Fronk:

Model Selection for Dags via RJMCMC for the Discrete and Mixed Case

Sonderforschungsbereich 386, Paper 271 (2002)

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

Projektpartner

Model Selection for Dags via RJMCMC for the Discrete and Mixed Case

Eva-Maria Fronk

Abstract

Based on a reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm which was developed by Fronk and Giudici (2000) to deal with model selection for Gaussian dags, we propose a new approach for the pure discrete case. Here, the main idea is to introduce latent variables which then allow to fall back on the already treated continuous case. This makes it also straightforward to tackle the mixed case, i.e. to deal simultaneously with continuous and discrete variables. The performance of the approach is investigated by means of a simulation study for different standard situations. In addition, a real data application is provided.

Keywords: Bayesian model selection; dag models; reversible jump Markov Chain Monte Carlo.

1 Introduction

Model selection for graphical models is confronted with the problem that the search space increases more than exponentially with the number of variables incorporated in the analysis. Due to the huge number of possible models, it is not feasible to judge them all by a goodness of fit criterion like the AIC or the BIC and to find the best model with respect to this criterion. Therefore, it makes sense to focus on parts of the search space, e.g. to reduce the search space to a special class of models, or to restrict it successively in course of the search. Madigan and Raftery (1994) e.g. reduce the search space by looking only at decomposable graphs. The latter concept can be found in approaches like the stepwise selection strategy of MIM (Edwards, 2000) or the Edwards–Havránek strategy (1985, 1987) .
In the context of directed acyclic graphs (dags), another problem arises by the so–called Markov equivalence, which means that different graphs can reflect the same statistical model. These equivalent graphs form equivalent classes which can be represented by one unique graph, the essential graph, which is a chain graph with special properties (Andersson et al. 1997a). Performing model selection in the space of essential graphs would reduce the search space enormously, but has to be bought by the price of a much more complex search algorithm which is again time consuming (Perlman, 2002).

In this paper, we propose a fully Bayesian approach using the reversible jump MCMC (RJMCMC) algorithm which was introduced by Green (1995) and which is able to deal with the changing dimension of the search space. In our opinion, this method offers several advantages. First, it traverses the search space randomly by visiting the different models according to their posterior probability. This probability provides a measure of goodness of fit which can be easily interpreted and makes it therefore possible to

compare different models in a sensible way. Another benefit of the MCMC approach is the possibility to combine the structural or qualitative learning (model selection) and the quantitative learning (estimation of the parameters) in a closed framework. Furthermore, it is very flexible and therefore easily extensible to problems which occur from e.g. missing data or latent variables.
Model selection by RJMCMC has been developed by Giudici and Green (1999) for pure continuous variables and by Giudici, Green, and Tarantola (1999) for the pure discrete case. Both approaches focus on undirected decomposable models which allow a factorisation by cliques and separators so that only local computations have to be performed. In the context of dags a reversible jump algorithm for the continuous case has been introduced by Fronk and Giudici (2000). Here, local computations are again possible because of the factorisation property of dags. As a disadvantage of this approach it has to be admitted that the equivalence classes are not taken into account, which has the consequence that the Markov chain moves in the space of all dags and not in the smaller one of their essential graphs.

Relying on the already developed methodology for Gaussian dags, here we present an approach to handle binary variables or even continuous and binary variables simultaneously. The paper is organized as follows. After a brief introduction into the main assumptions and the algorithm for the continuous case, the extension to binary variables is given in Section 2. The main idea consists in sampling a latent variable whose distribution has the same moments as the binary one. In order to estimate them, consequently the problem can be traced back to the already solved continuous case. To account for higher order interactions, which could not occur in the Gaussian case, in a second step the auxiliary construct of a so–called interaction graph is introduced and explained. We check the performance of the RJ algorithm by a simulation study in Section 3. Under the assumption of a conditional Gaussian (CG) distribution the simultaneous modeling of continuous and discrete variables is investigated in Section 4. A real data application is given in Section 5. Finally, we summarize the results and give an outlook to further research.

2 Algorithm for Binary Variables

For convenience, let us first recall the reversible jump algorithm for the Gaussian case by Fronk and Giudici (2000) before introducing the new approach for binary variables without and with interactions.

2.1 Gaussian Case

In Fronk and Giudici (2000), a reversible jump algorithm for model selection in Gaussian dags is proposed for which the p random variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ are assumed to follow a multivariate normal distribution. Thus, for each univariate conditional distribution $X_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}, d$ it holds that

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

where $\boldsymbol{x}_{pa(i)}$ denotes the known vector of the parent variables of X_i and $\boldsymbol{\beta}_{i|pa(i)}$ the vector of the corresponding coefficients. The conditioned variance is given by $\sigma^2_{i|pa(i)}$ and d is the underlying dag. The

analogy to a regression model is obvious. Further, the following assumptions are made

$$\begin{aligned}
\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)}, d &\sim \mathrm{N}_{|pa(i)|+1}\left(\boldsymbol{b}_{i|pa(i)}, \frac{1}{\alpha_i}\sigma^2_{i|pa(i)}\boldsymbol{I}\right), \\
\sigma^2_{i|pa(i)} \mid d &\sim \mathrm{IG}\left(\delta_{i|pa(i)}, \lambda_{i|pa(i)}\right), \\
p(d) &= 1/a,
\end{aligned}$$

that is the vector of coefficients $\boldsymbol{\beta}_{i|pa(i)}$ is also multivariate normally distributed, the variance $\sigma^2_{i|pa(i)}$ follows an inverse gamma distribution and for the dag d a discrete uniform distribution is assumed. The parameter vector $\boldsymbol{b}_{i|pa(i)}$ and the parameters α_i, $\delta_{i|pa(i)}$, and $\lambda_{i|pa(i)}$ have to be chosen sensibly, where a denotes the number of possible dags and $\boldsymbol{I}$ the identity matrix. Making use of the likelihood modularity and the global parameter independence (Geiger and Heckermann, 1999) the following factorisation of the joint distribution can be obtained:

$$\begin{aligned}
&p(\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) \\
&= p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\sigma}^2 \mid d)\, p(d) \\
&= \prod_{i=1}^{p} p(\boldsymbol{x}_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}) \prod_{i=1}^{p} p(\boldsymbol{\beta}_{i|pa(i)} \mid \sigma^2_{i|pa(i)}) \\
&\quad \prod_{i=1}^{p} p(\sigma^2_{i|pa(i)})\, p(d)
\end{aligned}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{1|pa}, \ldots, \boldsymbol{\beta}'_{p|pa})'$ and $\boldsymbol{\sigma}^2 = (\sigma^2_{1|pa}, \ldots, \sigma^2_{p|pa})'$.

