A Bayesian Approach to Causal Discovery

David Heckerman
Microsoft Research
Redmond, WA 98052
heckerma@microsoft.com

Christopher Meek
Microsoft Research
Redmond, WA 98052
meek@microsoft.com

Gregory Cooper
University of Pittsburgh
Pittsburgh, PA
gfc@smi.med.pitt.edu

February 1997

Technical Report
MSR-TR-97-05

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

We examine the Bayesian approach to the discovery of causal DAG models and compare it to the constraint-based approach. Both approaches rely on the Causal Markov condition, but the two differ significantly in theory and practice. An important difference between the approaches is that the constraint-based approach uses categorical information about conditional-independence constraints in the domain, whereas the Bayesian approach weighs the degree to which such constraints hold. As a result, the Bayesian approach has three distinct advantages over its constraint-based counterpart. One, conclusions derived from the Bayesian approach are not susceptible to incorrect categorical decisions about independence facts that can occur with data sets of finite size. Two, using the Bayesian approach, finer distinctions among model structures—both quantitative and qualitative—can be made. Three, information from several models can be combined to make better inferences and to better account for modeling uncertainty. In addition to describing the general Bayesian approach to causal discovery, we review approximation methods for missing data and hidden variables, and illustrate differences between the Bayesian and constraint-based methods using artificial and real examples.

1 Introduction

In this paper, we examine the Bayesian approach to the discovery of causal models in the family of directed acyclic graphs (DAGs). The Bayesian approach is related to the constraint-based approach, which is discussed in Chapters 1, 5, and 6 of this collection. In particular, both methods rely on the Causal Markov condition. Nonetheless, the two approaches differ significantly in theory and practice. An important difference between them is that the constraint-based approach uses categorical information about conditional-independence constraints in the domain, whereas the Bayesian approach weighs the degree to which such constraints hold. As a result, the Bayesian approach has three distinct advantages over its constraint-based counterpart. One, conclusions derived from the Bayesian approach are not susceptible to incorrect categorical decisions about independence facts that can occur with data sets of finite size. Two, using the Bayesian approach, finer distinctions among model structures—both quantitative and qualitative—can be made. Three, information from several models can be combined to make better inferences and to better account for modeling uncertainty.

In Sections 2 and 3, we review the Bayesian approach to model averaging and model selection and its application to the discovery of causal DAG models. In Section 4, we discuss methods for assigning priors to model structures and their parameters. In Section 5, we

compare the Bayesian and constraint-based methods for causal discovery for a small domain with complete data, highlighting some of the advantages of the Bayesian approach. In Section 6, we note computational difficulties associated with the Bayesian approach when data sets are incomplete—for example, when some variables are hidden—and discuss more efficient approximation methods including Monte-Carlo and asymptotic approximations. In Section 7, we illustrate the Bayesian approach on the data set of Sewall and Shah (1968) concerning the college plans of high-school students. Using this example, we show that the Bayesian approach can make finer distinctions among model structures than can the constraint-based approach.

2 The Bayesian Approach

In a constraint-based approach to the discovery of causal DAG models, we use data to make *categorical* decisions about whether or not particular conditional-independence constraints hold. We then piece these decisions together by looking for those sets of causal structures that are consistent with the constraints. To do so, we use the Causal Markov condition (Spirtes et. al, 1993) to link lack of cause with conditional independence.

In the Bayesian approach, we also use the Causal Markov condition to look for structures that fit conditional-independence constraints. In contrast to constraint-based methods, however, we use data to make *probabilistic* inferences about conditional-independence constraints. For example, rather than conclude categorically that, given data, variables X and Y are independent, we conclude that these variables are independent with some probability. This probability encodes our uncertainty about the presence or absence of independence. Furthermore, because the Bayesian approach uses a probabilistic framework, we no longer need to make decisions about individual independence facts. Rather, we compute the probability that the independencies associated with an entire causal structure are true. Then, using such probabilities, we can average a particular hypothesis of interest—such as, "Does X cause Y?"—over all possible causal structures.

Let us examine the Bayesian approach in some detail. Suppose our problem domain consists of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. In addition, suppose that we have some data $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, which is a random sample from some unknown probability distribution for $\mathbf{X}$. For the moment, we assume that each case $\mathbf{x}$ in D consists of an observation of all the variables in $\mathbf{X}$. We assume that the unknown probability distribution can be encoded by some causal model with structure $\mathbf{m}$. As in Spirtes et al. (1993), we assume that the structure of this causal model is a DAG that encodes conditional independencies via the Causal Markov condition. We are uncertain about the structure and parameters of the

model; and—using the Bayesian approach—we encode this uncertainty using probability. In particular, we define a discrete variable $\mathbf{M}$ whose states $\mathbf{m}$ correspond to the possible true models, and encode our uncertainty about $\mathbf{M}$ with the probability distribution $p(\mathbf{m})$. In addition, for each model structure $\mathbf{m}$, we define a continuous vector-valued variable Θ_m, whose values $\boldsymbol{\theta}_m$ correspond to the possible true parameters. We encode our uncertainty about Θ_m using the (smooth) probability density function $p(\boldsymbol{\theta}_m|\mathbf{m})$. The assumption that $p(\boldsymbol{\theta}_m|\mathbf{m})$ is a probability density function entails the assumption of faithfulness employed in constraint-based methods for causal discovery (Meek, 1995).

Given random sample D, we compute the posterior distributions for each $\mathbf{m}$ and $\boldsymbol{\theta}_m$ using Bayes' rule:

$$p(\mathbf{m}|D) = \frac{p(\mathbf{m})p(D|\mathbf{m})}{\sum_{m'} p(\mathbf{m}')p(D|\mathbf{m}')} \tag{1}$$

$$p(\boldsymbol{\theta}_m|D,\mathbf{m}) = \frac{p(\boldsymbol{\theta}_m|\mathbf{m})p(D|\boldsymbol{\theta}_m,\mathbf{m})}{p(D|\mathbf{m})} \tag{2}$$

where

$$p(D|\mathbf{m}) = \int p(D|\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|\mathbf{m})\ d\boldsymbol{\theta}_m \tag{3}$$

is called the *marginal likelihood*. Given some hypothesis of interest, h, we determine the probability that h is true given data D by averaging over all possible models and their parameters:

$$p(h|D) = \sum_m p(\mathbf{m}|D)p(h|D,\mathbf{m}) \tag{4}$$

$$p(h|D,\mathbf{m}) = \int p(h|\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|D,\mathbf{m})d\boldsymbol{\theta}_m \tag{5}$$

For example, h may be the event that the next case $\mathbf{X}_{N+1}$ is observed in configuration $\mathbf{x}_{N+1}$. In this situation, we obtain

$$p(\mathbf{x}_{N+1}|D) = \sum_m p(\mathbf{m}|D) \int p(\mathbf{x}_{N+1}|\boldsymbol{\theta}_m,\mathbf{m})\ p(\boldsymbol{\theta}_m|D,\mathbf{m})d\boldsymbol{\theta}_m \tag{6}$$

where $p(\mathbf{x}_{N+1}|\boldsymbol{\theta}_m,\mathbf{m})$ is the likelihood for the model. As another example, h may be the hypothesis that "X causes Y". We consider such a situation in detail in Section 5.

