

# The Posterior Probability of Bayes Nets with Strong Dependences

Gernot D. Kleiter

Institut für Psychologie, Universität Salzburg, Austria  
gernot.kleiter@sbg.ac.at

## Abstract

Stochastic independence is an idealized relationship located at one end of a continuum of values measuring degrees of dependence. Modeling real world systems, we are often not interested in the distinction between exact independence and any degree of dependence, but between weak ignorable and strong substantial dependence. Good models map significant deviance from independence and neglect approximate independence or dependence weaker than a noise threshold. This intuition is applied to learning the structure of Bayes nets from data.

We determine the conditional posterior probabilities of structures given that the degree of dependence at each of their nodes exceeds a critical noise level. Deviance from independence is measured by mutual information. Arc probabilities are determined by the amount of mutual information the neighbors contribute to a node, is greater than a critical minimum deviance from independence. A  $\chi^2$  approximation for the probability density function of mutual information is used. A large number of network structures in which the arc probabilities are evaluated, is generated by stochastic simulation. Finally, the probabilities of the whole graph structures are obtained by combining the individual arc probabilities with the number of possible construction sequences compatible with the partial ordering of the graph. While selecting models with large deviance from independence results in simple networks with few but important links, selecting models with small deviance results in highly connected networks containing also less important links.

## 1 Introduction

A Bayes net consists of two components, a qualitative and a quantitative one. The qualitative component represents dependence and independence between sets of variables by a directed acyclic graph (DAG). Its topology is visually represented in a diagram consisting of circles and arcs as shown in the examples in section 7 at the end of the paper. The quantitative component represents the numerical information about the conditional probability distributions by an associated set of tables. The structure of Bayes nets may be assessed by experts, extracted from data, or may be obtained from a combination of both. This paper considers data sourced learning. The extraction of a Bayes net from data is a statistical problem. Intuitively, we want to find those structures that, based on the observed data, are

well justified. For a review of the literature and tutorials on learning probabilistic networks see [6, 15, 16].

This paper presents a new method to extract structures from frequency data and to evaluate their probability. We select networks containing strong links and substantial deviance from independence. The critical level at which links are considered to be strong or weak, may vary and depend upon the actual problem at hand. The criterion works like a filter to select those links. While selection at large deviance from independence results in simple networks having few but important links, selection at small deviance from independence results in highly connected networks containing also less important links.

Several approaches have been described to identify Bayes nets or closely related structures from data (Table 1). *Descriptive methods* have been employed to *approximate* joint highly multidimensional probability distributions that are too complex to be processed or stored directly. An example of this approach is the seminal paper by Chow & Liu [10] in which the authors approximated a multidimensional probability distribution by a tree structure. They determined the mutual information at each of the branches of the tree and showed that a maximum likelihood estimator of the overall structure is obtained when the total sum of mutual information is maximized. Wong & Poon [47] considered a classification problem and showed that under certain assumptions the Chow & Liu criterion is equivalent to minimizing an upper bound of the Bayes error rate. Malvestuto [33] extended the Chow & Liu approach to decomposable models of a given complexity. He used a hill-climbing procedure for computing solutions that are locally optimal. An important relationship between the cross-entropy measure and the minimum description length (MDL) was described by Lam & Bacchus [31]. The coding length of a Bayes net is a monotonically increasing function of the Kullback-Leibler divergence and can be used to fit *simple* nets to the data. Jiroušek & Kleiter [21] have shown that the Lam & Bacchus algorithm does not fully exploit the information contained in the conditional probability distributions. To exploit all the information, they proposed to assign probabilities to cliques of a moral graph instead of nodes of an acyclic directed graph. Kjærulff [23] proposed to simplify and approximate Bayesian networks and related structures by the removal of weak dependences. Jensen & Anderson [20] developed a complementary method that annihilates small probabilities by setting small probabilities in the clique potential of a junction tree equal to zero predetermined threshold. They choose a threshold  $k$  such that the sum of the  $k$  smallest probabilities is less than a predetermined threshold. The intuition underlying our approach is related to these proposals as we also discard weak dependences and we also use a threshold below which we discard dependences.

In the descriptive approach the relative frequencies of the data are treated as if they were precisely known probabilities. This is different in the statistically oriented approaches, where a statistical model, a multinomial process, for example, is assumed to generate the data. The parameters of the model are inferred from the limited number of data actually observed. There are two different approaches to statistical inference, the *sampling theory* (frequentist) and the *Bayesian* (subjective) approach. While the sampling theory approach relies on the analysis of the sample space, the Bayesian approach is characterized by treating parameters as random variables and by introducing a probability distribution over the parameter space. In both approaches two different kinds of problems are analyzed, hypothesis testing and

Inference	Bayesian	Sampling theory
Interval judgment	IV. Posterior distribution of parameters	V. Confidence regions of parameters
Hypothesis testing	II. Bayesian significance test Bayes factor	III. Frequentist significance tests
Data description	I. Approximation of the joint distribution	

Table 1: Five approaches to investigate Bayes nets on the basis of sample data.

interval judgment. Hypothesis testing typically involves two hypotheses that are tested or compared and, furthermore, the losses induced by the selection of the right or the wrong hypothesis. Two hypotheses may also be compared by a *Bayes factor*. The Bayes factor is the ratio of posterior to prior odds. A “calibration” by linguistic labels such as “barely worth mentioning”, “positive”, or “strong” is sometimes proposed [37].

*Interval judgment* does not provide ready made decisions about the rejection or acceptance of hypotheses at a given significance level, but leaves the evaluation of the results of the statistical analysis to the user and the user’s judgment of the relevance with respect to the actual problem at hand. From a Bayesian point of view all the information in the data about the parameters of a model is contained in the posterior distribution. In the *distributional way of thinking* [25] probability distributions take on the role of a language to express partial knowledge in a flexible and coherent way. Probability intervals under the posterior distribution are an important criterion to evaluate the plausibility of conclusions.

The class of procedures applied most often in model selection is *frequentist significance testing*. Thus,  $\chi^2$  tests are usually performed to identify log-linear models or to test independence in contingency tables. Spirtes, Glymour, & Scheines, though, give the following critical comments to the use of significance testing in the context of independence structures: “The usual comforts of a statistical test are the significance level, which offers assurance as to the limiting frequency with which a true null hypothesis would erroneously be failed by the test, and the power against an alternative, which is a function of the limiting frequency with which a false null hypothesis would not be rejected when a specified alternative hypothesis is true. Except in very large samples, neither the significance level nor the power of tests used within the search algorithms to decide statistical dependence measures the long run frequency of anything interesting about the search.” ([44], p. 130/1) Moreover, significance testing in Bayes nets involves multiple tests and controversial P values raising additional problems [32]. A systematic reference investigating the role of significance testing in model selection within the domain of Bayes nets is not known to me, and the same holds for frequentist confidence regions.

The most prominent approach to the identification of Bayes nets is the Bayesian version of hypothesis testing introduced by Cooper & Herskovits [11] and further investigated by

Heckerman, Geiger, & Chickering [17] and Heckerman [15]. A tutorial is provided by Heckerman [16]. A treatment of Bayesian hypothesis testing from the perspective of Markov Chain Monte Carlo methods is given by Raftery [37]. As the method proposed in this paper is Bayesian, though not Bayesian hypothesis testing but Bayesian interval judgment, we give detailed description of the Cooper & Herskovits method to motivate our own treatment of model selection.

Consider a parameter space  $\Theta$  and subdivide it into two subsets,  $\theta_0$  and  $\theta_1$ , so that  $\theta_0$  (the null hypothesis) contains a particularly interesting set of values and  $\theta_1$  (the alternative hypothesis) is its complement,  $\theta_0 \cup \theta_1 = \Theta$ ,  $\theta_0 \cap \theta_1 = \emptyset$ . If the subset  $\theta_0$  consists of just one value,  $\theta_0$  is a *point null hypothesis* (simple hypothesis,  $H_0 : \theta = \theta_0$ ), otherwise it is a *composite hypothesis*. While a frequentist significance test decides upon the rejection of  $H_0$  on the basis of two error probabilities, the probability of rejecting  $H_0$  when in fact it is true (significance level  $\alpha$ , type I error) and the probability of accepting  $H_0$  when in fact it is false ( $\beta$ , type II error), a Bayesian test finds the posterior probability of  $\theta_0$  and  $\theta_1$ . The Bayesian version of hypothesis testing was introduced by Jeffreys [19]. In Bayesian statistics, though, hypothesis testing never played the dominant role it plays in the sampling theory approach.

