional edges have a surprisingly high frequency, e.g. for $n = 1000$ the variables X_6 and X_8 are connected with a probability of 0.65. That means that they are regarded as conditionally dependent in more than half of the cases. In the second simulation 2b, which again describes noisier data, the just mentioned tendencies become even clearer. As one would expect, edges with a higher partial correlation are detected more easily than those with a lower one. In any case summarizing inspection of the mean adjacency matrix from 2b it can be stated that the algorithms do not recognize the marginal independence of X_1, X_2, and X_3 and the partial independence of X_6, X_7, and X_8 from the data. The separating role of X_4 and X_5 is, however, well detected.
For further comparison of the two approaches, we focuse again on the case illustrated in 2a with $n = 100, 200, 1000$. Both algorithms run for 205000 iterations, of which 5000 are burned–in. Like before, we use consistent priors as in the first example. The results are summarized in Table 3. Note that both algorithms well detect the true underlying skeleton if data are sufficiently informative as e.g. for $n = 1000$. Furthermore, in the undirected case, edges are added to moralize the immoralities present in Figure 3. It is remarkable that the two reversible jump approximations are rather similar in terms of estimated edge presence probability, although the number of MCMC iterations considered is indeed lower than the number of possible models. For smaller sample sizes (e.g. 200 or 100) more edges are estimated to be present with a high probability.

7 Concluding remarks

We have presented a novel reversible jump MCMC algorithm that allows to perform both quantitative and structural learning in Gaussian directed graphical models and have compared it with the approach proposed in Giudici and Green (1999) for Gaussian undirected graphical models which was therefore slightly extended to a mean different from zero. This comparison constitutes a first step towards MCMC model selection for dag models in the space of essential graphs. For this purpose, however, more graph theoretical research is still needed.
We have tested our algorithms with artificial data, and the results are quite satisfactory: The two algorithms give very similar results and both approximate well the exact probabilities, if they can be calculated.
Besides extending our method to the general space of all essential graphs, we believe further research has to be carried out in terms of applications of the present approaches to real data. This would additionally call for appropriate convergence diagnostics of the algorithms.

	rj_dag			rj_udg	
edge	• •	•→•	•←•	• •	•—•
1, 2	.94 (.84, .59)	.03 (.09,.20)	.03 (.07,.21)	0	1(1,1)
1, 3	.94 (.83, .57)	.03 (.09,.26)	.03 (.08,.17)	0	1 (1,.99)
1, 4	0 (0, .08)	**1** (.99,.74)	0 (.01,.18)	0	1(1,.99)
1, 5	.97 (.90, .81)	.03 (.10,.13)	0 (0,.06)	.96	.04 (.08,.19)
1, 6	.96 (.93, .87)	.04 (.07,.10)	0 (0,.03)	.99	.01 (.01,.03)
1, 7	.93 (.84, .77)	.07 (.16,.19	0 (0,.04)	.99	.01 (.01,.06)
1, 8	.95 (.84, .89)	.05 (.16,.09)	0 (0,.02)	1	.00 (.01,.02)
2, 3	.94 (.80, .41)	.03 (.09,.34)	.03 (.11,.25)	0	1 (1,1)
2, 4	0 (.01, .15)	**1** (.96,.72)	0 (.03,.13)	0	1 (1,1)
2, 5	.78 (.32, .32)	.22 (.66,.54)	0 (.02,.14)	.80	.20 (.67,.70)
2, 6	.96 (.89, .84)	.04 (.11,.14)	0 (0,.02)	.98	.02 (.07,.14)
2, 7	.96 (.90, .57)	.04 (.10,.40)	0 (0,.03)	1	0 (.04,.30)
2, 8	.97 (.83, .72)	.03 (.17,.21)	0 (0,.07)	1	0 (.09,.14)
3, 4	0 (0,0)	**1** (.97,.73)	0 (.03,.27)	0	1 (1,1)
3, 5	.93 (.90, .87)	.07 (.10,.10)	0 (0,.03)	.93	.07 (.06,.13)
3, 6	.96 (.87, .60)	.04 (.13,.37)	0 (0,.03)	.99	.01 (.01,.06)
3, 7	.97 (.93, .86)	.03 (.07,.12)	0 (0,.02)	1	0 (.01,.03)
3, 8	.90 (.88, .87)	.10 (.12,.11)	0 (0,.02)	.99	.01 (.01,.02)
4, 5	0 (0,0)	**1** (.98,.82)	0 (.02,.18)	.08	.92 (1,1)
4, 6	.67 (.44, .51)	.33 (.36,.45)	0 (0,.04)	.70	.30 (.53,.82)
4, 7	.89 (.83, .61)	.11 (.17,.35)	0 (0,.04)	.99	.01 (.17,.47)
4, 8	.96 (.86, .84)	.04 (.14,.14)	0 (0,.02)	.96	.04 (.17,.19)
5, 6	0 (0,0)	**.89** (.89,.81)	.11 (.11,.19)	0	1 (1,1)
5, 7	0 (0,0)	**.96** (.95,.83)	.04 (.05,.17	0	1 (1,1)
5, 8	0 (0,0)	**.99** (.94,.91)	.01 (.06,.09)	0	1 (1,1)
6, 7	.9 (.87, .76)	.05 (.07,.14)	.05 (.06,.10)	.92	.08 (.20,.32)
6, 8	.35 (.34, .73)	.35 (.35,.15)	.30 (.31,.12)	.49	.51 (.80,.41)
7, 8	.60 (.66, .70)	.18 (.16,.15)	.22 (.18,.15)	.48	.52 (.48,.36)

Table 3: Posterior probabilities of the different egdes in the case of the complex model with equal noise for the two reversible jump algorithms. The first number outside the parentheses gives the estimated probability of an edge for a sample size of 1000 observations, the following two in parentheses at sample sizes for 200 and 100 observations.

8 Acknowledgements

This research was initiated during a visit of the first Author at the second, which was supported by a grant from the Highly Structured Stochastic Systems initiative of the European Science Foundation. We also acknowledge support from the University of Pavia, the German National Science Foundation, the Graduate College "Applied Algorithmic Mathematics" and the SFB 386. We thank Iris Pigeot for helpfull comments and Stefan Lang for incorporating the reversible jump algorithm for dags into the software package *BayesX*.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997a). A Characterization of Markov equivalence Classes for Acyclic Digraphs. *The Annals of Statistics*, **25**, 505–541.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997b). On the Markov equivalence of Chain Graphs, Undirected Graphs, and Acyclic Digraphs. *Scandinavian Journal of Statistics*, **24**, 81–102.

Brooks, S. P. (1998). Markov Chain Monte Carlo Method and its Application. *The Statistician*, **47**, 69–100.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman and Hall, London.

Geiger, D. and Heckerman, D. (1999). Parameter priors for directed acyclic graphical models and the characterisation of several probability distributions. Submitted for publication.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian Networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, 235–243.

Giudici, P. and Green, P. J. (1999). Decomposable Graphical Gaussian Model Determination. *Biometrika*, **86**, 785–801.

Giudici, P., Green, P. J., and Tarantola, C. (1999). Efficient Model Determination for Discrete Graphical Models. Submitted for publication.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711–32.

Lang, S. and Brezger, A. (2000). BayesX–Software for Bayesian Inference Based on Markov Chain Monte Carlo Simulation Techniques. Discussion Paper 187, Sonderforschungsbereich 386, Ludwig–Maximilians–Universität, München, Germany.

Schachter, R. and Kenley, C. (1989). Gaussian Influence Diagrams. *Management Science*, **35**, 527–550.