Under certain assumptions, these computations can be done efficiently and in closed form. One assumption is that the likelihood term $p(x|\boldsymbol{\theta}_m,\mathbf{m})$ factors as follows:

$$p(\mathbf{x}|\boldsymbol{\theta}_m,\mathbf{m}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i,\boldsymbol{\theta}_i,\mathbf{m}) \tag{7}$$

where each *local likelihood* $p(x_i|\mathbf{pa}_i,\boldsymbol{\theta}_i,\mathbf{m})$ is in the exponential family. In this expression, $\mathbf{pa}_i$ denotes the configuration of the variables corresponding to parents of node x_i, and $\boldsymbol{\theta}_i$

denotes the set of parameters associated with the local likelihood for variable x_i. One example of such a factorization occurs when each variable $X_i \in \mathbf{X}$ is discrete, having r_i possible values $x_i^1, \ldots, x_i^{r_i}$, and each local likelihood is a collection of multinomial distributions, one distribution for each configuration of $\mathbf{Pa}_i$—that is,

$$p(x_i^k | \mathbf{pa}_i^j, \boldsymbol{\theta}_i, \mathbf{m}) = \theta_{ijk} > 0 \tag{8}$$

where $\mathbf{pa}_i^1, \ldots, \mathbf{pa}_i^{q_i}$ ($q_i = \prod_{X_i \in \mathbf{Pa}_i} r_i$) denote the configurations of $\mathbf{Pa}_i$, and $\boldsymbol{\theta}_i = ((\theta_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$ are the parameters. The parameter θ_{ij1} is given by $1 - \sum_{k=2}^{r_i} \theta_{ijk}$. We shall use this example to illustrate many of the concepts in this paper. For convenience, we define the vector of parameters

$$\boldsymbol{\theta}_{ij} = (\theta_{ij2}, \ldots, \theta_{ijr_i})$$

for all i and j. A second assumption for efficient computation is that the parameters are mutually independent. For example, given the discrete-multinomial likelihoods, we assume that the parameter vectors $\boldsymbol{\theta}_{ij}$ are mutually independent.

Let us examine the consequences of these assumptions for our multinomial example. Given a random sample D that contains no missing observations, the parameters remain independent:

$$p(\boldsymbol{\theta}_m | D, \mathbf{m}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij} | D, \mathbf{m}) \tag{9}$$

Thus, we can update each vector of parameters $\boldsymbol{\theta}_{ij}$ independently. Assuming each vector $\boldsymbol{\theta}_{ij}$ has a conjugate prior[1]—namely, a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\theta}_{ij} | \alpha_{ij1}, \ldots, \alpha_{ijr_i})$—we obtain the posterior distribution for the parameters

$$p(\boldsymbol{\theta}_{ij} | D, \mathbf{m}) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} | \alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i}) \tag{10}$$

where N_{ijk} is the number of cases in D in which $X_i = x_i^k$ and $\mathbf{Pa}_i = \mathbf{pa}_i^j$. Note that the collection of counts N_{ijk} are sufficient statistics of the data for the model $\mathbf{m}$. In addition, we obtain the marginal likelihood (derived in Cooper and Herskovits, 1992):

$$p(D | \mathbf{m}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{11}$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. We then use Equation 1 and Equation 11 to compute the posterior probabilities $p(\mathbf{m} | D)$.

As a simple illustration of these ideas, suppose our hypothesis of interest is the outcome of $\mathbf{X}_{N+1}$, the next case to be seen after D. Also suppose that, for each possible outcome

[1] Bernardo and Smith (1994) provide a summary of likelihoods from the exponential family and their conjugate priors.

$\mathbf{x}_{N+1}$ of $\mathbf{X}_{N+1}$, the value of X_i is x_i^k and the configuration of $\mathbf{Pa}_i$ is $\mathbf{pa}_i^j$, where k and j depend on i. To compute $p(\mathbf{x}_{N+1}|D)$, we first average over our uncertainty about the parameters. Using Equations 4, 7, and 8, we obtain

$$p(\mathbf{x}_{N+1}|D,\mathbf{m}) = \int \left(\prod_{i=1}^{n} \theta_{ijk} \right) p(\boldsymbol{\theta}_m|D,\mathbf{m})\, d\boldsymbol{\theta}_m$$

Because parameters remain independent given D, we get

$$p(\mathbf{x}_{N+1}|D,\mathbf{m}) = \prod_{i=1}^{n} \int \theta_{ijk}\; p(\boldsymbol{\theta}_{ij}|D,\mathbf{m})\, d\boldsymbol{\theta}_{ij}$$

Because each integral in this product is the expectation of a Dirichlet distribution, we have

$$p(\mathbf{x}_{N+1}|D,\mathbf{m}) = \prod_{i=1}^{n} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \tag{12}$$

Finally, we average this expression for $p(\mathbf{x}_{N+1}|D,\mathbf{m})$ over the possible models using Equation 5 to obtain $p(\mathbf{x}_{N+1}|D)$.

3 Model Selection and Search

The full Bayesian approach is often impractical, even under the simplifying assumptions that we have described. One computation bottleneck in the full Bayesian approach is averaging over all models in Equation 4. If we consider causal models with n variables, the number of possible structure hypotheses is at least exponential in n. Consequently, in situations where we can not exclude almost all of these hypotheses, the approach is intractable. Statisticians, who have been confronted by this problem for decades in the context of other types of models, use two approaches to address this problem: *model selection* and *selective model averaging*. The former approach is to select a "good" model (i.e., structure hypothesis) from among all possible models, and use that model as if it were the correct model. The latter approach is to select a manageable number of good models from among all possible models and pretend that these models are exhaustive. These related approaches raise several important questions. In particular, do these approaches yield accurate results when applied to causal structures? If so, how do we search for good models?

The question of accuracy is difficult to answer in theory. Nonetheless, several researchers have shown experimentally that the selection of a single model that is likely a posteriori often yields accurate predictions (Cooper and Herskovits 1992; Aliferis and Cooper 1994; Heckerman et al., 1995) and that selective model averaging using Monte-Carlo methods can sometimes be efficient and yield even better predictions (Herskovits 1991; Madigan et al., 1996).

Chickering (1996a) has shown that for certain classes of prior distributions the problem of finding the model with the highest posterior is NP-Complete. However, a number of researchers have demonstrated that greedy search methods over a search space of DAGs works well. Also, constraint-based methods have been used as a first-step heuristic search for the most likely causal model (Singh and Valtorta, 1993; Spirtes and Meek, 1995). In addition, performing greedy searches in a space where Markov equivalent models (see definition below) are represented by a single model has improved performance (Spirtes and Meek, 1995; Chickering 1996b).

4 Priors

To compute the relative posterior probability of a model structure, we must assess the structure prior $p(\mathbf{m})$ and the parameter priors $p(\boldsymbol{\theta}_m|\mathbf{m})$. Unfortunately, when many model structures are possible, these assessments will be intractable. Nonetheless, under certain assumptions, we can derive the structure and parameter priors for many model structures from a manageable number of direct assessments.

4.1 Priors for Model Parameters

First, let us consider the assessment of priors for the parameters of model structures. We consider the approach of Heckerman et al. (1995) who address the case where the local likelihoods are multinomial distributions and the assumption of parameter independence holds.

Their approach is based on two key concepts: Markov equivalence and distribution equivalence. We say that two model structures for $\mathbf{X}$ are *Markov equivalent* if they represent the same set of conditional-independence assertions for $\mathbf{X}$ (Verma and Pearl, 1990). For example, given $\mathbf{X} = \{X, Y, Z\}$, the model structures $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, and $X \leftarrow Y \leftarrow Z$ represent only the independence assertion that X and Z are conditionally independent given Y. Consequently, these model structures are equivalent. Another example of Markov equivalence is the set of *complete model structures* on $\mathbf{X}$; a complete model is one that has no missing edge and which encodes no assertion of conditional independence. When $\mathbf{X}$ contains n variables, there are $n!$ possible complete model structures; one model structure for each possible ordering of the variables. All complete model structures for $p(\mathbf{x})$ are Markov equivalent. In general, two model structures are Markov equivalent if and only if they have the same structure ignoring arc directions and the same v-structures (Verma and Pearl, 1990). A *v-structure* is an ordered tuple (X, Y, Z) such that there is an arc from X to Y and from Z to Y, but no arc between X and Z.