To represent the dags and to check them for acyclicity we use the concept of adjacency matrices (Thulasiraman and Swamy, 1992). Moving through the search space by the sampled Markov chain three different changes of the current dag are allowed for. If two vertices are not connected a new edge can be inserted between them (birth step), whereas an already existing edge can be deleted (death step) or turned in its orientation (switch step). Based on the above assumptions and restrictions, one loop of the proposed RJMCMC algorithm consists of three different steps:

Reversible Jump Algorithm for Gaussian Case:

1. Updating of d by a birth, death or switch step; the first and the last need a check for acyclicity.
2. Updating of $\boldsymbol{\beta}_{i|pa}$, $i = 1, \ldots, p$.
3. Updating of $\sigma^2_{i|pa}$, $i = 1, \ldots, p$.

Note, that the first step corresponds to qualitative learning. Here, the change in dimension occurs if a death or birth step is carried out. Steps 2 and 3 just update the parameters of an already existing dag d, and therefore stand for quantitative learning.

2.2 Binary Variables

Now we consider the situation of p binary variables of which the joint distribution is assumed to be multinomial. The influence on a variable X_i from its known parents $\boldsymbol{x}_{pa(i)}$ shall be given by a probit model, i.e.

$$p_i \;=\; E(X_i \mid \boldsymbol{x}_{pa(i)}) \;=\; \Phi(\boldsymbol{x}'_{pa(i)}\boldsymbol{\beta}_{i|pa(i)}, \sigma^2_{i|pa(i)}), \tag{1}$$

where $i = 1, \ldots, p$ and $\Phi(\mu, \sigma^2)$ denotes the cdf of the normal distribution. In a first step, we ignore any interactions although they may be present as we have left the Gaussian case. Following an idea of Albert and Chib (1993) we reduce the above situation to the one of continuous variables by introducing latent variables Z_i with

$$Z_i \overset{iid}{\sim} \mathrm{N}(\boldsymbol{x}'_{pa(i)}\boldsymbol{\beta}_{i|pa(i)}, 1) \qquad \text{and} \qquad P(X_i = 1 \mid z_i) = \begin{cases} 1, & z_i \geq 0 \\ 0, & z_i < 0. \end{cases}$$

Therewith, we again obtain a Gaussian distribution for the full conditional of Z_i, i.e. $Z_i \mid \boldsymbol{x}_{pa(i)}, x_i, \boldsymbol{\beta}_{i|pa(i)} \sim \mathrm{N}(\boldsymbol{x}'_{pa(ki)}\boldsymbol{\beta}_{i|pa(i)}, 1)$, which is truncated at the left by 0 if $x_i = 1$ and otherwise, i.e. if $x_i = 0$, at the right. The joint distribution of $\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2$, and d is given by

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) &= p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, d) \\ &= \prod_{i=1}^{p} p(x_i \mid z_i, d)\, p(z_i \mid \boldsymbol{x}_{pa(i)}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d)\, p(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, d). \end{aligned}$$

Thus, we can extend the above algorithm to the discrete case by adding one additional step where we draw the latent variable Z_i from its full conditional for all binary variables X_i, $i = 1, \ldots, p$, and follow the remaining steps of the algorithm by using Z_i instead of X_i, which leads to:

Reversible Jump Algorithm for Binary Variables (Ignoring Interactions):

1. For X_i, $i = 1, \ldots, p$, draw Z_i from its full conditional $\mathrm{N}(\boldsymbol{x}'_{pa(i)}\boldsymbol{\beta}_{i|pa(i)}, 1)$, which is truncated at the left by 0 if $x_i = 1$ and at the right if $x_i = 0$.
2. Add, delete, or switch a directed edge like in the Gaussian case, but take the utility Z_i instead of X_i as response in ith regression model; the covariables $\boldsymbol{x}_{pa(i)}$ of the ith model remain unchanged.
3. Update $\boldsymbol{\beta}_{i|pa(i)}$, $i = 1, \ldots, p$.
4. Update $\sigma^2_{i|pa(i)}$, $i = 1, \ldots, p$.

2.3 Accounting for Interactions

the problem of present interactions is coped with by introducing a particular type of graph, which we call interaction graph. The general idea is that within this graph the interactions are treated as own variables, so the interaction graph can be regarded as an auxiliary graph the algorithm is able to deal with. Due to the enormous complexity we restrict the following considerations to two way interactions which seem

to be sufficient for most situations in practice. Further, we have to introduce the following notations for the interaction graph. We distinguish between the real (or main) parents $pa_m(i)$ of a variable i and those parents $pa_i a(i)$ that are caused by the interactions among the main parents. They are summarized in $pa(i) = pa_m(i) \cup pa_{ia}$. For the jth regression model, the interactions of the variable i with the other parents of j, $pa(j) \setminus i$, are denoted by $j(i)$.

In contrast to Equation (1), we now assume the following link between each variable X_i and its known parents $\boldsymbol{x}_{pa\_m(i)}$:

$$p_i \quad = \quad \Phi(\boldsymbol{\beta}'_{i|pa\_m(i)}\boldsymbol{x}_{pa\_m(i)} + \boldsymbol{\beta}'_{i|pa\_ia(i)}\boldsymbol{x}_{pa\_ia(i)}, \sigma^2_{i|pa(i)}), \tag{2}$$

where $\boldsymbol{x}_{pa\_ia(i)}$ is the vector of the $\binom{|\boldsymbol{x}_{pa\_m(i)}|}{2}$ interactions among the parents of X_i and $\boldsymbol{\beta}_{i|pa\_ia(i)}$ denotes the vector of the same dimension with the corresponding coefficients. The interaction graph, which represents $\boldsymbol{x}_{pa\_ia(i)}$ as own variables, holds two restrictions.

Proposition: Assuming Equation (2), directed acyclic graphs that include not only the regarded p variables but also their $\binom{p}{2}$ two way interactions as vertices have the following two properties:

1. A vertex that represents an interaction is a parent of another variable if and only if this is also true for its parents.
2. A vertex that represents an interaction has exactly two parents, namely those forming the represented interaction.

