Bayesian Analysis (2004) **1**, Number 1, pp. 1–44

Variational Bayesian Learning of Directed Graphical Models with Hidden Variables

Matthew J. Beal
Computer Science & Engineering
SUNY at Buffalo
New York 14260-2000, USA
mbeal@cse.buffalo.edu

Zoubin Ghahramani
Department of Engineering
University of Cambridge
Cambridge CB2 11PZ, UK
zoubin@eng.cam.ca.uk

Abstract. A key problem in statistics and machine learning is inferring suitable structure of a model given some observed data. A Bayesian approach to model comparison makes use of the marginal likelihood of each candidate model to form a posterior distribution over models; unfortunately for most models of interest, notably those containing hidden or latent variables, the marginal likelihood is intractable to compute.

We present the variational Bayesian (VB) algorithm for directed graphical models, which optimises a lower bound approximation to the marginal likelihood in a procedure similar to the standard EM algorithm. We show that for a large class of models, which we call conjugate exponential, the VB algorithm is a straightforward generalisation of the EM algorithm that incorporates uncertainty over model parameters. In a thorough case study using a small class of bipartite DAGs containing hidden variables, we compare the accuracy of the VB approximation to existing asymptotic-data approximations such as the Bayesian Information Criterion (BIC) and the Cheeseman-Stutz (CS) criterion, and also to a sampling based gold standard, Annealed Importance Sampling (AIS). We find that the VB algorithm is empirically superior to CS and BIC, and much faster than AIS. Moreover, we prove that a VB approximation can always be constructed in such a way that guarantees it to be more accurate than the CS approximation.

Keywords: Approximate Bayesian Inference, Bayes Factors, Directed Acyclic Graphs, EM Algorithm, Graphical Models, Markov Chain Monte Carlo, Model Selection, Variational Bayes

1 Introduction

Graphical models are becoming increasingly popular as tools for expressing probabilistic models found in various machine learning and applied statistics settings. One of the key problems is learning suitable structure of such graphical models from a data set, $\mathbf{y}$. This task corresponds to considering different model complexities — too complex a model will overfit the data and too simple a model underfits, with neither extreme generalising well to new data. Discovering a suitable structure entails determining which conditional dependency relationships amongst model variables are supported by the data. A Bayesian approach to model selection, comparison, or averaging relies on an important and difficult quantity: the *marginal likelihood* $p(\mathbf{y} \,|\, m)$ under each candidate

model, m. The marginal likelihood is an important quantity because, when combined with a prior distribution over candidate models, it can be used to form the posterior distribution over models given observed data, $p(m \,|\, \mathbf{y}) \propto p(m)p(\mathbf{y} \,|\, m)$. The marginal likelihood is a difficult quantity because it involves integrating out parameters and, for models containing hidden (or latent or missing) variables (thus encompassing many models of interest to statisticians and machine learning practitioners alike), it can be intractable to compute.

The marginal likelihood is tractable to compute for certain simple types of graphs; one such type is the case of *fully observed* discrete-variable directed acyclic graphs with Dirichlet priors on the parameters (Heckerman et al. 1995; Heckerman 1996). Unfortunately, if these graphical models include hidden variables, the marginal likelihood becomes intractable to compute even for moderately sized observed data sets. Estimating the marginal likelihood presents a difficult challenge for approximate methods such as asymptotic-data criteria and sampling techniques.

In this article we investigate a novel application of the variational Bayesian (VB) framework—first described in Attias (1999b)—to approximating the marginal likelihood of discrete-variable directed acyclic graph (DAG) structures that contain hidden variables. Variational Bayesian methods approximate the quantity of interest with a strict lower bound, and the framework readily provides algorithms to optimise the approximation. We describe the variational Bayesian methodology applied to a large class of graphical models which we call conjugate-exponential, derive the VB approximation as applied to discrete DAGs with hidden variables, and show that the resulting algorithm that optimises the approximation closely resembles the standard Expectation-Maximisation (EM) algorithm of Dempster et al. (1977). It will be seen for conjugate-exponential models that the VB methodology is an elegant Bayesian generalisation of the EM framework, replacing point estimates of parameters encountered in EM learning with *distributions* over parameters, thus naturally reflecting the uncertainty over the settings of their values given the data. Previous work has applied the VB methodology to particular instances of conjugate-exponential models, for example MacKay (1997), and Ghahramani and Beal (2000, 2001); Beal (2003) describes in more detail the theoretical results for VB in conjugate-exponential models.