The structure of a Bayes net may be stated as a null hypothesis and then be tested with a Bayesian significance test. The Bayesian significance test finds the posterior probability of the model. It is straightforward to extend the approach to more than two hypotheses and structures, respectively.

It is instructive to consider the most elementary case involving only two binary variables  $X$  and  $Y$ . If one denotes the probability of success in  $X$ ,  $Y$ , and in both  $X$  and  $Y$  as  $\theta_x$ ,  $\theta_y$ , and  $\theta_{xy}$ , independence would then be expressed by the null hypothesis  $H_0 : \theta_{xy} = \theta_x \theta_y$ . The null hypothesis specifies the set of all triples satisfying the product. Assuming that the joint prior distribution of the parameters is a Dirichlet distribution and that sampling is multinomial, then the joint posterior distribution (and also the two posterior marginals) follow a Dirichlet distribution.

Figure 1 represents the three parameters  $(\theta_x, \theta_y, \theta_{xy})$  on the axes of a unit cube. The interior of the cube contains a cloud of points that were obtained from 1,000 simulations of random triples drawn from a Dirichlet distribution. Also shown in the cube is the surface for which independence holds. The surface is the paraboloid defined by the equation  $\theta_{xy} = \theta_x \theta_y$ . It corresponds to the null hypothesis that  $X$  and  $Y$  are independent or, equivalently, to the empty graph. How do we find the posterior probability of the null hypothesis, that is, of independence? The integral over the densities contained on the surface of the paraboloid corresponds to the posterior density of the null hypothesis. The precise null hypothesis is of lower dimension than the full parameter space. As we are integrating over a surface, though, and not over a volume the integral and accordingly the probability of the null hypothesis is zero.

A different way to investigate independence is to introduce a metric that measures the distance of any point in the cube from the paraboloid. There are many proposals how to quantify deviance from independence. Mutual information is the quantity that has been used most often in Bayes nets [45]. In the present example containing two binary variables

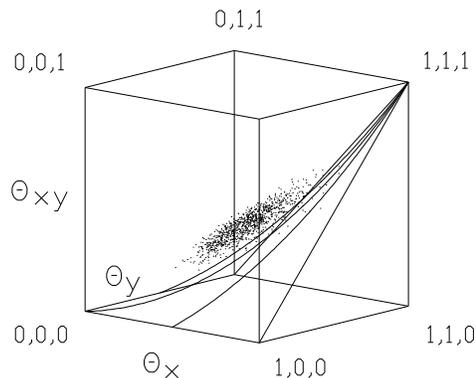


Figure 1: The cloud represents the probability density associated with the outcomes of two binary variables  $X$  and  $Y$ . The parameter space is defined by the unit cube of  $\theta_x, \theta_y$ , and  $\theta_{xy}$ . The independence relation between  $X$  and  $Y$  is represented by the paraboloid in the interior of the cube. The 1,000 points were obtained by random sampling from the Dirichlet distribution  $\text{Di}(15,5,15,5)$ .

mutual information is given by

$$\lambda = \theta_{xy} \log \frac{\theta_{xy}}{\theta_x \theta_y} + \theta_{x\bar{y}} \log \frac{\theta_{x\bar{y}}}{\theta_x \theta_{\bar{y}}} + \theta_{\bar{x}y} \log \frac{\theta_{\bar{x}y}}{\theta_{\bar{x}} \theta_y} + \theta_{\bar{x}\bar{y}} \log \frac{\theta_{\bar{x}\bar{y}}}{\theta_{\bar{x}} \theta_{\bar{y}}} .$$

The posterior probability density function of  $\lambda$  can be determined by stochastic simulation.

Now  $\lambda = 0$  is the null hypothesis that corresponds to perfect independence of  $X$  and  $Y$ . Again we run into the problem that the probability of the hypothesis is always zero. The marginal distribution of  $\lambda$  is continuous and “... a point null hypothesis  $H_0 : \theta = \theta_0$  cannot be tested under a *continuous* prior distribution.” ([38], p. 184). “In order to take seriously the problem of testing a point hypothesis, one must use a prior distribution in which  $Pr(\Theta = \theta_0) > 0$ .” ([40], p. 221) Jeffreys [19] proposed to assign a prior probability (a point mass) to the null hypothesis and distribute the remaining probability mass on the remaining parameter space. “Alternatively, one can replace the hypothesis with (what might be more reasonable) an interval hypothesis of the form  $H' : \Theta \in [\theta_0 - \epsilon, \theta_0 + \epsilon]$ ” ([40], p. 221) This is what we will do in this paper, except that we will consider the interval above a cutting score of  $\lambda$ .

We note that the prior probabilities assigned to  $H_0$  and to  $H_1$  in the Bayesian hypothesis testing approach constitute a second kind of prior probability in addition to the Dirichlet prior already introduced on the parameter space. One can easily extend the approach to more than two hypotheses. All possible structures for  $n$  nodes may be considered as the hypothesis space and prior probabilities may be introduced for all these structures accordingly.

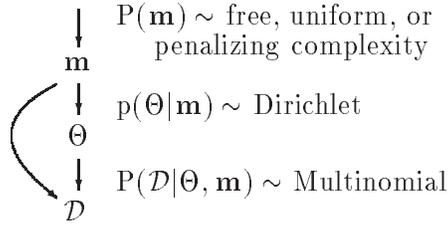


Figure 2: Cooper & Herskovits model,  $\mathbf{m}$  denotes the structural and  $\Theta$  the quantitative component of a Bayes net,  $\mathcal{D}$  the data.

Cooper & Herskovits [11] and Heckerman et al. [17] start from the following model

$$P(\mathbf{m}, \Theta, \mathcal{D}) = P(\mathbf{m})p(\Theta|\mathbf{m})P(\mathcal{D}|\Theta, \mathbf{m}). \quad (1)$$

Here  $\mathbf{m}$  denotes a hypothesis about the qualitative structure of a Bayes net,  $\Theta$  the quantitative part of the Bayes net, and  $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  the observed data. The hierarchical structure of the model is shown in Figure 2. Averaging out  $\Theta$  by integration leads to

$$P(\mathbf{m}, \mathcal{D}) = P(\mathbf{m}) \int_{\Theta} p(\Theta|\mathbf{m}) P(\mathcal{D}|\Theta, \mathbf{m}) d\Theta. \quad (2)$$

At each node  $X_i$  given its parents are instantiated at some configuration indexed by  $j$ , the conditional *data generating process* is assumed to be multinomial,

$$P(\mathcal{D}|\Theta, \mathbf{m}) \sim \text{Mu}(N_{ij}, \theta_{ij1}, \theta_{ij2}, \dots, \theta_{ijr_i}). \quad (3)$$

Let us first assume the hypothesis  $H_0 : \mathbf{m} = \mathbf{m}_0$  is true and that the *conditional prior* of the parameters (given the structure  $\mathbf{m}$  of the network) is Dirichlet

$$p(\Theta|\mathbf{m}) \sim \text{Di}(N'_{ij1}, N'_{ij2}, \dots, N'_{ijr_i}). \quad (4)$$

We note that the prior is different from the common analysis of parameters in contingency tables. Usually the joint prior distribution is assumed to be Dirichlet, here it is a conditional distribution. We denote the hypothetical sample sizes describing the prior distributions by  $N'_{ijk}$ , the observed sample size by  $N_{ijk}$ , and the posterior sample sizes accordingly by  $N''_{ijk} = N'_{ijk} + N_{ijk}$ , the *posterior distribution* is also Dirichlet

$$p(\Theta|\mathcal{D}, \mathbf{m}) \sim \text{Di}(N''_{ij1}, N''_{ij2}, \dots, N''_{ijr_i}). \quad (5)$$

where  $N''_{ij} = \sum_{k=1}^{r_j} N''_{ijk}$ .