As a result of these restrictions, the additional edges of an interaction graph are clearly determined, as all possible interactions of the parents are always regarded. It follows that the importance of an interaction is not indicated by its presence in the model but by the strength of its corresponding coefficient. As another consequence, the death, the birth, and especially the switch step now become much more complex since in the interaction graph more than one edge has typically to be changed at the same time. Furthermore, in most situations the switch step turns out to be a dimension changing step, too. We discuss the different steps in detail:

Birth and death step: Figure 1 shows a very simple situation where it can be seen that adding one edge in the real graph can result in adding several steps in the interaction graph. In general, it can be stated that adding an edge from i to j in the real graph d implicates $1 + |pa_m(j)|$ new edges in the interaction graph, where $|pa_m(j)|$ denotes the number of parents of j in the real graph. Consequently, the ith regression model is enlarged by $1+|pa_m(j)|$ new covariables. We denote the vector of their coefficients by $\boldsymbol{\beta}^\star_{ji} = (\beta^\star_{ji}, \boldsymbol{\beta}^{\star\prime}_{j(i)})'$. Using the terms of the reversible jump algorithm provided by Green (1995) the mapping from the former state space to the new one is then given by

$$g_B : \ (\boldsymbol{\beta}_{j|pa(j)}; \boldsymbol{u}_B) \quad \longmapsto \quad g_B(\boldsymbol{\beta}_{j|pa(j)}, \boldsymbol{\beta}^\star_{ji}) = \boldsymbol{\beta}_{j|pa^*(j)}. \tag{3}$$

The random variable $\boldsymbol{U}_B$ has the dimension $|pa_m(j)|+1$. We choose the normal distribution $\mathrm{N}_{1+|pa_m(j)|}(\boldsymbol{\mu}_u, \sigma^2_u \boldsymbol{I})$ as its proposal distribution where $\boldsymbol{\mu}_u$ denotes the least squares estimator under the restriction that

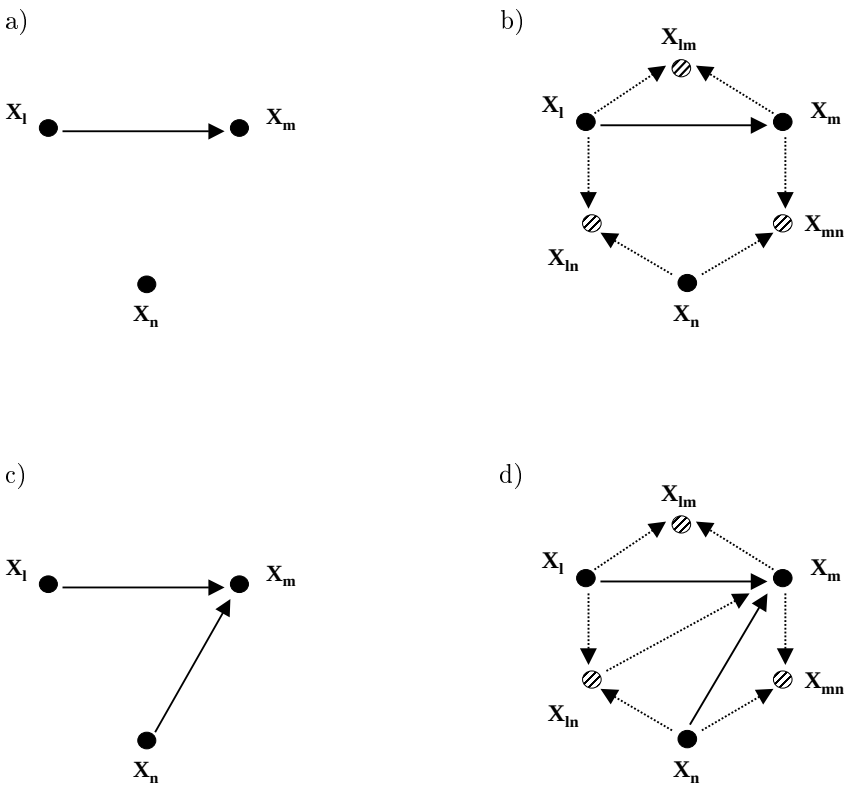

Figure 1: Example for a birth step in a graph with three vertices: The real graph with one resp. two edges is given in a) and c), the corresponding interaction graphs the algorithm works with are shown in b) resp. d).

the regression coefficients of the previous variables stay the same. Like in the Gaussian case, the probability of proposing the transition from d to $d^\star$ by a birth step is equal to proposing the opposite, namely moving from $d^\star$ to d by a death step. Thus, the proposal ratio reduces to

$$\mathcal{P}_B = \frac{r_D(d^\star)}{r_B(d)q(\boldsymbol{u}_B)} = \frac{1}{q(\beta^\star_{ji}, \boldsymbol{\beta}^\star_{j(i)})}. \tag{4}$$

As in addition the determinant of the Jacobi–matrix of (3) is equal to 1 we derive the acceptance ratio of a proposed dag $d^\star$ as

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

The acceptance ratio of the corresponding death step $\mathcal{A}_{\mathcal{D}}$ from $d^\star$ to d, where the edge from i to j is deleted in the dag $d^\star$, is given by the reciprocal value of $\mathcal{A}_{\mathcal{B}}$.

Switch step: As already mentioned, working with the interaction graphs the switch step turns out to be the most crucial one. As it is shown in Figure 2, switching the edge (j,i) into (i,j) the number of involved regression coefficients in the two regression models of i and j does not have to remain unchanged. As a consequence, working with the interaction graph the switch step can also change the dimension like the birth or death step. This always occurs if $|pa_m(i)| - 1 \neq |pa_m(j)|$. In the following, we assume that the total number of parameters is increasing, which we call a switch1 step. In analogy, speaking of a switch2 step reflects a situation where the total number of parameters is decreasing. Our investigations have shown that a good acceptance probability is only achieved when all parameters of the considered models are resampled. That is the coefficients of the actual variables, those of the interaction variables, and the variances. Thus, the new random vector $\boldsymbol{U}^{(1)}$ can be subdivided into the following components:

$$\begin{aligned}\boldsymbol{U}^{(1)} &= (\boldsymbol{\beta}^\star_{i|pa^\star(i)}, \boldsymbol{\beta}^\star_{j|pa(j)}, \beta^\star_{ji}, \boldsymbol{\beta}^\star_{j(i)}, \sigma^{\star 2}_{i|pa^\star(i)}, \sigma^{\star 2}_{j|pa^\star(j)}) \\ &= (\boldsymbol{\beta}^\star_{i|pa^\star(i)}, \boldsymbol{\beta}^\star_{j|pa^\star(j)}, \sigma^{\star 2}_{i|pa^\star(i)}, \sigma^{\star 2}_{j|pa^\star(j)}),\end{aligned}$$