The concept of distribution equivalence is closely related to that of Markov equivalence. Suppose that all causal models for $\mathbf{X}$ under consideration have local likelihoods in the family $\mathcal{F}$. This is not a restriction, per se, because $\mathcal{F}$ can be a large family. We say that two model structures $\mathbf{m}_1$ and $\mathbf{m}_2$ for $\mathbf{X}$ are *distribution equivalent with respect to (wrt) $\mathcal{F}$* if they represent the same joint probability distributions for $\mathbf{X}$—that is, if, for every $\boldsymbol{\theta}_{m1}$, there exists a $\boldsymbol{\theta}_{m2}$ such that $p(\mathbf{x}|\boldsymbol{\theta}_{m1}, \mathbf{m}_1) = p(\mathbf{x}|\boldsymbol{\theta}_{m2}, \mathbf{m}_2)$, and vice versa.

Distribution equivalence wrt some $\mathcal{F}$ implies Markov equivalence, but the converse does not hold. For example, when $\mathcal{F}$ is the family of generalized linear-regression models, the complete model structures for $n \geq 3$ variables do not represent the same sets of distributions. Nonetheless, there are families $\mathcal{F}$—for example, multinomial distributions and linear-regression models with Gaussian noise—where Markov equivalence implies distribution equivalence wrt $\mathcal{F}$ (Heckerman and Geiger, 1996). The notion of distribution equivalence is important, because if two model structures $\mathbf{m}_1$ and $\mathbf{m}_2$ are distribution equivalent wrt to a given $\mathcal{F}$, then it is often reasonable to expect that data can not help to discriminate them. That is, we expect $p(D|\mathbf{m}_1) = p(D|\mathbf{m}_2)$ for any data set D. Heckerman et al. (1995) call this property *likelihood equivalence*. Note that the constraint-based approach also does not discriminate among Markov equivalent structures.

Now let us return to the main issue of this section: the derivation of priors from a manageable number of assessments. Geiger and Heckerman (1995) show that the assumptions of parameter independence and likelihood equivalence imply that the parameters for any *complete* model structure $\mathbf{m}_c$ must have a Dirichlet distribution with constraints on the hyperparameters given by

$$\alpha_{ijk} = \alpha \; p(x_i^k, \mathbf{pa}_i^j | \mathbf{m}_c) \tag{13}$$

where α is the user's equivalent sample size[2], and $p(x_i^k, \mathbf{pa}_i^j|\mathbf{m}_c)$ is computed from the user's joint probability distribution $p(\mathbf{x}|\mathbf{m}_c)$. This result is rather remarkable, as the two assumptions leading to the constrained Dirichlet solution are qualitative.

To determine the priors for parameters of *incomplete* model structures, Heckerman et al. (1995) use the assumption of *parameter modularity,* which says that if X_i has the same parents in model structures $\mathbf{m}_1$ and $\mathbf{m}_2$, then

$$p(\boldsymbol{\theta}_{ij}|\mathbf{m}_1) = p(\boldsymbol{\theta}_{ij}|\mathbf{m}_2)$$

for $j = 1, \ldots, q_i$. They call this property parameter modularity, because it says that the distributions for parameters $\boldsymbol{\theta}_{ij}$ depend only on the structure of the model that is local to variable X_i—namely, X_i and its parents.

[2]Discussions of equivalent sample size can be found in Winkler (1967) and Heckerman et al. (1995).

Given the assumptions of parameter modularity and parameter independence, it is a simple matter to construct priors for the parameters of an arbitrary model structure given the priors on complete model structures. In particular, given parameter independence, we construct the priors for the parameters of each node separately. Furthermore, if node X_i has parents $\mathbf{Pa}_i$ in the given model structure, we identify a complete model structure where X_i has these parents, and use Equation 13 and parameter modularity to determine the priors for this node. The result is that all terms α_{ijk} for all model structures are determined by Equation 13. Thus, from the assessments α and $p(\mathbf{x}|\mathbf{m}_c)$, we can derive the parameter priors for all possible model structures. We can assess $p(\mathbf{x}|\mathbf{m}_c)$ by constructing a causal model called a *prior model*, that encodes this joint distribution. Heckerman et al. (1995) discuss the construction of this model.

4.2 Priors for Model Structures

Now, let us consider the assessment of priors on model structures. The simplest approach for assigning priors to model structures is to assume that every structure is equally likely. Of course, this assumption is typically inaccurate and used only for the sake of convenience. A simple refinement of this approach is to ask the user to exclude various structures (perhaps based on judgments of cause and effect), and then impose a uniform prior on the remaining structures. We illustrate this approach in Section 7.

Buntine (1991) describes a set of assumptions that leads to a richer yet efficient approach for assigning priors. The first assumption is that the variables can be ordered (e.g., through a knowledge of time precedence). The second assumption is that the presence or absence of possible arcs are mutually independent. Given these assumptions, $n(n-1)/2$ probability assessments (one for each possible arc in an ordering) determines the prior probability of every possible model structures. One extension to this approach is to allow for multiple possible orderings. One simplification is to assume that the probability that an arc is absent or present is independent of the specific arc in question. In this case, only one probability assessment is required.

An alternative approach, described by Heckerman et al. (1995) uses a prior model. The basic idea is to penalize the prior probability of any structure according to some measure of deviation between that structure and the prior model. Heckerman et al. (1995) suggest one reasonable measure of deviation.

Madigan et al. (1995) give yet another approach that makes use of imaginary data from a domain expert. In their approach, a computer program helps the user create a hypothetical set of complete data. Then, using techniques such as those in Section 2, they compute the posterior probabilities of model structures given this data, assuming the prior probabilities

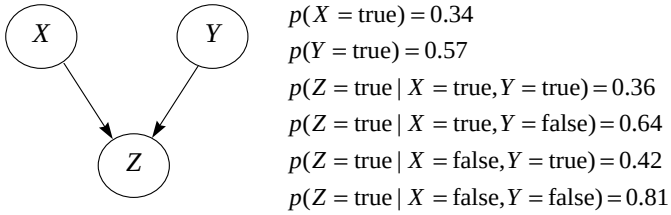

Figure 1: A causal model used to generate data.

of structures are uniform. Finally, they use these posterior probabilities as priors for the analysis of the real data.

5 Example

In this section, we provide a simple example that applies Bayesian model averaging and Bayesian model selection to the problem of causal discovery. In addition, we compare these methods with a constraint-based approach.

Let us consider a simple domain containing three binary variables X, Y, and Z. Let h denote the hypothesis that variable X causally influences variable Z. For brevity, we will sometimes state h as "X causes Z".

First, let us consider Bayesian model averaging. In this approach, we use Equation 4 to compute the probability that h is true given data D. Because our models are causal, the expression $p(D|\mathbf{m})$ reduces to an index function that is true when $\mathbf{m}$ contains an arc from node X to node Z. Thus, the right-hand-side of Equation 4 reduces to $\sum_{m''} p(\mathbf{m}''|D)$, where the sum is taken over all causal models $\mathbf{m}''$ that contain an arc from X to Z. For our three-variable domain, there are 25 possible causal models and, of these, there are eight models containing an arc from X to Z.

To compute $p(\mathbf{m}|D)$, we apply Equation 1, where the sum over $\mathbf{m}'$ is taken over the 25 models just mentioned. We assume a uniform prior distribution over the 25 possible models, so that $p(\mathbf{m}') = 1/25$ for every $\mathbf{m}'$. We use Equation 11 to compute the marginal likelihood $p(D|\mathbf{m})$. In applying Equation 11, we use the prior given by $\alpha_{ijk} = 1/r_i q_i$, which we obtain from Equation 13 using a uniform distribution for $p(\mathbf{x}|\mathbf{m}_c)$ and an equivalent sample $\alpha = 1$. Because this equivalent sample size is small, the data strongly influences the posterior probabilities for h that we derive.