Cooper & Herskovits now invoke the *predictive distribution* [1]. Denote by  $\mathbf{x}_{r+1}$  the configuration of the next case to be observed after having observed a sample of  $r$  previous cases. The probability of the next case to obtain a certain configuration given a sample of

previous cases (the parameters averaged out) is called a predictive probability. It is well known that the predictive distribution for a fixed number of further cases in the Dirichlet-multinomial model is a Dirichlet-multinomial distribution [1],  $\text{DiMu}(N''_{ij1}, N''_{ij2}, \dots, N''_{ijr_i})$ . Usually, the distribution is derived for several further observations. In the case of one further observation it is especially simple [24]. The probability for the next case falling into category  $k$  is equal to the relative frequency of this category (or more correctly in the present case, to  $N''_{ijk} / \sum_k N''_{ijk}$ ). Thus, to predict the first data vector  $\mathbf{x}_1$  when yet no previous data was observed, we make use of the mean of the prior distribution

$$P(\mathbf{x}_1 | \mathbf{m}) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{N'_{ijk}}{N'_{ij}}, \quad (6)$$

while for the case  $(l+1)$  the predictive probabilities are based on the means of the posterior distribution after  $l$  previous observations,

$$P(\mathbf{x}_{l+1} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l, \mathbf{m}) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{N''_{ijk}^{[l]}}{N''_{ij}^{[l]}}. \quad (7)$$

The double product results from the fact that we consider all nodes and all their associated parents. The probability of the sequence of *all* the data in the observed sample is just the product of the probabilities in Eq. 7 for the cases 1 to  $N$ . The probability of the data given  $\mathbf{m}$  is the *marginal likelihood function* of the network structures averaged over the model parameters. It is obtained from the product of the predictive probabilities of all the cases in the sample

$$P(\mathcal{D} | \mathbf{m}) = \prod_{l=1}^N P(\mathbf{x}_l | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{l-1}, \mathbf{m}). \quad (8)$$

The product can be rearranged algebraically so that the order of the observed data turns out to be irrelevant, and we have

$$P(\mathcal{D} | \mathbf{m}) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{? (N'_{ij})}{? (N''_{ij})} \prod_{k=1}^{r_i} \frac{? (N''_{ijk})}{? (N'_{ijk})}. \quad (9)$$

This is the *Cooper-Herskovits scoring function*.

The asymptotic approximation of the marginal likelihood as the number of observations  $N \rightarrow \infty$  has been shown by Bouckaert [4] to be

$$P(\mathcal{D} | \mathbf{m}) \approx H(\mathbf{m}, \mathcal{D})N - \frac{1}{2} \dim(\mathbf{m}) \log(N), \quad (10)$$

where  $\dim(\mathbf{m})$  is the number of parameters in  $\mathbf{m}$  and  $H(\mathbf{m}, \mathcal{D})$  is the entropy of the probability distribution obtained by projecting the frequencies of the observed cases into the conditional probability tables of  $\mathbf{m}$ . The asymptotic marginal likelihood function is called the *Bayesian Information Criterion* (BIC). It was first derived by Schwarz [41] as an asymptotic approximation of the posterior distribution. The criterion is equivalent to the

MDL. This is an important relationship showing that different approaches can converge to equivalent solutions.

When we introduce a *prior distribution*  $P(\mathbf{m})$  over the set of network structures  $\mathcal{M}$  (hypothesis space), then Bayes' Theorem leads to the *posterior distribution*

$$P(\mathbf{m}|\mathcal{D}) = \frac{P(\mathbf{m}) P(\mathcal{D}|\mathbf{m})}{\sum_{\mathbf{m} \in \mathcal{M}} P(\mathbf{m}) P(\mathcal{D}|\mathbf{m})}. \quad (11)$$

The posterior probability of a model  $\mathbf{m}$  can easily be determined when a simple null hypothesis is tested against a simple alternative hypothesis. This happens, for example, in the behavioral sciences when two competitive theories are tested against each other. There are many cases, though, in which we do not want to restrict our selection to a small set of models. When we know little about the domain under investigation we want the data to propose which structures are plausible and which ones are not. When we extend the set of hypotheses  $\mathcal{M}$  to include all possible Bayes net structures, though, the normalizing sum in the denominator of the posterior probability in Eq. 11 cannot be determined as the number of structures in  $\mathcal{M}$  is too large. The number of possible structures increases over-exponentially with the number of variables. Robinson [39] gives the following recursive function for the number of labelled acyclic directed graphs for  $n$  nodes

$$f(n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i), \quad f(0) = 1.$$

We obtain  $f(2) = 2$ ,  $f(3) = 25$ ,  $f(4) = 543$ ,  $f(5) = 29,281$ ,  $f(10) = 4,175,098,976,430,598,143$ . The motivation for the development of many methods in graphical modeling was to overcome the problems caused by the exponential growth of the number of parameters with the number of state variables. The growth of parameters, though, is much slower than the growth of network structures! In the case of ten binary variables, for example, we have “only”  $2^{10}=1,024$  parameters, but  $4.1 \times 10^{18}$  graph structures.

Thus, in problems with more than a few variables the posterior probability of the structures can only be evaluated up to a standardizing constant. With unstandardized probabilities *Bayes factors* can be obtained. Large Bayes factors, though, can result from the comparison of two structures both having small posterior probabilities. Likewise, the marginal likelihood function allows the selection of the “best” model, but not the evaluation how good the model really is. In addition, the asymptotic results raise the question whether we really want to assume that the sample size approaches infinity. Are we not moving in circles then and taking observed relative frequencies to be equal to probability parameters and thus eliminating the whole bussiness of inferential statistics? The equivalence of the BIC metric and the MDL criterion beautifully indicates that under the assumption of infinite sample sizes the inferential apparatus is not needed (we are back on the descriptive level in box I in Table 1).

A second difficulty enters when we want to introduce a prior distribution on the set of possible models. Which distribution should we choose? A uniform distribution giving each model the same prior probability is convenient but leads to problems. Lam & Bacchus raise the following point with respect to uniform priors on the set of all possible network

structures: “Unfortunately, it seems to us that this is the wrong choice. By choosing this prior their [Cooper & Herskovits’] method would prefer a more accurate network, even if that network is *much* more complex and only slightly more accurate. Given that we must perform learning with only a limited amount of data, this insistence on accuracy is questionable.” ([31], p. 273) De Groot (in the discussion of [42]) recommends: “when diffuse prior distributions are used in Bayesian inference, they must be used with care. Although they can serve as convenient and useful approximations in some estimation problems, they are never appropriate for tests of significance. Under no circumstances should they be regarded as representing ignorance.” Heckerman et al. [17] proposed a prior that penalizes the number of arcs in the graph. The proposal might be improved by characterizing the complexity of a graph by its MDL and by taking into account that many different structures are equivalent in term of probabilities. But do we need a prior distribution on the set of possible models at all? In the method proposed below we only need the prior on the parameter space but not on the model space, and this is also what is usually done in Bayesian analysis of contingency tables.

In Bayesian statistics a hypothesis is accepted if it obtains a high posterior probability. Because in the Cooper & Herskovits approach the prior probabilities are very small (and the sample size is not approaching infinity) the posterior probabilities may be very close to zero. Even the “best” model may obtain a probability that is extremely small. This poses a serious problem. We should not select a model obtaining an infinitesimally small posterior probability, and afterwards use this model to propagate probabilities in single case inferences. The probability of the selected model should be close to 1 and not close to 0. In probability propagation the structure and the estimated probabilities of the Bayes net are treated as if they were perfectly known and given as a fact. By doing this the uncertainty due to model selection is swept under the carpet. Similar points were raised by Madigan & Raftery [32]. They discard all models obtaining low Bayes factors and arrive at a much smaller set of models for the model selection process. We believe that *fully standardized posterior probabilities are indispensable for a proper judgment of the plausibility of a model.*

It is well known that many of the graphs contained in the set of possible graphs are probabilistically equivalent. To take a trivial example, for a given number of nodes all complete DAGs are factorizations of the same joint distribution. This property of Bayes nets has not been treated in a satisfactory way in model selection. Unfortunately, the method proposed below suffers from the same weakness.

We next give a short preview of the approach to be described below. We are often not interested in *precise* independence *versus* any degree of dependence but in *ignorable* dependence *versus relevant* dependence. We are interested in dependence that is larger than a theoretically or practically relevant value that depends on the purpose to be served by our analysis. If we express the strength of a dependence by an effect size, the effect size must be larger than a noise level that in the context of a given problem is only of little interest. The graph structure of a Bayes net can be represented in an adjacency matrix containing 0 and 1 values. The 1s indicate that an arc is present, the 0s that it is absent. When we extract a network from data we cannot be absolutely certain about the presence or absence of the arcs. The 0 and 1 values are uncertain and will be treated as Boolean random variables. With each arc we associate a probability  $P_{ij}$  that the indicator in the

adjacency matrix is 1 or, equivalently, that the associated arc in the graph is present. The way of modeling the problem is analogous to the Bayesian investigation of a set of binary variables.