where $pa^\star$ indicates that the parent structure of the proposed dag $d^\star$ and not of the current dag d is referred to. The vector $\boldsymbol{U}^{(1)}$ is of dimension

$$m_1 = |pa^\star(i)| + |pa^\star(j)| + 4.$$

Let $\boldsymbol{\theta}^{(1)}$ denote all parameters of the regression models that are due to the underlying graph d whereas $\boldsymbol{\theta}^{(2)}$ refers to those of the graph $d^\star$. A realisation of $\boldsymbol{U}^{(1)}$ is then mapped to $(\boldsymbol{\theta}^{(2)}; \boldsymbol{u}^{(2)})$ by

$$\begin{aligned} g_1 : \quad \mathbb{R}^{n_1+m_1} &\longrightarrow \mathbb{R}^{n_2+m_2} \\ (\boldsymbol{\theta}^{(1)}; \boldsymbol{u}^{(1)}) &\longmapsto (\boldsymbol{\theta}^{(2)}; \boldsymbol{u}^{(2)}) \end{aligned}$$

with $\boldsymbol{u}^{(2)}$ containing all components of $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^{n_1}$ which change or vanish in the parameter vector $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^{n_2}$, i.e.

$$\boldsymbol{u}^{(2)} = (\boldsymbol{\beta}_{i|pa(i)}, \boldsymbol{\beta}_{j|pa(j)}, \sigma^2_{i|pa(i)}, \sigma^2_{j|pa(j)}).$$

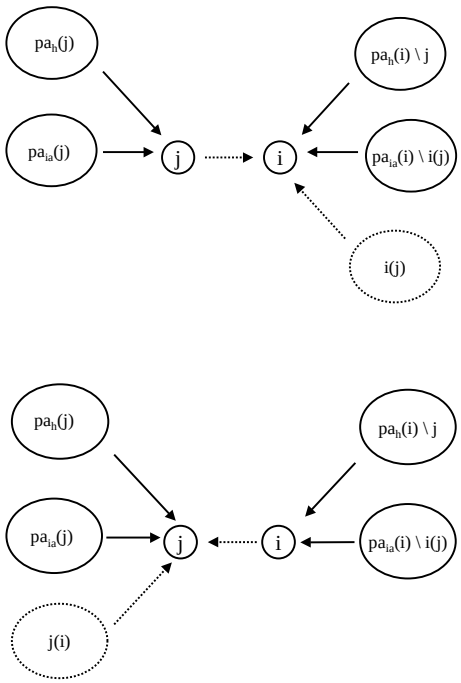

Figure 2: Changes in the involved subgraph of the interaction graph that are caused by a switch step. In both cases, the assignment of the parents refers to the upper graph. The parent sets of i and j do not need to be disjunctive.

The dimension of $\boldsymbol{u}^{(2)}$ results in $m_2 = |pa(i)| + |pa(j)| + 4$, which implies that the matching of the dimensions is given as $n_1 + m_1 = n_2 + m_2$. The reversal step from $d^\star$ to d is now denoted as switch2 step and is represented by

$$\begin{array}{rccc} g_2: & I\!R^{n_2+m_2} & \longrightarrow & I\!R^{n_1+m_1} \\ & (\boldsymbol{\theta}^{(2)}; \boldsymbol{u}^{(2)}) & \longmapsto & (\boldsymbol{\theta}^{(1)}; \boldsymbol{u}^{(1)}). \end{array}$$

The proposal is chosen in analogy to the Gaussian case without interactions, i.e. the variances are drawn from their distributions given the parents and their interactions and the coefficients are drawn from their conditional distribution given the same plus the variances. Thus, we get again a normal distribution for the proposal of the regression coefficients and an inverse gamma distribution for the variances. For more details see Fronk and Giudici (2000). The probability of proposing a move from d to $d^\star$ in the switch1 step is equal to the one of the reversal move of the switch2 step, namely $\frac{1}{p(p-1)}$, and the determinant of the Jacobi matrix, $\mathcal{J} = \frac{\partial g_B(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\partial(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}$, again results in 1. Altogether, we in principle get the same acceptance ratio like in the Gaussian case without interactions. The acceptance ratio of the reversal switch2 step is again obtained by the reciprocal value.

It should, however, be noticed that the acceptance ratio of the switch step is in general very low: In

many situations the Markov chain already stays in a graphical model which describes the data very well. The new statistical model, which is proposed and represented by the dag $d^\star$, can now be an equivalent or a different model. If the latter is true there is few reason to change the model, because the association structure of the data is already very well characterized and the probability that the new model is even better is small. Consequently, the acceptance ratio will be low. If otherwise $d^\star$ represents a model which is equivalent to d the proposal has to be much better than the already fitted model of d to have a chance to be accepted. Thus, the chance to switch is low again. As another drawback of this procedure it has to be mentioned that drawing from the above proposal distributions is very time consuming as it is always performed for all variables involved.
To tackle these problems, we suggest an alternative approach for the switch step: If the proposed model $d^\star$ is equivalent to d it should not be possible to distinguish between them statistically at least in theory. The differences that occur in reality are due to insufficient updates resp. proposals. It seems therefore adequate to accept a switch step into an equivalent model with a fixed acceptance probability of 0.5. This accelerates the process enormously because no proposals and no acceptance ratios have to be calculated. There are two possible versions of this procedure: The first one consists in performing the switch step only into equivalent models using the acceptance ratio of 0.5. This includes the fact that different statistical models have to be reached by a birth or death step. The second possibility is to work with the complex switch step for movements out of an equivalent class and take the simpler alternative otherwise.

3 Simulation Study

In the following, different standard situations of marginal and conditional independence are simulated. We sample data sets of different sizes and also distinguish between the model selection with and without accounting for interactions. For a given underlying graph and sample size each data set is sampled 20 times and the averaged results are presented. To average the posterior probabilities of the selected models we only take the ten best models of each run into account. We choose a runtime of 55000 iterations from which the first 5000 are not considered at all (burn–in). We then take the values of every 20th iteration for our estimation of the probability distribution $p(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, d \mid \boldsymbol{X})$. The data are generated by the software package *BayesX* (Lang and Brezger, 2001).