To generate data, we first selected the model structure $X \longrightarrow Z \longleftarrow Y$ and randomly sampled its probabilities from a uniform distribution. The resulting model is shown in

Table 1: A summary of data used in the example.

number	sufficient statistics							
of cases	$\bar{x}\bar{y}\bar{z}$	$\bar{x}\bar{y}z$	$\bar{x}y\bar{z}$	$\bar{x}yz$	$x\bar{y}\bar{z}$	$x\bar{y}z$	$xy\bar{z}$	xyz
150	5	36	38	15	7	16	23	10
250	10	60	51	27	15	25	41	21
500	23	121	103	67	19	44	79	44
1000	44	242	222	152	51	80	134	75
2000	88	476	431	311	105	180	264	145

Table 2: Bayesian model averaging, Bayesian model selection, and constrain-based results for an analysis of whether "X causes Z" given data summarized in Table 1.

number of cases	$p(\text{"}X \text{ causes } Z\text{"}\|D)$	output of Bayesian model selection	output of PC algorithm
150	0.036	X and Z unrelated	X and Z unrelated
250	0.123	X and Z unrelated	X causes Z
500	0.141	X causes Z or Z causes X	X and Z unrelated (with inconsistency)
1000	0.593	X causes Z	X causes Z
2000	0.926	X causes Z	X causes Z

Figure 1. Next, we sampled data from the model according to its joint distribution. As we sampled the data, we kept a running total of the number cases seen in each possible configuration of $\{X, Y, Z\}$. These counts are sufficient statistics of the data for any causal model $\mathbf{m}$. These statistics are shown in Table 1 for the first 150, 250, 500, 1000, and 2000 cases in the data set.

The second column in Table 2 shows the results of applying Equation 4 under the assumptions stated above for the first N cases in the data set. When $N = 0$, the data set is empty, in which case probability of hypothesis h is just the prior probability of "X causes Z": 8/25=0.32. Table 2 shows that as the number of cases in the database increases, the probability that "X causes Z" increases monotonically as the number of cases increases. Although not shown, the probability increases toward 1 as the number of cases increases beyond 2000.

Column 3 in Table 2 shows the results of applying Bayesian model selection. Here, we

list the causal relationship(s) between X and Z found in the model or models with the highest posterior probability $p(\mathbf{m}|D)$. For example, when $N = 500$, there are three models that have the highest posterior probability. Two of the models have Z as a cause of X; and one has X as a cause of Z.

Column 4 in Table 2 shows the results of applying the PC constraint-based causal discovery algorithm (Spirtes et al., 1993), which is part of the Tetrad II system (Scheines et al., 1994). PC is designed to discover causal relationships that are expressed using DAGs.[3] We applied PC using its default settings, which include a statistical significance level of 0.05. Note that, for $N = 500$, the PC algorithm detected an inconsistency. In particular, the independence tests yielded (1) X and Z are dependent, (2) Y and Z are dependent, (3) X and Y are independent given Z, and (4) X and Z are independent given Y. These relationships are not consistent with the assumption underlying the PC algorithm that the only independence facts found to hold in the sample are those entailed by the Causal Markov condition applied to the generating model. In general, inconsistencies may arise due to the use of thresholds in the independence tests.

There are several weaknesses of the Bayesian-model-selection and constraint-based approaches illustrated by our results. One is that the output is categorical—there is no indication of the strength of the conclusion. Another is that the conclusions may be incorrect in that they disagree with the generative model. Model averaging (column 2) does not suffer from these weaknesses, because it indicates the strength of a causal hypothesis.

Although not illustrated here, another weakness of constraint-based approaches is that their output depends on the threshold used in independence tests. For causal conclusions to be correct asymptotically, the threshold must be adjusted as a function of sample size (N). In practice, however, it is unclear what this function should be.

Finally, we note that there are practical problems with model averaging. In particular, the domain can be so large that there are too many models over which to average. In such situations, the exact probabilities of causal hypotheses can not be calculated. However, we can use selective model averaging to derive approximate posterior probabilities, and consequently give some indication of the strength of causal hypotheses.

6 Methods for Incomplete Data and Hidden Variables

Among the assumptions that we described in Section 2, the one that is most often violated is the assumption that all variables are observed in every case. In this section, we examine

[3]The algorithm assumes that there are no *hidden* variables. See Sections 6 and 7 for a discussion of hidden-variable models and methods for learning them.

Bayesian methods for relaxing this assumption.

An important distinction for this discussion is that of hidden versus observable variable. A *hidden variable* is one that is unknown in all cases. An *observable variable* is one that is known in some (but not necessarily all) of the cases. We note that constraint-based and Bayesian methods differ significantly in the way that they missing data. Whereas constraint-based methods typically throw out cases that contain an observable variable with a missing value, Bayesian methods do not.

Another important distinction concerning missing data is whether or not the absence of an observation is dependent on the actual states of the variables. For example, a missing datum in a drug study may indicate that a patient became too sick—perhaps due to the side effects of the drug—to continue in the study. In contrast, if a variable is hidden, then the absence of this data is independent of state. Although Bayesian methods and graphical models are suited to the analysis of both situations, methods for handling missing data where absence is independent of state are simpler than those where absence and state are dependent. Here, we concentrate on the simpler situation. Readers interested in the more complicated case should see Rubin (1978), Robins (1986), Cooper (1995), and Spirtes et al. (1995).

Continuing with our example using discrete-multinomial likelihoods, suppose we observe a single incomplete case. Let $\mathbf{Y} \subset \mathbf{X}$ and $\mathbf{Z} = \mathbf{X} \backslash \mathbf{Y}$ denote the observed and unobserved variables in the case, respectively. Under the assumption of parameter independence, we can compute the posterior distribution of $\boldsymbol{\theta}_{ij}$ for model structure $\mathbf{m}$ as follows:

$$\begin{aligned} p(\boldsymbol{\theta}_{ij}|\mathbf{y},\mathbf{m}) &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{y},\mathbf{m})\ p(\boldsymbol{\theta}_{ij}|\mathbf{y},\mathbf{z},\mathbf{m}) \\ &= (1 - p(\mathbf{pa}_i^j|\mathbf{y},\mathbf{m}))\ \{p(\boldsymbol{\theta}_{ij}|\mathbf{m})\} + \sum_{k=1}^{r_i} p(x_i^k, \mathbf{pa}_i^j|\mathbf{y},\mathbf{m}) p(\boldsymbol{\theta}_{ij}|x_i^k, \mathbf{pa}_i^j, \mathbf{m}) \end{aligned} \tag{14}$$

(See Spiegelhalter and Lauritzen, 1990, for a derivation.) Each term $p(\boldsymbol{\theta}_{ij}|x_i^k, \mathbf{pa}_i^j, \mathbf{m})$ in Equation 14 is a Dirichlet distribution. Thus, unless both X_i and all the variables in $\mathbf{Pa}_i$ are observed in case $\mathbf{y}$, the posterior distribution of $\boldsymbol{\theta}_{ij}$ will be a linear combination of Dirichlet distributions—that is, a Dirichlet mixture with mixing coefficients $(1 - p(\mathbf{pa}_i^j|\mathbf{y},\mathbf{m}))$ and $p(x_i^k, \mathbf{pa}_i^j|\mathbf{y},\mathbf{m}), k = 1, \ldots, r_i$.

When we observe a second incomplete case, some or all of the Dirichlet components in Equation 14 will again split into Dirichlet mixtures. That is, the posterior distribution for $\boldsymbol{\theta}_{ij}$ will become a mixture of Dirichlet mixtures. As we continue to observe incomplete cases, each missing values for $\mathbf{Z}$, the posterior distribution for $\boldsymbol{\theta}_{ij}$ will contain a number of components that is exponential in the number of cases. In general, for any interesting set of local likelihoods and priors, the exact computation of the posterior distribution for $\boldsymbol{\theta}_m$

will be intractable. Thus, we require an approximation for incomplete data.

6.1 Monte-Carlo Methods

One class of approximations is based on Monte-Carlo or sampling methods. These approximations can be extremely accurate, provided one is willing to wait long enough for the computations to converge.