We are then faced with two problems: (i) To obtain the individual arc probabilities  $P_{ij}$  from the data, and (ii) to obtain the probability of a whole graph from the individual arc probabilities. The first problem is tackled by calculating the probability that the strength of conditional dependence given the data in the neighborhood of a node is larger than a given effect size. The arc probabilities are set equal to the probability that the mutual information (Kullback-Leibler divergence)  $\lambda$  is larger than a minimum value  $\lambda_*$ . This analysis is associated with the quantitative part of the network and will be treated in section 3. The second problem is tackled by applying random graph methods. Here the probability of the whole graph structure is obtained from a uniform prior distribution on the set of all possible construction sequences of the graphs together with the individual arc probabilities. This problem will be treated in sections 4 and 5.

## 2 Basic Model

We consider a set  $V$  of  $n$  discrete variables  $(X_1, \dots, X_n)$ . Each variable is associated with a finite domain of  $k_i$  possible values  $Val(X_i)$ . The sample space  $\mathcal{X}$  is given by the Cartesian product of the domains of possible values,  $\mathcal{X} = \times_{i=1}^n Val(X_i)$ . Its cardinality is  $D = \prod_{i=1}^n k_i$ . If all variables are binary we have  $D = 2^n$ .  $D$  corresponds to the number of cells in a multiway contingency table. We observe a sample of  $N$  realizations  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,n})$ ,  $i = 1, \dots, N$ , of the variables  $(X_1, \dots, X_n)$ . We assume no observations are missing and that the realizations are generated independently with identical probabilities  $\theta_1, \dots, \theta_D$ . As a consequence, the cell frequencies  $r_1, \dots, r_d$ , where  $\sum_{i=1}^D r_i = N$  and  $d = D - 1$ , follow a multinomial distribution that is characterized by the  $d$  multinomial parameters  $\theta_1, \dots, \theta_d$ , where  $\sum_{i=1}^D \theta_i = 1$ ,

$$P(r_1, \dots, r_d | N, \theta_1, \dots, \theta_d) = \binom{N}{r_1, \dots, r_d} \theta_1^{r_1} \dots \theta_d^{r_d} (1 - \sum_1^d \theta_i)^{r_D}. \quad (12)$$

As a shorthand notation we write  $[r_1, \dots, r_d | N, \theta_1, \dots, \theta_d] \sim \text{Mu}(N, \theta_1, \dots, \theta_d)$ . Note that in the vector of probabilities we “lose” one dimension because the probabilities add up to 1 and the “last” value is not random but determined by the preceding  $d$  values. We therefore do not have  $D$  but  $d = D - 1$  random variables only. We assume the model does not contain any additional hidden variables.

To make inferences about the parameters  $\theta_1, \dots, \theta_d$  we introduce a prior distribution. It is convenient to work with a Dirichlet prior distribution. The multinomial parameters constitute a random vector on the simplex  $\mathcal{S}^d = \{(\theta_1, \dots, \theta_d) | \theta_1 > 0, \dots, \theta_d > 0; \sum_1^d \theta_i < 1\}$ . Let  $(\nu_1, \dots, \nu_D)$  be a vector of reals with  $(\nu_1 > 0, \dots, \nu_D > 0)$ . If the density of the vector is given by

$$p(\theta_1, \dots, \theta_d) = \frac{?(\nu_1 + \dots + \nu_D)}{?(\nu_1) \dots ?(\nu_D)} \theta_1^{\nu_1 - 1} \dots \theta_d^{\nu_d - 1} (1 - \sum_1^d \theta_i)^{\nu_D - 1} \quad (13)$$

the parameters follow a Dirichlet distribution. We write  $[\theta_1, \dots, \theta_d] \sim \text{Di}(\nu_1, \dots, \nu_D)$ . A uniform prior distribution on the simplex is described by the parameters  $\nu_1 = 1, \dots, \nu_D = 1$ . These parameters may be treated as hypothetical frequencies expressing the number of observations the prior knowledge is worth. Working with a Dirichlet *prior*, the *posterior* distribution of the multinomial parameters also follows a Dirichlet distribution. It is characterized by the sums of corresponding hypothetical and actual frequencies, that is by  $\alpha_1 = \nu_1 + r_1, \dots, \alpha_D = \nu_D + r_D$ , so that

$$[\theta_1, \dots, \theta_d | N, r_1, \dots, r_d] \sim \text{Di}(\alpha_1, \dots, \alpha_D). \quad (14)$$

If the joint distribution on the simplex is a Dirichlet distribution then all marginals and conditional distributions are Dirichlet distributions [46].

We work with a second order probability density function on the simplex of first order probabilities. The distribution tells us which values in the parameter space, given the observed data, are plausible and which ones are not. Because of the high number of dimensions, even for a moderate number of variables  $n$ , these considerations seem to be of theoretical interest only. In the following section, though, we show how to apply the “distributional” approach to mutual information, the quantity employed to measure the strength of dependence between variables.

### 3 Probability Distribution of Mutual Information

A widely used measure that expresses the strength of probabilistic dependence at the nodes of Bayes net is *mutual information* (Kullback-Leibler divergence). For an  $r \times s$  two-way contingency table (we will need only two-way tables) mutual information is defined as

$$\lambda = \sum_{i=1}^r \sum_{j=1}^s \theta_{ij} \log \frac{\theta_{ij}}{\theta_{i+} \theta_{+j}}, \quad (15)$$

where  $\theta_{i+}$  and  $\theta_{+j}$  denote the marginal probabilities obtained by summing over the indices replaced by the “+” signs.

If the probabilities on the right hand side are not perfectly known, the measure on the left hand side cannot perfectly be known either. In our model  $\theta_{ij}$ ,  $\theta_{i+}$ , and  $\theta_{+j}$  follow Dirichlet distributions. In [29] we proposed a  $\chi^2$  approximation for the distribution of  $\lambda$  that is slightly improved here. It is well known [9] that the relative entropy distance can be approximated by  $\chi^2$ . This leads to the mean of the distribution. The variance was obtained by heuristic numerical methods.

Let  $X_1$  and  $X_2$  be two discrete random variables with  $r$  and  $s$  possible values,  $r \leq s$ ,  $D = r \times s$ , and  $d = D - 1$ . If the joint distribution of their probabilities in the  $r \times s$  table is Dirichlet,  $[\theta_{1,1}, \theta_{1,2}, \dots, \theta_{r,s-1}] \sim \text{Di}(\alpha, \alpha_{1,1}, \alpha_{1,2}, \dots, \alpha_{r,s-1})$ , then the distribution of mutual information  $\lambda$  is approximately  $\chi^2(I + (s - 1)/2, 1/\alpha)$  distributed, where

$$I = \alpha \log(\alpha) + \sum_i \sum_j \alpha_{ij} \log(\alpha_{ij}) - \sum_i \alpha_{i+} \log(\alpha_{i+}) - \sum_j \alpha_{+j} \log(\alpha_{+j}). \quad (16)$$

and  $\alpha = \sum_{i=1}^r \sum_{j=1}^s \alpha_{i,j}$ . A random variable  $X$  is  $\chi^2$  distributed with  $\nu$  degrees of freedom and scaling factor  $\omega$ ,  $[X] \sim \chi^2(\nu, \omega)$  for short, if its probability density function is

$$P(x|\nu, \omega) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \frac{x^{\nu/2-1}}{\omega^{\nu/2}} \exp\left(-\frac{x}{2\omega}\right), 0 \leq x < \infty, \nu > 0, \omega > 0. \quad (17)$$

The quantity  $I$  is well known in the statistical literature from testing the goodness of fit. The mean and the variance of the distribution are

$$\mu = \omega\nu = \frac{I + (s-1)/2}{\alpha} \quad \text{and} \quad \sigma^2 = 2\omega^2\nu = 2\frac{I + (s-1)/2}{\alpha^2}. \quad (18)$$

If in the sample  $X_1$  and  $X_2$  are independent,  $I$  is zero and the posterior distribution is  $\chi^2$  with  $(s-1)/2$  degrees of freedom and the variance approaching zero as the sample size  $\alpha \rightarrow \infty$ . Usually, the degrees of freedom in  $\chi^2$  applications in contingency tables are related to the number of cells. Note that in the present context they are related to the strength of the dependence. The scaling factor  $1/\alpha$  expresses the accuracy of the dependence. This implies that for large  $\alpha$  the mean of the distribution is close to that value that is obtained when the mutual information is directly estimated from the relative frequencies.