3.1 Marginal Independence

We first investigate the situation of marginal independence of three variables X_0, X_1, X_2 that become dependent conditioning on a fourth one, namely X_3. The corresponding equivalence class 1 is represented by the dag in Figure 3. We sample two kinds of data sets that have this structure as their underlying graph. The first one, data type 1a, shows no interaction in its sampling scheme:

$$\begin{aligned} X_{l0} &\sim B(1, 0.5); \qquad X_{l1} \sim B(1, 0.5); \qquad X_{l2} \sim B(1, 0.5); \\ X_{l3} &\sim B(1, p_{l3}), \qquad p_{l3} = \Phi(x_{l0} + x_{l1} + x_{l2}), \end{aligned}$$

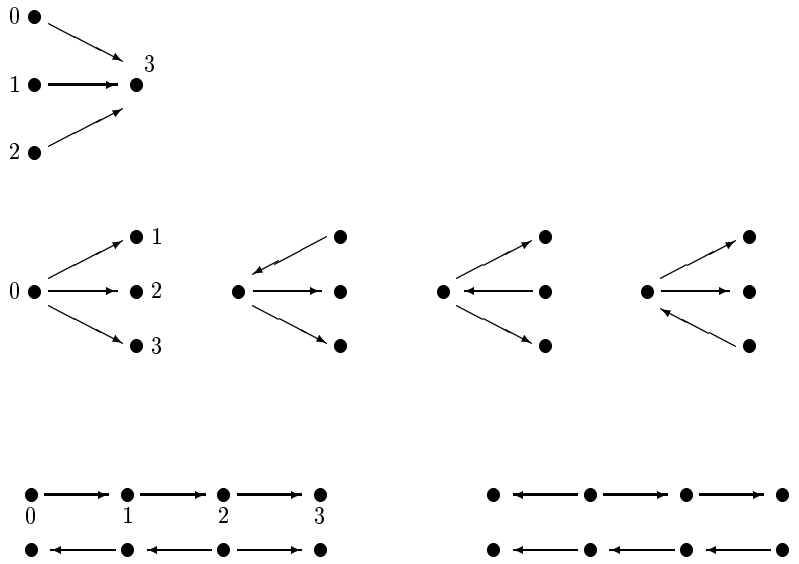

Figure 3: Equivalence classes 1 (above), 2 (middle), and 3 (below); equivalence classes 1 is represented by one dag, the others are consisting of four equivalent dags.

where $l = 1, \ldots, n$ and $n = 100, 200, 500, 1000$. In the second type of data set, 1b, the edges of the underlying dag are only due to the interactions, as it has the sampling scheme

$$\begin{aligned} X_{l0} &\sim B(1, 0.5); \qquad X_{l1} \sim B(1, 0.5); \qquad X_{l2} \sim B(1, 0.5); \\ X_{l3} &\sim B(1, p_{l3}), \qquad p_{l3} = \Phi(x_{l0}x_{l1} + x_{l0}x_{l2} + x_{l1}x_{l2}). \end{aligned}$$

The outcomes are summarized in Table 1. It is not surprising that in all cases an increasing number of observations increases the security of finding the underlying model which is indicated by a higher posterior probability as well as by a greater difference to the second best model. It can also be concluded that in nearly all cases the true model is found. Except for the data set of type 1a with 1000 observations, the model selection accounting for interactions is always better regardless of the fact whether the data contain interactions or not.

To gain more information about the performance of the model selection we have a closer look at the average of the 20 averaged adjacency matrices of the selected models. As there exists no sensible averaging of models – at least not by averaging the adjacency matrices of the sampled models – these matrices of course cannot be interpreted as models. They just give the probability of each possible edge to appear in a selected dag, but tell nothing about the combination of their appearance.

Considering these averaged adjacency matrices and comparing the two selection types, it turns out that a higher probability of the correct edges, i.e. of edges that exist in the underlying model, does not have to go along with a lower probability of the wrong ones. The adjacency matrices in Figure 4 may serve as an example for a sample size of $n = 1000$ observations.

	search without ia		search with ia	
	$\bar{\hat{p}}$ (r)	$\bar{\hat{p}_a}$	$\bar{\hat{p}}$ (r)	$\bar{\hat{p}_a}$
	data type 1a (without ia)			
$n = 100$	0.04 (1)	0.03	0.10 (1)	0.09
$n = 200$	0.11 (1)	0.05	0.22 (1)	0.12
$n = 500$	0.22 (1)	0.10	0.25 (1)	0.11
$n = 1000$	0.44 (1)	0.08	0.28 (1)	0.05
	data type 1b (with ia)			
$n = 100$	–	0.03	0.06 (2)	0.07
$n = 200$	0.09 (1)	0.07	0.22 (1)	0.15
$n = 500$	0.23 (1)	0.08	0.37 (1)	0.16
$n = 1000$	0.29 (1)	0.09	0.45 (1)	0.15

Table 1: Averaged results of the 20 data sets of each combination: The averaged estimated posterior probability of the underlying model is denoted by $\bar{\hat{p}}$ and its rank under the ten best models by r. The highest averaged estimated posterior probability of an alternative model is given by $\bar{\hat{p}_a}$.

3.2 Conditional Independence

The algorithm should also be able to detect situations of conditional independence. Thus, we consider two situations: The first one is given by equivalence class 2, the second by equivalence class 3, which both come up with four different but equivalent dags as it can be seen in Figure 3. The data sets with equivalence classes 2 and 3 are generated by the sampling schemes

$$\begin{aligned} X_{l0} &\sim B(1, 0.5), & X_{l2} &\sim B(1, p_{l2}),\ p_{l2} = \Phi(x_{l0}), \\ X_{l1} &\sim B(1, p_{l1}),\ p_{l1} = \Phi(x_{l0}), & X_{l3} &\sim B(1, p_{l3}),\ p_{l3} = \Phi(x_{l0}), \end{aligned}$$

and

$$\begin{aligned} X_{l0} &\sim B(1, 0.5), & X_{l2} &\sim B(1, p_{l2}),\ p_{l2} = \Phi(x_{l1}), \\ X_{l1} &\sim B(1, p_{l1}),\ p_{l1} = \Phi(x_{l0}), & X_{l3} &\sim B(1, p_{l3}),\ p_{l3} = \Phi(x_{l2}) \end{aligned}$$

with again $l = 1, \ldots, n$ for varying sample sizes n of 100, 200, 500, and 1000 observations. In contrast to the marginal case, here it is not possible to sample data sets with interactions between the variables. The averaged results of the 20 runs for each combination are summarized in Table 2.