In this section, we discuss one of many Monte-Carlo methods known as *Gibbs sampling*, introduced by Geman and Geman (1984). Given variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ with some joint distribution $p(\mathbf{x})$, we can use a Gibbs sampler to approximate the expectation of a function $f(\mathbf{x})$ with respect to $p(\mathbf{x})$ as follows. First, we choose an initial state for each of the variables in $\mathbf{X}$ somehow (e.g., at random). Next, we pick some variable X_i, unassign its current state, and compute its probability distribution given the states of the other $n-1$ variables. Then, we sample a state for X_i based on this probability distribution, and compute $f(\mathbf{x})$. Finally, we iterate the previous two steps, keeping track of the average value of $f(\mathbf{x})$. In the limit, as the number of cases approach infinity, this average is equal to $\mathrm{E}_{p(\mathbf{x})}(f(\mathbf{x}))$ provided two conditions are met. First, the Gibbs sampler must be *irreducible*. That is, the probability distribution $p(\mathbf{x})$ must be such that we can eventually sample any possible configuration of $\mathbf{X}$ given any possible initial configuration of $\mathbf{X}$. For example, if $p(\mathbf{x})$ contains no zero probabilities, then the Gibbs sampler will be irreducible. Second, each X_i must be chosen infinitely often. In practice, an algorithm for deterministically rotating through the variables is typically used. Introductions to Gibbs sampling and other Monte-Carlo methods—including methods for initialization and a discussion of convergence—are given by Neal (1993) and Madigan and York (1995).

To illustrate Gibbs sampling, let us approximate the probability density $p(\boldsymbol{\theta}_m|D, \mathbf{m})$ for some particular configuration of $\boldsymbol{\theta}_m$, given an incomplete data set $D = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ and a causal model for discrete variables with independent Dirichlet priors. To approximate $p(\boldsymbol{\theta}_m|D, \mathbf{m})$, we first initialize the states of the unobserved variables in each case somehow. As a result, we have a complete random sample D_c. Second, we choose some variable X_{il} (variable X_i in case l) that is not observed in the original random sample D, and reassign its state according to the probability distribution

$$p(x'_{il}|D_c \setminus x_{il}, \mathbf{m}) = \frac{p(x'_{il}, D_c \setminus x_{il}|\mathbf{m})}{\sum_{x''_{il}} p(x''_{il}, D_c \setminus x_{il}|\mathbf{m})}$$

where $D_c \setminus x_{il}$ denotes the data set D_c with observation x_{il} removed, and the sum in the denominator runs over all states of variable X_{il}. As we have seen, the terms in the numerator and denominator can be computed efficiently (see Equation 11). Third, we repeat this

reassignment for all unobserved variables in D, producing a new complete random sample D'_c. Fourth, we compute the posterior density $p(\boldsymbol{\theta}_m|D'_c, \mathbf{m})$ as described in Equations 9 and 10. Finally, we iterate the previous three steps, and use the average of $p(\boldsymbol{\theta}_m|D'_c, \mathbf{m})$ as our approximation.

Monte-Carlo approximations are also useful for computing the marginal likelihood given incomplete data. One Monte-Carlo approach, described by Chib (1995) and Raftery (1996), uses Bayes' theorem:

$$p(D|\mathbf{m}) = \frac{p(\boldsymbol{\theta}_m|\mathbf{m})\ p(D|\boldsymbol{\theta}_m, \mathbf{m})}{p(\boldsymbol{\theta}_m|D, \mathbf{m})} \tag{15}$$

For any configuration of $\boldsymbol{\theta}_m$, the prior term in the numerator can be evaluated directly. In addition, the likelihood term in the numerator can be computed using causal-model inference (Jensen et al., 1990). Finally, the posterior term in the denominator can be computed using Gibbs sampling, as we have just described. Other, more sophisticated Monte-Carlo methods are described by DiCiccio et al. (1995).

6.2 The Gaussian Approximation

Monte-Carlo methods yield accurate results, but they are often intractable—for example, when the sample size is large. Another approximation that is more efficient than Monte-Carlo methods and often accurate for relatively large samples is the *Gaussian approximation* (e.g., Kass et al., 1988; Kass and Raftery, 1995).

The idea behind this approximation is that, for large amounts of data, $p(\boldsymbol{\theta}_m|D, \mathbf{m})$ $\propto p(D|\boldsymbol{\theta}_m, \mathbf{m}) \cdot p(\boldsymbol{\theta}_m|\mathbf{m})$ can often be approximated as a multivariate-Gaussian distribution. In particular, let

$$g(\boldsymbol{\theta}_m) \equiv \log(p(D|\boldsymbol{\theta}_m, \mathbf{m}) \cdot p(\boldsymbol{\theta}_m|\mathbf{m})) \tag{16}$$

Also, define $\tilde{\boldsymbol{\theta}}_m$ to be the configuration of $\boldsymbol{\theta}_m$ that maximizes $g(\boldsymbol{\theta}_m)$. This configuration also maximizes $p(\boldsymbol{\theta}_m|D, \mathbf{m})$, and is known as the *maximum a posteriori* (MAP) configuration of $\boldsymbol{\theta}_m$. Using a second degree Taylor polynomial of $g(\boldsymbol{\theta}_m)$ about the $\tilde{\boldsymbol{\theta}}_m$ to approximate $g(\boldsymbol{\theta}_m)$, we obtain

$$g(\boldsymbol{\theta}_m) \approx g(\tilde{\boldsymbol{\theta}}_m) - \frac{1}{2}(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)A(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)^t \tag{17}$$

where $(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)^t$ is the transpose of row vector $(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)$, and A is the negative Hessian of $g(\boldsymbol{\theta}_m)$ evaluated at $\tilde{\boldsymbol{\theta}}_m$. Raising $g(\boldsymbol{\theta}_m)$ to the power of e and using Equation 16, we obtain

$$\begin{aligned} p(\boldsymbol{\theta}_m|D, \mathbf{m}) &\propto p(D|\boldsymbol{\theta}_m, \mathbf{m})\ p(\boldsymbol{\theta}_m|\mathbf{m}) \\ &\approx p(D|\tilde{\boldsymbol{\theta}}_m, \mathbf{m})\ p(\tilde{\boldsymbol{\theta}}_m|\mathbf{m})\ \exp\{-\frac{1}{2}(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)A(\boldsymbol{\theta}_m - \tilde{\boldsymbol{\theta}}_m)^t\} \end{aligned} \tag{18}$$

Hence, the approximation for $p(\boldsymbol{\theta}_m|D, \mathbf{m})$ is Gaussian.

To compute the Gaussian approximation, we must compute $\tilde{\boldsymbol{\theta}_m}$ as well as the negative Hessian of $g(\boldsymbol{\theta}_m)$ evaluated at $\tilde{\boldsymbol{\theta}_m}$. In the following section, we discuss methods for finding $\tilde{\boldsymbol{\theta}_m}$. Meng and Rubin (1991) describe a numerical technique for computing the second derivatives. Raftery (1995) shows how to approximate the Hessian using likelihood-ratio tests that are available in many statistical packages. Thiesson (1995) demonstrates that, for multinomial distributions, the second derivatives can be computed using causal-model inference.

Using the Gaussian approximation, we can also approximate the marginal likelihood. Substituting Equation 18 into Equation 3, integrating, and taking the logarithm of the result, we obtain the approximation:

$$\log p(D|\mathbf{m}) \approx \log p(D|\tilde{\boldsymbol{\theta}_m}, \mathbf{m}) \ + \log p(\tilde{\boldsymbol{\theta}_m}|\mathbf{m}) \ + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A| \tag{19}$$

where d is the dimension of $g(\boldsymbol{\theta}_m)$. For a causal model with multinomial distributions, this dimension is typically given by $\prod_{i=1}^{n} q_i(r_i - 1)$. Sometimes, when there are hidden variables, this dimension is lower. See Geiger et al. (1996) for a discussion of this point. This approximation technique for integration is known as *Laplace's method*, and we refer to Equation 19 as the *Laplace approximation*. Kass et al. (1988) have shown that, under certain regularity conditions, the relative error of this approximation is $O_p(1/N)$, where N is the number of cases in D. Thus, the Laplace approximation can be extremely accurate. For more detailed discussions of this approximation, see—for example—Kass et al. (1988) and Kass and Raftery (1995).