The approximation does not only apply to two variables but also to the situation where we have one focus variable  $X_i$  and a set of  $c$  conditioning variables  $X_C = \{X_1, \dots, X_c\}$ . The conditioning variables enter the analysis in the form of their  $c$ -fold Cartesian product. In case of binary variables we obtain a  $2 \times 2^c$  table. It is analyzed in the same way as a two variable problem in which the first variable has two values and the second  $2^c$  values.

With the help of the  $\chi^2$  distribution it is possible to state how sure we can be that the influence of a set of conditioning variables upon a dependent variable is larger than a given effect size. We take the area under the  $\chi^2$  distribution that is on the right hand side of the effect size  $\lambda_*$

$$P(\lambda > \lambda_*) = \int_{\lambda_*}^{\infty} f(\lambda|I, \alpha) d\lambda, \quad (19)$$

where  $X_i$  is the dependent variable and  $X_C$  is the set of conditioning variables. We can now express how sure we are that the dependence between a set of conditioning neighbors and a dependent focus node is greater than a given effect size. In the stochastic simulation process described below, we will credit this probability temporarily to all arcs in the neighborhood of the focus node. Before we describe this process in more detail we will discuss the function of arc probabilities in Bayesian networks.

## 4 Random Directed Acyclic Graphs

While several different kinds of random graphs are distinguished in the literature [3, 22] we consider only random graphs in which the nodes are fixed (non-random) and the edges are random. Usually, random graphs are introduced with undirected graphs. Let  $V$  be a set of  $n$  (finite labelled) nodes,  $E$  a set of  $m$  undirected edges (pairs of nodes in  $V$ ), and  $G(V, E)$  the associated graph of order  $n$  and degree  $m$ . When multiple edges and loops are not admitted, there are  $n(n-1)/2$  edges that can be drawn joining pairs of nodes. Let  $\mathcal{G}$

be the set of all  $2^{n(n-1)/2}$  possible undirected graphs on  $V$ . Let  $\mathbf{P}$  be an  $n \times n$  matrix of probabilities  $p_{ij}$ . An undirected *random graph*  $G(n, \mathbf{P})$  is a probability space on  $\mathcal{G}$ .

For a *directed* graph  $G$  without loops and multiple arcs we may, for example, specify that the probability of an arc from node  $j$  to node  $i$  is  $P[(i, j) \in G] = P_{ij}$ . We say that an edge  $(i, j)$  is *on* if  $(i, j) \in G$  and *off* otherwise. In the literature the  $P_{ij}$  are usually assumed to be equal for all edges and the *on/off* states of the edges are assumed to be independent. We will attach probabilities to the arcs of a Bayes net. The arcs will *not* have equal probabilities and the arcs will *not* be assumed to be independent (they will be assumed to be conditionally independent, though, given a total ordering of the nodes).

The probability to produce a whole graph from the individual arc probabilities is determined by a *graph generating model*. Before we turn to our graph generating model, we have to introduce some terminology.

The qualitative part of a Bayes net is a partial ordering. One specific partial ordering may arise from several total or linear orderings. Any *permutation* of the nodes of a Bayes net corresponds to a total ordering. With  $n$  elements there are  $(n!)$  permutations. Following Shafer [43] we call a permutation that is compatible with the partial ordering of the actual Bayes net a *construction sequence*. In parts of the literature a permutation is also conceived as a complete *transitive tournament*. A construction sequence is a *linear extension* of a partial ordering.

We use the following *graph generating model* for random DAGs:

1. Select one of the  $(n!)$  permutations with probability  $1/(n!)$ . A short algorithm generating random permutations with probability  $1/(n!)$  is given in [34]. The adjacency matrix of a graph that corresponds to a total ordering can always be arranged such that the lower triangular matrix contains 1s, and that the diagonal and the upper triangular matrix contain 0s. The lower triangular matrix has  $n(n-1)/2$  1s, each one corresponding to an arc. The associated DAG is completely connected.
2. For all arcs in the graph keep an arc switched *on* with probability  $P_{ij}$  or switch it *off* with probability  $1 - P_{ij}$ , respectively. Some entries in the lower triangular matrix will be set to 0. The entries in the upper triangular matrix do not change.

Step (1) guarantees that the directed random graph is acyclic. We note that once a permutation is selected in step (1) we assume conditional independence of the arcs in step (2). The equal probabilities in step (1) may be justified by the fact that in a Bayes net each total ordering represents a probabilistically equivalent factorization of the joint probability distribution. It is thus reasonable to assign equal selection probabilities to all orderings.

When we extract the structure of a Bayes net from data we of course do not know a total ordering which produced the dependence in our data. Thus, the extraction process proceeds in the opposite way. Up to now we can estimate the strength of certain dependence and derive arc probabilities by the use of the  $\chi^2$  approximation. We now therefore turn to the question of inferring the DAG probability from the arc probabilities.

Let us first consider the case of a *common arc probability*  $p_*$ . This case arises when we do not fix a minimum effect size  $\lambda_*$  but a minimum probability level  $p_*$  and let the effect sizes vary. The situation is in a way analogous to selecting links on the basis of a fixed

significance levels and attaching effect sizes to the links. Let the number of arcs in a model be denoted by  $\#(\mathbf{m})$ . Obviously the conditional probability distribution of the *number of arcs* (not taking the structure of the graph into account) given a construction sequence  $C$  is binomial,

$$P(\#(\mathbf{m}) = r|C, n, p_*) = \binom{n(n-1)/2}{r} p_*^r (1-p_*)^{n(n-1)/2-r}. \quad (20)$$

If the probability of each of the construction sequences is the same, namely  $1/(n!)$ , then the marginal distribution (with respect to  $C$ ) of the number of arcs  $P(\#(\mathbf{m}) = r|n, p_*)$  is also binomial. It is a probability mixture of binomial distributions with equal weights.

The probability of a model  $\mathbf{m}$  is obtained by the theorem of total probability, that is by summing up the conditional probabilities the specific structure obtains given its construction sequences,  $P(\mathbf{m}|C, n, p_*)$ , multiplied by the probability of the construction sequences,  $P(C)$ . Because of the common arc probability all graphs, having the same number of nodes  $r$ , have the same probability. To determine this probability we find the number of ways in which  $\mathbf{m}$  can be extended to a construction sequence  $C$ . If one denotes the number of such extensions by  $Q$ , then  $Q \times 1/(n!)$  is the probability of  $\mathbf{m}$  when every construction sequence has the same probability and the probability of  $\mathbf{m}$  in a graph generating model with a common probability is

$$P(\mathbf{m}|n, p_*) = Q \frac{1}{n!} p_*^r (1-p_*)^{n(n-1)/2-r}, \quad (21)$$

where  $r$  is the number of arcs in the specific graph and  $n(n-1)/2$  is the number of possible arcs in a construction sequence.

If the arcs have different probabilities the probability of the whole structure cannot be expressed by a binomial but by a product of the individual arc probabilities only. Before we turn to this case, though, we have to find  $Q$ , the number of construction sequences of a DAG.

## 5 The Number of Linear Extensions of a Partial Ordering

Using a more general terminology, a construction sequence is a total ordering, the qualitative structure of a Bayes net is a partially ordered set (poset), and the number of construction sequences of a given DAG is equivalent to the number of linear extensions of given partial ordering.

Determining the number of linear extensions of a partially ordered set is of considerable relevance. Theoretically it is fundamental to the theory of ordered sets and it is also of practical importance in such areas as computer sciences, for example, because of its close relationship with sorting problems. Brightwell & Winkler [5] have shown that the problem is #P-complete. Pruesse & Ruskey presented an algorithm that *generates* the linear extensions in time proportional to the number of linear extensions (constant amortized time). It would be most efficient to have an algorithm that produces successively linear extensions that differ only by a transposition of two of their elements. Unfortunately, as the authors show, this is not possible with all posets. To overcome this problem Pruesse & Ruskey generate

each linear extension twice and each extension is flagged by a plus or a minus sign. They represent the set of linear extensions by a transposition graph in which each linear extensions corresponds to one vertex and two vertices are adjacent whenever the corresponding linear extensions differ by a single transposition. Pruesse & Ruskey prove that there exists a Hamiltonian path through the transposition graph and this proof underlies their algorithm.

We have proposed an algorithm that determines the number of linear extensions without generating all the extensions [27, 28]. We used a coding scheme that in the generation process allows to jump over large subsets of linear extensions. As for the present purpose only the number of linear extensions is of interest and not the actual listing of all instances, this is an advantage.