For a small number of $n = 100$ observations the algorithm is not able to find the underlying model irrespective from the kind of model selection (with or without interactions). Anyhow, no other model is clearly preferred, as the posterior probability of the best alternative model is not very high and does not show a great difference to the next best models either. This misbehaviour vanishes, however, with increasing sample size. This especially holds for the model selection that incorporates possible interactions.

In Figure 5, the adjacency and corresponding skeleton matrices of equivalence class 3 for 100 resp. 1000 observations are shown. We call S the skeleton matrix that belongs to the adjacency matrix A if

$$\overline{\bar{A}_{1a}} = \begin{pmatrix} 0 & .09(.11) & .09(.15) & .86(.70) \\ .08(.10) & 0 & .11(.15) & .86(.70) \\ .10(.14) & .11(.14) & 0 & .86(.70) \\ .14(.03) & .14(.03) & .14(.03) & 0 \end{pmatrix}$$

$$\overline{\bar{A}_{1b}} = \begin{pmatrix} 0 & .11(.15) & .09(.09) & .77(.93) \\ .10(.14) & 0 & .12(.13) & .72(.93) \\ .09(.10) & .14(.14) & 0 & .77(.93) \\ .23(.07) & .28(.07) & .23(.07) & 0 \end{pmatrix}$$

Figure 4: The average over the 20 runs of the averaged adjacency matrix for the data type 1a (above) and the data type 1b (below) with 1000 observations. Model selection has been carried out accounting for interactions and not, the results of the latter are in brackets.

	search without ia		**search with ia**	
	$\bar{\hat{\boldsymbol{p}}}$ (r)	$\bar{\hat{\boldsymbol{p}}}_{\boldsymbol{a}}$	$\bar{\hat{\boldsymbol{p}}}$ (r)	$\bar{\hat{\boldsymbol{p}}}_{\boldsymbol{a}}$
	equivalence class 2			
$n = 100$	0.08 (4)	0.11	0.06 (4)	0.10
$n = 200$	0.16 (1)	0.16	0.17 (1)	0.13
$n = 500$	0.30 (1)	0.15	0.43 (1)	0.16
$n = 1000$	0.48 (1)	0.14	0.73 (1)	0.09
	equivalence class 3			
$n = 100$	–	0.05	0.03 (5)	0.08
$n = 200$	0.15 (1)	0.10	0.16 (1)	0.14
$n = 500$	0.27 (1)	0.11	0.45 (1)	0.09
$n = 1000$	0.43 (1)	0.13	0.72 (1)	0.09

Table 2: Simulation results for the equivalence classes 2 and 3; same notation as in Figure 1.

$[S]_{i,j} = [A]_{ij} + [A]_{ji}$ for $i, j = 1, \ldots, p$. Thus, the skeleton matrix gives information about the general structure of the graph but disregards the orientation of the edges. It can be seen that also for the small sample size of $n = 100$ the edges of the true model are more often represented in the selected models than the others.

This is particularly observed for the selection strategy which allows for interactions. Otherwise, again the occurrence of wrong edges is higher. This disadvantage disappears for a simple size of $n = 1000$ observations, where the probability of wrong edges is obviously lower. The correct edges are always found with a higher probability in all selected models for both selection strategies. Summarizing, it can be stated that the selection strategy performs in a satisfying way. It cannot be concluded that regarding or disregarding the interaction results in a better performance, though there seems to be a tendency to favor the former. In addition to the posterior probabilities of the selected models, information can be gained by the adjacency matrices which is of special importance if no model is clearly preferred.

$$\overline{\bar{A}_3} = \begin{pmatrix} 0 & .51(.54) & .20(.27) & .20(.24) \\ .49(.45) & 0 & .36(.41) & .20(.24) \\ .19(.20) & .32(.38) & 0 & .39(.41) \\ .25(.22) & .22(.28) & .46(.48) & 0 \end{pmatrix}$$

$$\overline{\bar{S}_3} = \begin{pmatrix} 0 & .99(.99) & .39(.47) & .45(.47) \\ \star & 0 & .68(.79) & .42(.52) \\ \star & \star & 0 & .85(.89) \\ \star & \star & \star & 0 \end{pmatrix}$$

$$\overline{\bar{A}_3} = \begin{pmatrix} 0 & .38(.29) & .13(.06) & .09(.03) \\ .62(.71) & 0 & .48(.47) & .12(.05) \\ .13(.07) & .52(.53) & 0 & .65(.74) \\ .08(.03) & .11(.05) & .35(.26) & 0 \end{pmatrix}$$

$$\overline{\bar{S}_3} = \begin{pmatrix} 0 & 1(1) & .26(.12) & .17(.07) \\ \star & 0 & 1(1) & .24(.10) \\ \star & \star & 0 & 1(1) \\ \star & \star & \star & 0 \end{pmatrix}$$

Figure 5: Average of the averaged adjacency (A) and skeleton (S) matrices of the 20 runs of equivalence class 3 with 100 (above) and 1000 (below) observations. The results outside (inside) the brackets are obtained from a model selection without (with) interactions.

4 Algorithm for the Mixed Case

We now briefly address the mixed case, where we assume the considered continuous and binary variables to follow a conditional Gaussian (CG) distribution (cf. Tate, 1954; Dempster, 1973; Lauritzen, 1996; Lauritzen and Wermuth, 1989). For our purpose, we assume the mixed vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ to follow a homogeneous CG distribution, which implies that we do not allow for squared interactions among the continuous variables. Based on the factorisation property $f(\boldsymbol{x}) = \prod_{l=i}^{p} f(x_i \mid \boldsymbol{x}_{pa(i)})$ the estimation problem can again be reduced to the univariate distributions $f(x_i \mid \boldsymbol{x}_{pa(i)})$, which are now CG regressions with either a continuous or binary response variable. These distributions can be represented by a normal regression resp. a probit model with mixed covariables. In the case of a binary response variable we allow for pairwise discrete or mixed interactions. This leads to the following algorithm for the discrete case:

Reversible Jump Algorithm for the Mixed Case:

1. For all variables X_i, $i = 1, \ldots, p$,

   If X_i is discrete,

   For all observations X_{ki}, $k = 1, \ldots, n$,

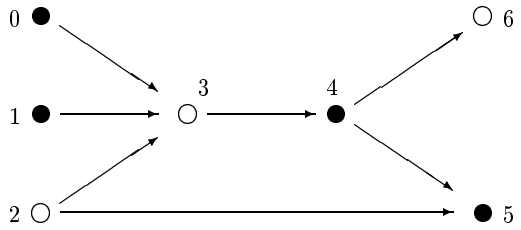

Figure 6: Equivalence class 4: The open circles denote continuous, the black ones binary variables.

draw utility Z_{ki} from full conditional $Z_{ki} \mid x_{ki}, \boldsymbol{x}_{kpa(i)}\boldsymbol{\beta}_{i|pa(i)}$

2. Update d, i.e. cancel, add, or switch the directed edge $X_j \to X_i$; distinguish
   - Response X_i is continuous:
     (a) Take the algorithm for the Gaussian case, where now the covariables $pa(X_i)$ can be continuous or binary
   - Response X_i is discrete
     (a) Replace binary response X_i by continuous utility Z_i
     (b) Consider the new or vanishing interactions among the parents of X_i and possibly X_j, that can be pairwise discrete or mixed
     (c) Carry out birth, death or switch step
3. Update $\boldsymbol{\beta}_{i|pa(i)}$, $i = 1, \ldots, p$
4. Update $\sigma^2_{i|pa(i)}$, $i = 1, \ldots, p$

The performance of the algorithm is again tested by a simulation study. Based on the dag of Figure 6 as underlying graph, we sample twenty data sets with seven variables and $n = 100, 200, 500, 1000$ observations. They are generated by the sampling scheme

$$\begin{aligned}
X_{i0} &\sim B(1, 0.5); \\
X_{i1} &\sim B(1, 0.5); \\
X_{i2} &\sim \mathrm{N}(0, 1); \\
X_{i3} &\sim \mathrm{N}(X_{i0} + X_{i1} + X_{i2}, 1); \\
X_{l4} &\sim B(1, p_{l4}), \quad p_{l4} = \Phi(X_{l3}); \\
X_{l5} &\sim B(1, p_{l5}), \quad p_{l5} = \Phi(X_{l2} + X_{l4}); \\
X_{l6} &\sim \mathrm{N}(X_{l4}, 1).
\end{aligned}$$

We denote this data type which contains no interactions by 4a and the alternative by 4b which differs in the generation of X_{l5} by

$$X_{l5} \sim B(1, p_{l5}), \qquad p_{l5} = \Phi(X_{l2} + X_{l4} + X_{l2}X_{l4})$$

	data typ 4a (without ia)		**data typ 4b** (with ia)	
	$\hat{p}$ (r)	$\hat{p}_a$	$\hat{p}$ (r)	$\hat{p}_a$
$n = 100$	–	0.001	–	0.001
$n = 200$	0.005 (1)	0.002	–	0.005
$n = 500$	0.051 (1)	0.023	0.107 (1)	0.033
$n = 1000$	0.194 (1)	0.054	0.194 (1)	0.050

Table 3: Simulation results for equivalence class 4, the results are averaged over the simulation runs of the 20 data sets of each type.

which thus implies an interaction. The model search is only performed by taking interactions into account. Otherwise, too many restrictions would be imposed on the CG distribution. The satisfying results of the study are given in Table 3.

With seven variables involved in a data set, a sample size of $n = 100$ or 200 observations is obviously too small for the algorithm to prefer one particular dag. For $n = 500$ the underlying model is more or less clearly detected with a posterior probability of 0.051 resp. 0.107. It is then definitely found for a sample size of $n = 1000$. Like in the pure discrete case, even for a small number of observations the general structure of the graph is reflected by the corresponding adjacency and skeleton matrices. We refrain from showing these situations here.

5 Example: Women and Mathematics

The data set of interest here, which became quite famous under the name "Women and Mathematics", stems from a survey by Lacampagne (1979). The aim was to analyze the success of a special "women and mathematics" lecture. For this purpose, 1190 students at eight high schools were asked some demographic variables, their attitude towards mathematics and if they had taken part in the program. Table 4 shows the six variables to be investigated. The data has already been analyzed by several authors as for instance Upton (1991), Madigan and Raftery (1994), and Giudici et al. (1999). We will compare our results with those of Giudici et al. (1999) who used a reversible jump algorithm, too, but restricted their model search to undirected decomposable graphs. This class of graphical models can be regarded as a subset of the dags, namely those which do not contain immoralities.

var.	**question**	**answer**
X_0	WAM lecture attendance	yes/no
X_1	sex	male/female
X_2	school type	urban/suburban
X_3	"I need mathematics for my future"	agree/disagree
X_4	subject preference	math and science/liberal arts
X_5	future plans	college/job

Table 4: Variables of the women and mathematics data set.

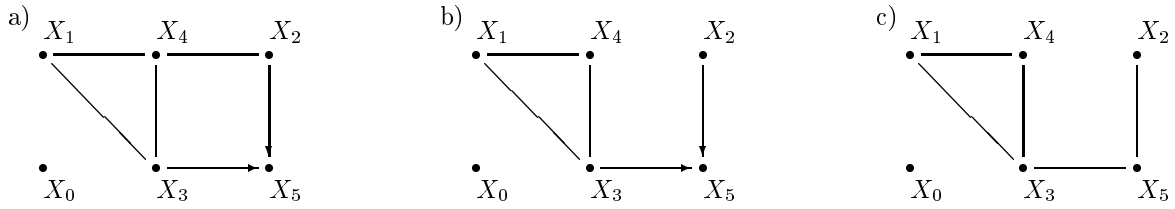

Figure 7: The essential graphs with the highest posterior probabilities of 0.20 (a) and 0.10 (b) by the reversible jump algorithm for dags and the undirected decomposable graph with the highest posterior probability (0.52) by the hierarchical reversible jump algorithm of Giudici et al. (c).