Although Laplace's approximation is efficient relative to Monte-Carlo approaches, the computation of $|A|$ is nevertheless intensive for large-dimension models. One simplification is to approximate $|A|$ using only the diagonal elements of the Hessian A. Although in so doing, we incorrectly impose independencies among the parameters, researchers have shown that the approximation can be accurate in some circumstances (see, e.g., Becker and Le Cun, 1989, and Chickering and Heckerman, 1997). Another efficient variant of Laplace's approximation is described by Cheeseman and Stutz (1995) and Chickering and Heckerman (1997).

We obtain a very efficient (but less accurate) approximation by retaining only those terms in Equation 19 that increase with N: $\log p(D|\tilde{\boldsymbol{\theta}_m}, \mathbf{m})$, which increases linearly with N, and $\log|A|$, which increases as $d \log N$. Also, for large N, $\tilde{\boldsymbol{\theta}_m}$ can be approximated by $\hat{\boldsymbol{\theta}_m}$, the maximum likelihood configuration of $\boldsymbol{\theta}_m$ (see the following section). Thus, we obtain

$$\log p(D|\mathbf{m}) \approx \log p(D|\hat{\boldsymbol{\theta}_m}, \mathbf{m}) \ - \frac{d}{2}\log N \tag{20}$$

This approximation is called the *Bayesian information criterion* (BIC). Schwarz (1978)

has shown that the relative error of this approximation is $O_p(1)$ for a limited class of models. Haughton (1988) has extended this result to curved exponential models. Kass and Wasserman (1995) and

The BIC approximation is interesting in several respects. First, roughly speaking, it does not depend on the prior. Consequently, we can use the approximation without assessing a prior.[4] Second, the approximation is quite intuitive. Namely, it contains a term measuring how well the parameterized model predicts the data ($\log p(D|\hat{\boldsymbol{\theta}}_m, \mathbf{m})$) and a term that punishes the complexity of the model ($d/2\, logN$). Third, the BIC approximation is exactly minus the Minimum Description Length (MDL) criterion described by Rissanen (1987).

6.3 The MAP and ML Approximations and the EM Algorithm

As the sample size of the data increases, the Gaussian peak will become sharper, tending to a delta function at the MAP configuration $\tilde{\boldsymbol{\theta}}_m$. In this limit, we can replace the integral over $\boldsymbol{\theta}_m$ in Equation 5 with $p(h|\tilde{\boldsymbol{\theta}}_m, \mathbf{m})$. A further approximation is based on the observation that, as the sample size increases, the effect of the prior $p(\boldsymbol{\theta}_m|\mathbf{m})$ diminishes. Thus, we can approximate $\tilde{\boldsymbol{\theta}}_m$ by the maximum *maximum likelihood* (ML) configuration of $\boldsymbol{\theta}_m$:

$$\hat{\boldsymbol{\theta}}_m = \arg\max_{\boldsymbol{\theta}_m} \{p(D|\boldsymbol{\theta}_m, \mathbf{m})\}$$

One class of techniques for finding a ML or MAP is gradient-based optimization. For example, we can use gradient ascent, where we follow the derivatives of $g(\boldsymbol{\theta}_m)$ or the likelihood $p(D|\boldsymbol{\theta}_m, \mathbf{m})$ to a local maximum. Russell et al. (1995) and Thiesson (1995) show how to compute the derivatives of the likelihood for a causal model with multinomial distributions. Buntine (1994) discusses the more general case where the likelihood comes from the exponential family. Of course, these gradient-based methods find only local maxima.

Another technique for finding a local ML or MAP is the expectation–maximization (EM) algorithm (Dempster et al., 1977). To find a local MAP or ML, we begin by assigning a configuration to $\boldsymbol{\theta}_m$ somehow (e.g., at random). Next, we compute the *expected* sufficient statistics for a complete data set, where expectation is taken with respect to the joint distribution for $\mathbf{X}$ conditioned on the assigned configuration of $\boldsymbol{\theta}_m$ and the known data D. In our discrete example, we compute

$$\mathrm{E}_{p(\mathbf{x}|D,\boldsymbol{\theta}_s,\mathbf{m})}(N_{ijk}) = \sum_{l=1}^{N} p(x_i^k, \mathbf{pa}_i^j | \mathbf{y}_l, \boldsymbol{\theta}_m, \mathbf{m}) \tag{21}$$

[4] One of the technical assumptions used to derive this approximation is that the prior is bounded and bounded away from zero around $\hat{\boldsymbol{\theta}}_m$.

where $\mathbf{y}_l$ is the possibly incomplete lth case in D. When X_i and all the variables in $\mathbf{Pa}_i$ are observed in case $\mathbf{x}_l$, the term for this case requires a trivial computation: it is either zero or one. Otherwise, we can use any causal-model inference algorithm to evaluate the term. This computation is called the *expectation step* of the EM algorithm.

Next, we use the expected sufficient statistics as if they were actual sufficient statistics from a complete random sample D_c. If we are doing an ML calculation, then we determine the configuration of $\boldsymbol{\theta}_m$ that maximizes $p(D_c|\boldsymbol{\theta}_m, \mathbf{m})$. In our discrete example, we have

$$\theta_{ijk} = \frac{\mathrm{E}_{p(\mathbf{x}|D,\boldsymbol{\theta}_s,\mathbf{m})}(N_{ijk})}{\sum_{k=1}^{r_i} \mathrm{E}_{p(\mathbf{x}|D,\boldsymbol{\theta}_s,\mathbf{m})}(N_{ijk})}$$

If we are doing a MAP calculation, then we determine the configuration of $\boldsymbol{\theta}_m$ that maximizes $p(\boldsymbol{\theta}_m|D_c, \mathbf{m})$. In our discrete example, we have[5]

$$\theta_{ijk} = \frac{\alpha_{ijk} + \mathrm{E}_{p(\mathbf{x}|D,\boldsymbol{\theta}_s,\mathbf{m})}(N_{ijk})}{\sum_{k=1}^{r_i}(\alpha_{ijk} + \mathrm{E}_{p(\mathbf{x}|D,\boldsymbol{\theta}_s,\mathbf{m})}(N_{ijk}))}$$

This assignment is called the *maximization step* of the EM algorithm. Under certain regularity conditions, iteration of the expectation and maximization steps will converge to a local maximum. The EM algorithm is typically applied when sufficient statistics exist (i.e., when local likelihoods are in the exponential family), although generalizations of the EM algorithm have been used for more complicated local distributions (see, e.g., McLachlan and Krishnan, 1997).

7 A Case Study

To further illustrate the Bayesian approach and differences between it and the constraint-based approach, let us consider the following example. Sewell and Shah (1968) investigated factors that influence the intention of high school students to attend college. They measured the following variables for 10,318 Wisconsin high school seniors: *Sex* (SEX): male, female; *Socioeconomic Status* (SES): low, lower middle, upper middle, high; *Intelligence Quotient* (IQ): low, lower middle, upper middle, high; *Parental Encouragement* (PE): low, high; and *College Plans* (CP): yes, no. Our goal here is to understand the causal relationships among these variables.

The data are described by the sufficient statistics in Table 3. Each entry denotes the number of cases in which the five variables take on some particular configuration. The

[5]The MAP configuration $\tilde{\boldsymbol{\theta}}_m$ depends on the coordinate system in which the parameter variables are expressed. The MAP given here corresponds to the *canonical* coordinate system for the multinomial distribution (see, e.g., Bernardo and Smith, 1994, pp. 199–202).

Table 3: Sufficient statistics for the Sewall and Shah (1968) study.