The set of all possible orderings is equivalent to the set of all permutations of the first  $n$  natural numbers. The set builds a permutation group. Permutations can be coded in several ways. The coding by cycles is, for example, well known. We code a permutation by the number of inversions each of its elements introduces. An inversion is when a larger number precedes a smaller one. For each element in the permutation code we count the number of smaller elements on its right hand side. Thus, the permutation  $(1, 2, 3, 4)$  is coded by  $(0, 0, 0, 0)$ , the permutation  $(4, 1, 3, 2)$  by  $(3, 0, 1, 0)$ , or  $(3, 2, 4, 1)$  by  $(2, 1, 1, 0)$  etc.. The advantage of the inversion code is that each dimension starts at 0. The inversion numbers of the nodes  $X_1, X_2, \dots, X_n$  are denoted by  $(x_1, x_2, \dots, x_n)$ .

A nice way to visually represent the set of all permutations in their integer coding is to draw a permutohedron (or convexpolyhedron) [2]. Each permutation is represented by a point in the  $n$ -dimensional Euclidian space. The permutohedron is a structure on which a probability function can easily be defined. To find the cardinality of certain subsets though is not easy at all. For this purpose the handling of the dual space of inversion coded permutations seems easier to deal with.

The empty graph with  $n$  nodes and  $m = 0$  arcs has  $(n!)$  construction sequences or permutations. Using the inversion code of the permutations and denoting the inversion numbers by  $x_1, x_2, \dots, x_n$  we may obtain this result by the sum

$$Q = \sum_{x_1=0}^{n-1} \sum_{x_2=0}^{n-2} \dots \sum_{x_n=0}^0 1 = n \times (n-1) \times \dots \times 2 \times 1 = n!. \quad (22)$$

The sum counts all possible permutations in the permutohedron. This corresponds to the number of linear extensions of the empty graph. When the graph is not empty each arc that is introduced in the graph excludes a set of permutations in the permutohedron. We count the number of remaining permutations by the same sequence of sums but with adjusted lower and upper bounds. Each new arc narrows the range between the lower and the upper bounds of the sums. Unfortunately, the upper and lower bounds change dynamically as a function of some of the other bounds. Under certain conditions, though, the bounds of inner sums are constant and in these cases the counting process can replace summation by multiplication. Obviously, this speeds up the process dramatically. For more details the reader is referred to [27].

## 6 Stochastic Simulation

In a Bayes net the conditional probabilities of the states of node  $X_i$  depend on the states of its neighbors and its neighbors only. The neighbors build the *Markov blanket* of the focus node  $X_i$ . The Markov blanket  $mb(X_i)$  consists of the set of parents  $pa(X_i)$ , children  $ch(X_i)$ , and parents of the children  $pa(ch(X_i))$  of  $X_i$ . A variable occurs only once in the Markov blanket, though it simultaneously may be a parent of  $X_i$  and a parent of a child of  $X_i$ . Moreover, the focus node is not itself a member of the Markov blanket, though it clearly is a parent of its children. Hrycej [18] has shown that stochastic simulation of the Markov blankets of a Bayes net is a Gibbs sampler of the involved probability distributions. A similar technique was employed to propagate probabilities and their precisions in Bayes nets [26]. Before the simulation starts we fix the effect size  $\lambda_*$ , the number  $B$  of initial burn-in iterations, and the number  $T$  of iterations in the main stage.  $B = T = 1,000$  usually lead to results that change little when  $T$  is increased to 2,000 or 5,000. Here it is possible to insert causal zeros, that is arcs for which the direction is known perfectly. “High blood pressure”, for example, may be caused by “family history”, but not vice versa. The arc probabilities are initialized and may, for example, be set to  $2/n(n-1)$ .

In each iteration  $t$  we first generate a random DAG according to the graph generating model described in section 4. That is, we select a permutation corresponding to a completely connected DAG with probability  $1/(n!)$ , and then switch each of the arcs on with the current arc probability  $P_{ij}^{(t)}$ .

Next, for each node  $X_i$  we compute the posterior mutual information  $I_i''$  given the neighbors in its Markov blanket  $mb(X_i)$ . From the data we extract a two way table  $Z_i$  containing the number of cases falling into any of the combinations of possible states of  $X_i$  and the Cartesian product of possible states in the Markov blanket.  $Z_i$  is a temporary two-dimensional table. The rows correspond to the states of the focus node, and the columns to the possible states in the Cartesian product of the states in the Markov blanket. The number of columns is thus exponential in the number of nodes contained in the Markov blanket.

We now incorporate the prior knowledge about the probability parameters. We assume the prior Dirichlet distribution is assessed by the method of hypothetical sample sizes. An excellent introduction into the assessment of prior distributions by the method of hypothetical samples sizes is contained in [35]. It is convenient, but not necessary, to work with a uniform distribution containing 1s in each cell. To obtain a table of posterior coefficients  $Z_i''$  the hypothetical sample sizes in  $Z_i'$  are added to the frequency counts in  $Z_i$ .  $I_i''$  is determined according to equation 16. Next, we compute the posterior probability that the mutual information  $\lambda_i$  is larger than the effect size  $\lambda_*$  on the basis of the values contained in  $Z_i''$ :

$$P(\lambda_i > \lambda_* | N'', I_i'') \approx \int_{\lambda_*}^{\infty} \chi^2 \left( \frac{I_i'' + (s-1)}{2}, \frac{1}{N''} \right) d\lambda, \quad (23)$$

where  $N''$  is equal to the sum of the total actual and the total hypothetical sample sizes, that is,  $N'' = N + N'$ .

Temporarily all arcs in the neighborhood of  $X_i$  obtain the same probability. That is, the

same credit is given to all arcs that define  $mb(X_i)$ . At the beginning of the simulation, a series of burn-in iterations stabilize the process. During the main iterations,  $t = 1, 2, \dots, T$ , the number of times each arc is in the *on* state is counted and the current relative frequency at each arc is determined. Denoting the 0/1 state of an arc from  $X_j$  to  $X_i$  by  $A_{ij} \in \{0, 1\}$ , the current relative frequency determines the probabilities for an arc to be *on* in the next iteration:

$$P_{ij}^{(t+1)} = \frac{1}{t} \sum_1^t A_{ij}^{(t)}. \quad (24)$$

We label the nodes of the model  $\mathbf{m}$  such that all 1s corresponding to the presence of an arc in the model are contained in the lower triangular part of its adjacency matrix. Such a labeling corresponds a topological sorting. We next build a second lower triangular matrix  $\mathbf{P}^*$  for the associated probabilities of the presence or absence of the arcs.  $P_{ij}^{(T)}$  is put into the matrix if there is an arc from node  $j$  to node  $i$ , and  $1 - P_{ij}^{(T)}$  if there is no arc from  $j$  to  $i$ .

According to the graph generating model described in section 4 we obtain the conditional probability of  $\mathbf{m}$  given  $\lambda_*$  and  $\mathcal{D}$  by combining three terms. The first term is  $1/(n!)$  and corresponds to the probability of selecting a specific construction sequence  $C$ ,  $P(C) = 1/(n!)$ . The second term is  $Q$  and corresponds to the number of construction sequences compatible with  $\mathbf{m}$ .  $Q$  is the cardinality of the set  $\mathcal{C}$  of linear extensions (or construction sequences) of the partial ordering of  $\mathbf{m}$ . There are  $Q$  linear orderings that may have produced  $\mathbf{m}$ ,  $Q = |\mathcal{C}|$ . The third term gives the conditional probability of the pattern of arcs in  $\mathbf{m}$  given a construction sequence, the minimum deviance from independence  $\lambda_*$  and the data  $\mathcal{D}$ . According to the random graph model, the probability that an arc is selected is  $P_{ij}^{(T)}$  and the probability that an arc is not selected (dropped from the total ordering) is  $1 - P_{ij}^{(T)}$ , respectively. These probabilities depend on the data  $\mathcal{D}$  and the level of the minimum deviance from independence  $\lambda_*$ . Given a specific construction sequence  $C$  that is compatible with  $\mathbf{m}$  the overall probability to obtain the on/off pattern of the arcs in  $\mathbf{m}$  is just the product of the entries in the lower triangular matrix of  $\mathbf{P}^*$ , that is  $P(\mathbf{m}|C, \lambda_*, \mathcal{D}) = \prod_{i=2}^n \prod_{j=1}^i P_{ij}^*$ . To obtain the conditional probability of a model  $\mathbf{m}$  given  $\lambda_*$  and  $\mathcal{D}$  we combine the three terms by multiplying the selection probability  $1/(n!)$  and the conditional pattern probability  $\prod_{i=2}^n \prod_{j=1}^i P_{ij}^*$ , and build the sum over all possible linear orderings that may have produced the structure of  $\mathbf{m}$ . As there are  $Q$  such orderings this amounts to multiplying by  $Q$ .