Our results are based on a runtime of 205000 iterations where the first 5000 are treated as burn–in and then every 20th is picked out to estimate the parameters. Furthermore, we use that type of switch step which only allows the movement into an equivalent model and accepts this with a probability of 0.5. To summarize the different dags of one equivalence class we present them by their essential graph. Figure 7 shows the two models which get the highest posterior probability with 0.20 and 0.10. The best model of Giudici et al. (1999), which is searched for in the space of undirected decomposable graphs, looks quite similar (see Figure 7) and is selected with a probability of 0.52. Both selection strategies detect an independence of the lecture attendance and the remaining variables. The variables X_1, X_3, and X_4 (alias sex, self–assessed importance, and preferences) form a clique. The main difference between the results of the two reversible jump approaches lies in the fact that our strategy supports an independence of X_2 and X_3 (importance and school type) given X_4 (preferences) which vanishes conditioning on X_5 (future plans). This is indicated by the immorality $X_3 \rightarrow X_5 \leftarrow X_2$, which can also be seen in the averaged adjacency matrix of Figure 8. As a marginal independence between two variables that vanishes by conditioning on a third cannot be represented by undirected graphs, the model selection of Giudici et al. (1999) just detects a conditional independence.
The general association structure becomes quite clear by the skeleton matrix (see also Figure 8) where those vertices that are connected in Figure 7b) and c) are linked by an edge with a probability of 1. In our approach, an edge between X_2 (school type) and X_4 (preferences) occurs in nearly half of the cases, which is again in constrast to the results of Giudici et al. (1999). More information on the substantial background is needed to decide on the model which describes reality better.

6 Conclusion

We introduced a reversible jump algorithm for model selection in the space of all dags for binary variables extending an algorithm already known for the continuous case. The extension to the mixed case as a combination of the continuous and the binary proceeding is then straightforward. For both cases, simulation studies have been carried out to analyze the performance of the algorithm. The results turn out to be satisfying. The example of the women and mathematics data set shows that also for real data sensible results are obtained.

$$\bar{A}_{wam} = \begin{pmatrix} 0 & .02 & .03 & .03 & .03 & .01 \\ .04 & 0 & .02 & .40 & .36 & .00 \\ .05 & .02 & 0 & .06 & .13 & .77 \\ .07 & .60 & .22 & 0 & .45 & .82 \\ .05 & .64 & .33 & .55 & 0 & .21 \\ .07 & .00 & .23 & .18 & .11 & 0 \end{pmatrix}$$

$$\bar{S}_{wam} = \begin{pmatrix} \star & .06 & 0 & .09 & .08 & .09 \\ \star & \star & .04 & \mathbf{1} & \mathbf{1} & .01 \\ \star & \star & \star & .27 & .46 & \mathbf{1} \\ \star & \star & \star & \star & \mathbf{1} & \mathbf{1} \\ \star & \star & \star & \star & \star & .32 \\ \star & \star & \star & \star & \star & \star \end{pmatrix}$$

Figure 8: Averaged adjacency and skeleton matrix for the women and mathematics data set. In the latter the edges of the essential graphs in Figure 7 are printed in bold.

We have not yet investigated the behaviour of the algorithm in terms of convergence and mixture. This is a very crucial task where we refer for a discussion of the general problem and possible solution to Brooks and Giudici (1999) and Brooks et al. (2001). As already mentioned, another drawback of our approach lies in the fact that the search takes place in the space of all dags instead of restricting to the essential graphs. Thus, long run times occur due to the huge search space. But the continuing developing of faster machines will hopefully improve the performance. Nevertheless, for small data sets up to 10 variables helpful information about the association structure can currently be gained at the moment. Besides the selected model, insights are also obtained from the averaged adjacency matrix.

The algorithm is implemented in the software package *BayesX*, which contains several methods for Bayesian inference based on MCMC (see Lang and Brezger, 2001). It can be downloaded from `http://www.stat.uni-muenchen.de/~lang/bayesx/bayesx.html`.

There are many possible extensions of this algorithm. The flexible MCMC design offers the possibility to also allow for latent or missing variables. The former has been considered by Giudici and Stanghellini (1999) for undirected graphs; to get an idea of the latter see for example Schafer (1997). In many practical cases, some of the possible influences and dependencies are already known. Therefore, it would be useful to state edges as fixed during the model search which, of course, also implies a reduction in run time, since it enormously decreases the search space.

Acknowledgements

This research was supported by the German National Science Foundation, the Graduate College "Applied Algorithmic Mathematics" and the SFB 386. We thank Angelika Blauth and Iris Pigeot for helpfull comments and Stefan Lang for incorporating the reversible jump algorithm into the software package *BayesX*.

References

Albert, J. H. and S. Chib (1993). Bayesian Analysis of Binary Polychotomous Response Data. *Journal of the American Statistical Association 88*, 669–679.

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

Brooks, S. and P. Giudici (1999). Convergence Assessment for Reversible Jump MCMC Simulations. In J. Berger, J. M. Bernardo, A. P. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6*, pp. 733–742. Oxford University Press.

Brooks, S., P. Giudici, and A. Philippe (2001). Nonparametric Convergence Assessment for MCMC Model Selection. Paper submitted.

Dempster, A. P. (1973). Aspects of the Multinomial Logit Model. In P. R. Krishnaiah (Ed.), *Multivariate Analysis III*. Academic Press, New York.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2 ed.). Springer, New York.

Edwards, D. and T. Havránek (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika 72*, 339–351.

Edwards, D. and T. Havránek (1987). A Fast Model Selection Procedure for Large Families of Models. *Journal of the American Statistical Association 82*, 205 – 213.

Fronk, E.-M. and P. Giudici (2000). Markov Chain Monte Carlo Model Selection for DAG Models. Paper submitted.

Geiger, H. and D. Heckerman (1999). Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. Technical report, Microsoft Research.

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

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

Giudici, P. and E. Stanghellini (1999). Bayesian Inference for Graphical Factor Analysis Models. Paper submitted.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika 82*, 711–732.

Lacampagne, C. B. (1979). *An Evaluation of the Women and Mathematics (WAM) Program and Associated Sex–Related Differences in the Teaching, Learning and Counseling of Mathematics.* Ed. D. Thesis, Columbia University Teachers College, USA.

Lang, S. and A. Brezger (2001). Bayesian P-splines. SFB 386 Discussion Paper 236, University of Munich.

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

Lauritzen, S. L. and N. Wermuth (1989). Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *The Annals of Statistics 17*, 31–57.

Madigan, D. and A. E. Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the American Statistical Association 89*, 1535–1546.

Perlman, M. D. (2002). Graphical Model Search via Essential Graphs. *Contemporary Mathematics (to appear)*.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

Tate, R. F. (1954). Correlation Between a Discrete and Continuous Variable. *Annals of Mathematical Statistics 25*, 603–607.

Thulasiraman, K. and M. N. S. Swamy (1992). *Graphs: Theory and Algorithms*. Wiley, New York.

Upton, G. J. G. (1991). The Exploratory Analysis of Survey Data Using Log–Linear Models. *The Statistician 40*, 169–182.