4	349	13	64	9	207	33	72	12	126	38	54	10	67	49	43
2	232	27	84	7	201	64	95	12	115	93	92	17	79	119	59
8	166	47	91	6	120	74	110	17	92	148	100	6	42	198	73
4	48	39	57	5	47	123	90	9	41	224	65	8	17	414	54
5	454	9	44	5	312	14	47	8	216	20	35	13	96	28	24
11	285	29	61	19	236	47	88	12	164	62	85	15	113	72	50
7	163	36	72	13	193	75	90	12	174	91	100	20	81	142	77
6	50	36	58	5	70	110	76	12	48	230	81	13	49	360	98

first entry corresponds to the configuration SEX=male, SES=low, IQ=low, PE=low, and CP=yes. The remaining entries correspond to configurations obtained by cycling through the states of each variable such that the last variable (CP) varies most quickly. Thus, for example, the upper (lower) half of the table corresponds to male (female) students.

First, let us analyze the data under the assumption that there are no hidden variables. To generate priors for model parameters, we use the method described in Section 4.1 with an equivalent sample size of 5 and a prior model where $p(\mathbf{x}|\mathbf{m}_c)$ is uniform. (The results are not sensitive to the choice of parameter priors. For example, none of the results reported in this section change qualitatively for equivalent sample sizes ranging from 3 to 40.) For structure priors, we assume that all model structures are equally likely, except, on the basis of prior causal knowledge about the domain, we exclude structures where SEX and/or SES have parents, and/or CP have children. Because the data set is complete, we use Equation 11 to compute the posterior probabilities of model structures. The two most likely model structures found after an exhaustive search over all structures are shown in Figure 2. Note that the most likely graph has a posterior probability that is extremely close to one so that model averaging is not necessary.

If we adopt the Causal Markov condition and also assume that there are no hidden variables, then the arcs in both graphs can be interpreted causally. Some results are not surprising—for example the causal influence of socioeconomic status and IQ on college plans. Other results are more interesting. For example, from either graph we conclude that sex influences college plans only indirectly through parental influence. Also, the two graphs

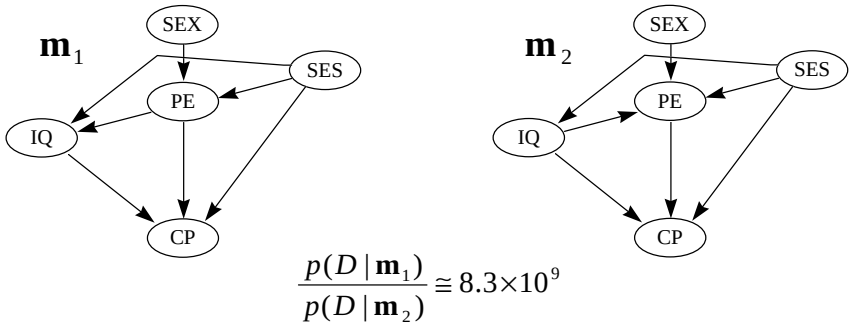

Figure 2: The a posteriori most likely model structures without hidden variables.

differ only by the orientation of the arc between PE and IQ. Either causal relationship is plausible.

We note that the second most likely graph was selected by Spirtes et al. (1993), who used the constraint-based PC algorithm with essentially identical assumptions. The only differences in the independence facts entailed by the most likely graph and the second most likely graphs are that the most likely graph entails SEX and IQ are independent given SES and PE whereas the second most likely graph entails SEX and IQ are marginally independent. Although both Bayesian and classical independence tests indicate that the conditional independence holds more strongly given the data, the PC algorithm chooses the second most likely graph due to its greedy nature. In particular, after the PC algorithm decides that SEX and IQ are marginally independent (at the threshold used by Spirtes et al.), it never considers the independence of SEX and IQ given SES and PE.

Returning to our analysis, the most suspicious result is the suggestion that socioeconomic status has a direct influence on IQ. To question this result, let us consider new models obtained from the models in Figure 2 by replacing this direct influence with a hidden variable pointing to both SES and IQ. Let us also consider models where (1) the hidden variable points to two or three of SES, IQ, and PE, (2) none, one, or both of the connections SES—PE and PE—IQ are removed, and (3) no variable has more than three parents. For each structure, we vary the number of states of the hidden variable from two to six.

We approximate the posterior probability of these models using the Cheeseman-Stutz (1995) variant of the Laplace approximation. To find the MAP $\tilde{\boldsymbol{\theta}}_m$, we use the EM algorithm, taking the largest local maximum from among 100 runs with different random initializations of $\boldsymbol{\theta}_m$. The model with the highest posterior probability is shown in Figure 3. This model is $2 \cdot 10^{10}$ times more likely that the best model containing no hidden variable. The next most likely model containing a hidden variable, which has one additional

arc from the hidden variable to PE, is $5 \cdot 10^{-9}$ times less likely than the best model. Thus, if we again adopt the Causal Markov condition and also assume that we have not omitted a reasonable model from consideration, then we have strong evidence that a hidden variable is influencing both socioeconomic status and IQ in this population—a sensible result. In particular, according to the probabilities in Figure 3, both SES and IQ are more likely to be high when H takes on the value 1. This observation suggests that the hidden variable represents "parent quality".

It is possible for constraint-based methods which use independence constraints to discriminate between models with and without hidden variables and to indicate the presence of latent variables (see Spirtes et al., 1993). Constraint-based methods also use non-independence constraints—for example, tetrad constraints—to make additional discriminations. However, these methods cannot distinguish between the model in Figure 3 and the most likely graph without hidden variables (the network on the left in Figure 1). Conditional independence constraints alone cannot be used to distinguish the models, because the two graphs entail the same set of independence facts on the observable variables. Independence constraints in combination with known non-independence constraints also fail to discriminate between the models. In addition, as this study illustrates, Bayesian methods can sometimes be used to determine the number of classes for a latent variable. A constraint-based method using only independence constraints can never determine the number of classes for a latent variable. We conjecture that any distinction among causal structures which can be made by constraint-based methods, even those not restricted to independence constraints, can be made using Bayesian methods. In addition, we conjecture that, asymptotically, when a constraint-based method chooses one model over another, the Bayesian approach will make the same choice, provided the Causal Markov condition and the assumption of faithfulness hold.

8 Open Issues

The Bayesian framework gives us a conceptually simple framework for learning causal models. Nonetheless, the Bayesian solution often comes with a high computational cost. For example, when we learn causal models containing hidden variables, both the exact computation of marginal likelihood and model averaging/selection can be intractable. Although the approximations described in Section 6 can be applied to address the difficulties associated with the computation of the marginal likelihood, model averaging and model selection remain difficult. The number of possible models with hidden variables is significantly larger than the number of possible DAGs over a fixed set of variables. Without constraining the

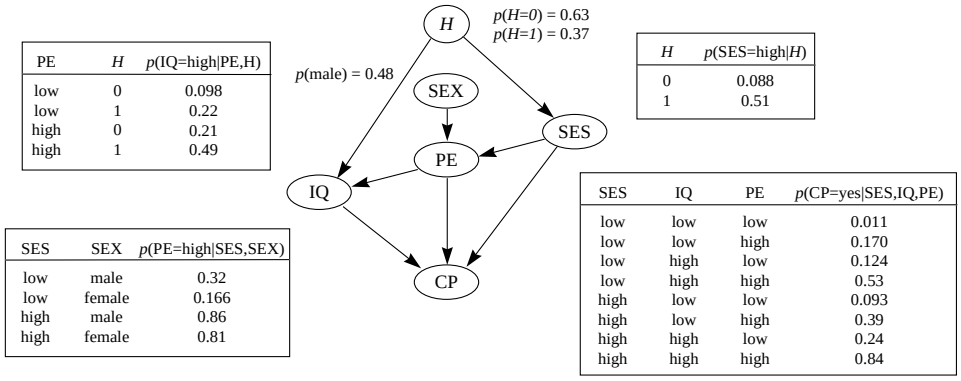

Figure 3: The a posteriori most likely model structure with a hidden variable. Probabilities shown are MAP values. Some probabilities are omitted for lack of space.

set of possible models with hidden variables—for instance, by restricting the number of hidden variables—the number of possible models is infinite. On a positive note, Spirtes et al. (1993) have shown that constraint-based methods under suitable assumptions can sometimes indicate the existence of a hidden common cause between two variables. Thus, it may be possible to use the constraint-based methods to suggest an initial set of plausible models containing hidden variables that can then be subjected to a Bayesian analysis.