To summarize, the conditional posterior probability of a model  $\mathbf{m}$  given the minimum deviance from independence  $\lambda_*$  and the data  $\mathcal{D}$  is

$$P(\mathbf{m}|\lambda_*, \mathcal{D}) = \sum_{C \in \mathcal{C}(\mathbf{m})} P(C)P(\mathbf{m}|C, \lambda_*, \mathcal{D}) \quad (25)$$

$$= Q \frac{1}{n!} \prod_{i=2}^n \prod_{j=1}^{i-1} P_{ij}^*, \quad (26)$$

where  $P_{ij}^* = P_{ij}^{(T)}$  if in the adjacency matrix  $A_{ij} = 1$  and  $P_{ij}^* = 1 - P_{ij}^{(T)}$  if  $A_{ij} = 0$ .

For large  $\lambda_*$  the empty graph having zero arcs obtains a probability close to 1. As  $\lambda_* \rightarrow 0$  the  $P_{ij}^{(T)}$  becomes 1 and each construction sequence obtains a probability close to  $1/(n!)$ .

If, on one hand, we look out for very strong effect sizes only, we can be almost sure that nearly no connections will pass the effect size filter, while, on the other hand, if we admit very weak effect sizes, practically every node is connected with every other one.

## 7 Examples

Recently, Edwards [12] p. 9, discussed the Florida murder data 1976-1977 originally published by Radelet (Table 2). There are three binary variables: the colors of victims (black, white), the color of murderers (black, white), and the sentences (death or other). Note that (perhaps contrary to prejudice) in the marginal cross classification *murderer*  $\times$  *sentence* the death sentence for white murderers (12.5 %) is slightly higher than that for black murderers (11.4 %).

	A		$\neg A$	
	B	$\neg B$	B	$\neg B$
C	6	0	11	19
$\neg C$	97	9	52	132

Table 2: Florida murder data 1976-1977;  $A$ =black victim,  $\neg A$ =white victim,  $B$  = black murderer,  $\neg B$ =white murderer,  $C$ =death sentence,  $\neg C$ =other sentence.

We analyzed the data at various  $\lambda_*$  levels. For each analysis 2,000 iterations were run. Probability estimates were obtained from the relative frequencies of the random graphs between iteration 1,000 and 2,000 (Table 3). Graphs with very small probabilities are not listed. For  $\lambda_* \rightarrow 0$  the process oscillates between the six possible complete graphs. Each clique has approximately the same probability. Intuitively, this is a consequence of admitting arcs that in the limit contribute zero mutual information. For large  $\lambda_*$  the empty graph obtains with probability close to 1. No arcs are selected at all because there are no frequency tables in the data resulting in probabilities higher than the critical  $\lambda_*$ . The interesting structures are found between these two extremes. For relatively small  $\lambda_*$  there are practically just two graphs obtaining probabilities .5 each. They differ only by the direction of the arcs between  $A$  and  $B$ . They are probabilistically equivalent and the random graphs oscillate between both structures. For a relatively large  $\lambda_*$  the empty graph obtains a considerable probability, about .55 for  $\lambda_* = .22$ . For effect sizes around  $\lambda_* = .17$  the limiting structure oscillates between two one-arc graphs. In both the sentence is independent of the color of the murderer and the victim. There is a strong relationship between the color of the murderers and victims. There is no graph with two arcs that obtains a reasonable posterior probability under any of the effect sizes.

For each graph the posterior probability may be plotted as a function of the effect size. The resulting curve can show steep rises and falls. The complete graph  $B \rightarrow A, C \rightarrow A, C \rightarrow B$ , for example, has the probabilities .167 at  $\lambda_* \rightarrow 0$ , .321 at  $\lambda_* = .01$ , .427 at  $\lambda_* = .02$ , .479 at  $\lambda_* = .07$ , .493 at  $\lambda_* = .12$ , .212 at  $\lambda_* = .13$ , .068 at  $\lambda_* = .14$ , and .018 at  $\lambda_*$

Arcs	Mutual Information					
	.02	.07	.12	.17	.22	.27
Empty graph	.000	.000	.000	.042	.549	.910
B→A,C→A, C→B	.427	.479	.493	.009	.000	.000
A→B,C→A, C→B	.422	.514	.495	.003	.000	.000
A→B,A→C	.017	.000	.003	.000	.000	.000
B→A	.000	.000	.000	.383	.196	.037
A→B	.000	.000	.000	.363	.202	.027
A→B,A→C	.000	.000	.000	.067	.005	.000
B→A,C→B	.000	.000	.000	.069	.002	.000

Table 3: Posterior probabilities for various graph structures and sizes of minimum mutual information for the Florida murder data 1976-1977; A = victim, B = murderer, C = sentence.

= .15.

We next analyze data from an investigation on coronary heart disease that has recently been also discussed by Edwards [12], p. 24 ff.. The investigation included six binary variables, Smoker (A, yes/no), Mental Work (B, strenuous yes/no), Physical Work (C, strenuous yes/no), Systolic Blood Pressure (D, less than 140 yes/no), Lipoprotein Ratio (E, ratio of beta to alpha lipoproteins less than 3), and Family History (F, of coronary heart disease). The sample consisted of 1841 cases. The  $2^6 = 64$  cell counts are

44, 40, 112, 67, 129, 145, 12, 23, 35, 12, 80, 33, 109,  
67, 7, 9, 23, 32, 70, 66, 50, 80, 7, 13, 24, 25, 73, 57,  
51, 63, 7, 16, 5, 7, 21, 9, 9, 17, 1, 4, 4, 3, 11, 8, 14, 17,  
5, 2, 7, 3, 14, 14, 9, 16, 2, 3, 4, 0, 13, 11, 5, 14, 4, 4.

The last index of the list  $ABCDEF$  changes fastest. We add 1 to each cell.

We have considerable prior knowledge about the causal direction of edges. Smoking does not cause mental or physical work, it has no influence on the family history of coronary diseases. The coronary variables have no influence on smoking behavior. We put the causal zeros in the adjacency matrix shown in Table 4.

For small effect sizes we obtain many highly connected networks, but all with small posterior probabilities. At  $\lambda_* = .075$  the two nine arc structures shown in Figure 3 have the probabilities .056 and .078, but there are many other structures having probabilities around .01. The two structures shown are probabilistically equivalent and it thus makes a lot of sense that they both obtain very similar posterior probabilities. As  $\lambda_*$  increases the two structures become more and more probable. At  $\lambda_* = .13$  they have probabilities about .5 each. At about  $\lambda_* = .1526$  two new structures containing five arcs only appear, both again probabilistically equivalent. They are extremely unstable, though, and from  $\lambda_* = .157$  on

		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
Smoker	<i>A</i>	*			0	0	
Mental Work	<i>B</i>		*		0	0	
Physical Work	<i>C</i>			*	0	0	
Systolic BP	<i>D</i>				*		
Lipoprotein Ratio	<i>E</i>					*	
Family History	<i>F</i>		0	0	0	0	*

Table 4: “Causal zeroes” in the adjacency matrix of the coronary heart disease example; the direction of the arcs is from columns to rows; thus, high systolic blood pressure (*D*) does not lead to a certain family history (*F*), but the family history may lead to high or low blood pressure.

only structures with the two arcs  $E \rightarrow D$  or  $D \rightarrow E$  arise. As  $\lambda_*$  further increases the empty graph gets more and more probable. For example, at  $\lambda_* = .2$  its probability is .667 and the probability for the two other structures is .169 and .164, respectively. For effect sizes larger than  $\lambda_* = .24$  the structure transmutes itself to a completely unconnected network.

## 8 Conclusions

We discussed a method to find the conditional posterior probabilities of Bayes net structures given minimum deviance from independence. We first employed statistical properties of mutual information to obtain arc probabilities given the data. Secondly, we used random graph methods and stochastic simulation methods to obtain the network probabilities.