Another problem associated with learning causal models containing hidden variables is the assessment of parameter priors. The approach in Section 4 can be applied in such situations, although the assessment of a joint distribution $p(\mathbf{x}|\mathbf{m}_c)$ in which $\mathbf{x}$ includes hidden variables can be difficult. Another approach may be to employ a property called *strong likelihood equivalence* (Heckerman, 1995). According to this property, data should not help to discriminate among two models that are distribution equivalent with respect to the non-hidden variables. Heckerman (1995) showed that any method that uses this property will yield priors that differ from those obtained using a prior network.[6]

One possibility for avoiding this problem with hidden-variable models, when the sample size is sufficiently large, is to use BIC-like approximations. Such approximations are commonly used (Crawford, 1994; Raftery, 1995). Nonetheless, the regularity conditions that guarantee $O_p(1)$ or better accuracy do not typically hold when choosing among causal models with hidden variables. Additional work is needed to obtain accurate approximations for the marginal likelihood of these models.

[6]In particular, Heckerman (1995) showed that strong likelihood equivalence is not consistent with parameter independence and parameter modularity.

Even in models without hidden variables there are many interesting issues to be addressed. In this paper we discuss only discrete variables having one type of local likelihood: the multinomial. Thiesson (1995) discusses a class of local likelihoods for discrete variables that use fewer parameters. Geiger and Heckerman (1994) and Buntine (1994) discuss simple linear local likelihoods for continuous nodes that have continuous and discrete variables. Buntine (1994) also discusses a general class of local likelihoods from the exponential family for nodes having no parents. Nonetheless, alternative likelihoods for discrete and continuous variables are desired. Local likelihoods with fewer parameters might allow for the selection of correct models with less data. In addition, local likelihoods that express more accurately the data generating process would allow for easier interpretation of the resulting models.

Acknowledgments

We thank Max Chickering for implementing the software used in our analysis of the Sewall and Shah (1968) data.

References

[Aliferis and Cooper, 1994] Aliferis, C. and Cooper, G. (1994). An evaluation of an algorithm for inductive learning of Bayesian belief networks using simulated data sets. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 8–14. Morgan Kaufmann.

[Becker and LeCun, 1989] Becker, S. and LeCun, Y. (1989). Improving the convergence of back-propagation learning with second order methods. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 29–37. Morgan Kaufmann.

[Bernardo and Smith, 1994] Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* John Wiley and Sons, New York.

[Buntine, 1991] Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence,* Los Angeles, CA, pages 52–60. Morgan Kaufmann.

[Buntine, 1994] Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225.

[Cheeseman and Stutz, 1995] Cheeseman, P. and Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and

Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press, Menlo Park, CA.

[Chib, 1995] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

[Chickering, 1996a] Chickering, D. (1996a). Learning Bayesian netowrks is NP-complete. In Fisher, D. and Lenz, H., editors, *Learning from Data*, pages 121–130. Springer-Verlag.

[Chickering, 1996b] Chickering, D. (1996b). Learning equivalence classes of Bayesian-network structures. In *Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence,* Portland, OR. Morgan Kaufmann.

[Chickering and Heckerman, 1997] Chickering, D. and Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29:181–212.

[Cooper, 1995] Cooper, G. (1995). Causal discovery from data in the presence of selection bias. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 140–150, Fort Lauderdale, FL.

[Cooper and Herskovits, 1992] Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

[Crawford, 1994] Crawford, S. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89:259–267.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

[DiCiccio et al., 1995] DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L. (July, 1995). Computing Bayes factors by combining simulation and asymptotic approximations. Technical Report 630, Department of Statistics, Carnegie Mellon University, PA.

[Geiger and Heckerman, 1994] Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 235–243. Morgan Kaufmann.

[Geiger and Heckerman, 1995] Geiger, D. and Heckerman, D. (Revised February, 1995). A characterization of the Dirichlet distribution applicable to learning Bayesian networks. Technical Report MSR-TR-94-16, Microsoft Research, Redmond, WA.

[Geiger et al., 1996] Geiger, D., Heckerman, D., and Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence,* Portland, OR, pages 283–290. Morgan Kaufmann.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–742.

[Haughton, 1988] Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16:342–355.

[Heckerman, 1995] Heckerman, D. (1995). A Bayesian approach for learning causal networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 285–295. Morgan Kaufmann.

[Heckerman and Geiger, 1996] Heckerman, D. and Geiger, D. (Revised, November, 1996). Likelihoods and priors for Bayesian networks. Technical Report MSR-TR-95-54, Microsoft Research, Redmond, WA.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

[Herskovits, 1991] Herskovits, E. (1991). *Computer-based probabilistic network construction.* PhD thesis, Medical Information Sciences, Stanford University, Stanford, CA.

[Jensen et al., 1990] Jensen, F., Lauritzen, S., and Olesen, K. (1990). Bayesian updating in recursive graphical models by local computations. *Computational Statisticals Quarterly*, 4:269–282.

[Kass and Raftery, 1995] Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

[Kass et al., 1988] Kass, R., Tierney, L., and Kadane, J. (1988). Asymptotics in Bayesian computation. In Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3*, pages 261–278. Oxford University Press.

[Kass and Wasserman, 1995] Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

[Madigan et al., 1995] Madigan, D., Garvin, J., and Raftery, A. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics: Theory and Methods*, 24:2271–2292.

[Madigan et al., 1996] Madigan, D., Raftery, A., Volinsky, C., and Hoeting, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models,* Portland, OR.

[Madigan and York, 1995] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232.

[McLachlan and Krishnan, 1997] McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley.

[Meek, 1995] Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 411–418. Morgan Kaufmann.

[Meng and Rubin, 1991] Meng, X. and Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86:899–909.

[Neal, 1993] Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

[Raftery, 1995] Raftery, A. (1995). Bayesian model selection in social research. In Marsden, P., editor, *Sociological Methodology*. Blackwells, Cambridge, MA.

[Raftery, 1996] Raftery, A. (1996). *Hypothesis testing and model selection*, chapter 10. Chapman and Hall.

[Rissanen, 1987] Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:223–239 and 253–265.

[Robins, 1986] Robins, J. (1986). A new approach to causal interence in mortality studies with sustained exposure results. *Mathematical Modelling*, 7:1393–1512.

[Rubin, 1978] Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

[Russell et al., 1995] Russell, S., Binder, J., Koller, D., and Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* Montreal, QU, pages 1146–1152. Morgan Kaufmann, San Mateo, CA.

[Scheines et al., 1994] Scheines, R., Spirtes, P., Glymour, C., and Meek, C. (1994). *Tetrad II: User's Manual.* Lawrence Erlbaum, Hillsdale, N.J.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

[Sewell and Shah, 1968] Sewell, W. and Shah, V. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 73:559–572.

[Singh and Valtorta, 1993] Singh, M. and Valtorta, M. (1993). An algorithm for the construction of Bayesian network structures from data. In *Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence,* Washington, DC, pages 259–265. Morgan Kaufmann.

[Spiegelhalter and Lauritzen, 1990] Spiegelhalter, D. and Lauritzen, S. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag, New York.

[Spirtes and Meek, 1995] Spirtes, P. and Meek, C. (1995). Learning Bayesian networks with discrete variables from data. In *Proceedings of First International Conference on Knowledge Discovery and Data Mining,* Montreal, QU. Morgan Kaufmann.

[Spirtes et al., 1995] Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 499–506. Morgan Kaufmann.

[Thiesson, 1995] Thiesson, B. (1995). Score and information for recursive exponential models with incomplete data. Technical report, Institute of Electronic Systems, Aalborg University, Aalborg, Denmark.

[Verma and Pearl, 1990] Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence,* Boston, MA, pages 220–227. Morgan Kaufmann.

[Winkler, 1967] Winkler, R. (1967). The assessment of prior distributions in Bayesian analysis. *American Statistical Association Journal*, 62:776–800.