The conditional posterior probability of a DAG given a minimum deviance from independence depends on two factors: the number of its linear extensions  $Q$  and the product of its arc probabilities. Highly specific structures containing many arcs have relatively low  $Q$  values. To obtain a high probability such structures must compensate the low  $Q$  values

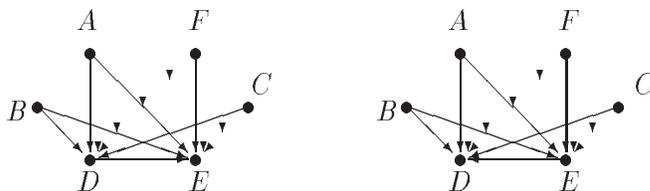


Figure 3: Low effect size  $\lambda_* = .075$  generating two highly connected and probabilistically equivalent Bayes net structures with posterior probability .056 (left) and .078 (right); only the arcs  $D \rightarrow E$  and  $E \rightarrow D$  are oscillating. At  $\lambda_* = .13$  both structures obtain the probabilities .528 and .472, respectively.

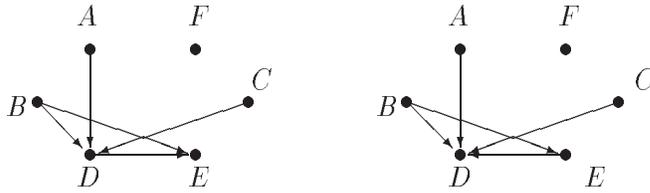


Figure 4: Two highly unstable probabilistic equivalent structures at effect size  $\lambda_* = .1526$  with posterior probability close to .5 each.

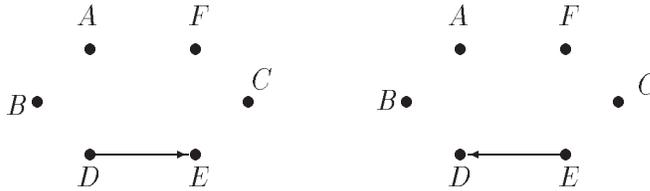


Figure 5: Effect sizes between  $\lambda_* = .16$  and  $\lambda_* = .19$  generate two probabilistically equivalent structures with one arc only. For  $\lambda_* > .2$  the posterior probability for the empty graph is larger than .7.

by high effect sizes. The balance between specificity on one hand and effect size on the other hand reflects two errors: overfitting the data on one hand and missing important relationships on the other hand.

We emphasize that there is usually not just one “best” or “optimal” structure (comparable to a point estimation) that can be extracted from the data. On the contrary, it is quite common that there are several alternative graph structures (probabilistically non-equivalent ones) that obtain similar support from the data. From the theory of random graphs it is known that the probabilities for structural properties of graphs (for example the number of components) is very sensitive to the arc probabilities. Small changes in the arc probabilities may lead to large changes in the probability of the according property. Especially in large networks the estimated structures may be “brittle”. The development of methods to investigate the behavior of graph estimates from data should help us to improve our judgment about the instruments we use in probabilistic reasoning.

## References

- [1] Aitchison, J. & Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.
- [2] Berge, C. (1971). *Principles of Combinatorics*. Academic Press: New York.

- [3] Bollobás, B. (1985). *Random Graphs*. Academic Press: New York.
- [4] Bouckaert, R. (1995). *Bayesian belief networks: From construction to inference*. PhD thesis, University of Utrecht.
- [5] Brightwell, G. & Winkler, P. (1991). Counting linear extensions. *Order*, **8**, 225-242.
- [6] Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions of Knowledge and Data Engineering*, **8**, 195–210.
- [7] Canfield, E. R. & Williamson, S. G. (1995). A loop-free algorithm for generating the linear extensions of a poset. *Order*, **12**, 57 – 75.
- [8] Chickering, D. M. & Heckerman, D. (1997, revised version). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Technical Report MSR-TR-96-08. Microsoft Research, Advanced Technology Division, Microsoft Corporation.
- [9] Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory*. New York: Wiley.
- [10] Chow, C. K. & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *14*, 462–467.
- [11] Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309-347.
- [12] Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer.
- [13] Golumbic, M. C. (1980). *Algorithmic Graph Theory and Perfect Graphs*. Boston: Academic Press.
- [14] Geiger, D., Heckerman, D. & Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. Technical Report MSR-TR-96-07, Microsoft Research, Advanced Technology Division, Microsoft Corporation.
- [15] Heckerman, D. (1996). Bayesian networks for knowledge discovery. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.), *Advances in Knowledge Discovery and Data Mining*. Menlo Park: AAAI Press/The MIT Press, 273–305.
- [16] Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (ed.), *Learning in Graphical Models*, pp. 301–354. Dordrecht: Kluwer.
- [17] Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning* **20**: 197-243.

- [18] Hrycej, T. (1990). Gibbs sampling in Bayesian networks. *Artificial Intelligence*, **46**, 351–363.
- [19] Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.
- [20] Jensen, F. & Andersen, S. K. (1990). Approximations in Bayesian belief universes for knowledge based systems. *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence*, Cambridge, MA.
- [21] Jiroušek, R. and Kleiter, G. D. (1995). A note on learning Bayesian networks. Proceedings of ECML-95, Kodratoff, Y. Nakhaeizadeh, G., and Taylor, C. (eds.) *Statistics, Machine Learning and Knowledge Discovery in Databases*, 148-153.
- [22] Karoński, M. (1995). Random graphs. In R. L. Graham, M. Grötschel, & L. Lovász, *Handbook of Combinatorics, Vol. 1*, Amsterdam: Elsevier, 351-380.
- [23] Kjærulff, U. (1994). Reduction of computational complexity in Bayesian networks through removal of weak dependences. Institute for Electronic Systems. Department of Mathematics and Computer Science, Aalborg, Denmark, R-94-2009.
- [24] Kleiter, G. D. (1980). *Bayes Statistik*. Berlin/New York: De Gruyter.
- [25] Kleiter, G. D. (1992). Bayesian diagnosis in expert systems. *Artificial Intelligence*, **54**, 1–32.
- [26] Kleiter, G. D. (1996). Propagating imprecise probabilities in Bayesian networks. *Artificial Intelligence*, **88**, 143 - 161.
- [27] Kleiter, G. D. (1997). The number of linear extensions of a directed acyclic graph. Institut für Psychologie, Universität Salzburg.
- [28] Kleiter, G. D. (1998). Structural uncertainty in Bayes nets. Institut für Psychologie, Universität Salzburg.
- [29] Kleiter, G. D. & Jiroušek, R. (1996). Learning Bayesian networks under the control of mutual information. In: S. Moral et al. (eds.), Sixth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System, Granada, 985-990.
- [30] Kolman, B. & Busby, R. (1987). *Discrete Mathematical Structures for Computer Science*. Englewood Cliffs: Prentice-Hall.
- [31] Lam, W. & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, **10**, 269–293.
- [32] Madigan, D. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, **89**, 1535-1546.

- [33] Malvestuto, F. M. (1991). Approximating discrete probability distributions with decomposable models. *IEEE Transactions on Systems, Man, and Cybernetics*, **21**, 1287–1294.
- [34] Nijenhuis, A. & Wilf, H. S. (1978). *Combinatorial Algorithms*. New York: Academic Press.
- [35] Novick, M. R. & Jackson, P. H. (1974). *Statistical Methods for Educational and Psychological Research*. New York: MacGraw-Hill.
- [36] Pruesse, G. & Ruskey, F. (1994). Generating linear extensions fast. *SIAM Journal of Computing*, **23**, 373-386.
- [37] Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 163–187. London: Chapman & Hall.
- [38] Robert, C. P. (1994). *The Bayesian Choice*. Berlin: Springer.
- [39] Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In C. H. C. Little (Ed.), *Combinatorial Mathematics V, Lecture Notes in Mathematics 622*, New York: Springer.
- [40] Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.
- [41] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- [42] Shafer, G. (1982). Lindley’s paradox. *Journal of the American Statistical Association*, **77**, 325–351.
- [43] Shafer, G. (1996). *Probabilistic Expert Systems*. Philadelphia: SIAM.
- [44] Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. New York/Berlin: Springer.
- [45] Studeny, M. & Vejnarova, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In M. I. Jordan (ed), *Learning in Graphical Models*, pp. 261–297. Dordrecht: Kluwer.
- [46] Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley
- [47] Wong, S. K. M. & Poon, F. C. S. (1989). Comments on approximating discrete probability distributions with dependence trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 333-335.