We also briefly outline and compute the Bayesian Information Criterion (BIC) and Cheeseman-Stutz (CS) approximations to the marginal likelihood for DAGs (Schwarz 1978; Cheeseman and Stutz 1996), and compare these to VB in a particular model selection task. The particular task we have chosen is that of finding which of several possible structures for a simple graphical model (containing hidden and observed variables) has given rise to a set of observed data. The success of each approximation is measured by how it ranks the true model that generated the data amongst the alternatives, and also by the accuracy of the marginal likelihood estimate.

As a gold standard, against which we can compare these approximations, we consider sampling estimates of the marginal likelihood using the Annealed Importance Sampling (AIS) method of Neal (2001). We consider AIS to be a "gold standard" in the sense that we believe it is one of the best methods to date for obtaining reliable estimates of

the marginal likelihoods of the type of models explored here, given sufficient sampling computation. To the best of our knowledge, the AIS analysis we present constitutes the first serious case study of the tightness of variational Bayesian bounds. An analysis of the limitations of AIS is also provided. The aim of the comparison is to establish the reliability of the VB approximation as an estimate of the marginal likelihood in the general incomplete-data setting, so that it can be used in larger problems — for example embedded in a (greedy) structure search amongst a much larger class of models.

The remainder of this article is arranged as follows. Section 2 begins by examining the model selection question for discrete directed acyclic graphs, and shows how exact marginal likelihood calculation becomes computationally intractable when the graph contains hidden variables. In Section 3 we briefly cover the EM algorithm for maximum likelihood (ML) and maximum a posteriori (MAP) parameter estimation in DAGs with hidden variables, and derive and discuss the BIC and CS asymptotic approximations. We then introduce the necessary methodology for variational Bayesian learning, and present the VBEM algorithm for variational Bayesian lower bound optimisation of the marginal likelihood — in the case of discrete DAGs we show that this is a straightforward generalisation of the MAP EM algorithm. In Section 3.6 we describe an Annealed Importance Sampling method for estimating marginal likelihoods of discrete DAGs. In Section 4 we evaluate the performance of these different approximation methods on the simple (yet non-trivial) model selection task of determining which of all possible structures within a class generated a data set. Section 5 provides an analysis and discussion of the limitations of the AIS implementation and suggests possible extensions to it. In Section 6 we consider the CS approximation, which is one of the state-of-the-art approximations, and extend a result due to Minka (2001) that shows that the CS approximation is a lower bound on the marginal likelihood in the case of mixture models, by showing how the CS approximation can be constructed for any model containing hidden variables. We complete this section by proving that there exists a VB bound that is guaranteed to be at least as tight or tighter than the CS bound, independent of the model structure and type. Finally, we conclude in Section 7 and suggest directions for future research.

2 Calculating the marginal likelihood of DAGs

We focus on discrete-valued Directed Acyclic Graphs, although all the methodology described in the following sections is readily extended to models involving real-valued variables. Consider a data set of size n, consisting of independent and identically distributed (i.i.d.) observed variables $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_n\}$, where each $\mathbf{y}_i$ is a vector of discrete-valued variables. We model this observed data $\mathbf{y}$ by assuming that it is generated by a discrete directed acyclic graph consisting of hidden variables, $\mathbf{s}$, and observed variables, $\mathbf{y}$. Combining hidden and observed variables we have $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ $= \{(\mathbf{s}_1, \mathbf{y}_1), \ldots, (\mathbf{s}_n, \mathbf{y}_n)\}$. The elements of each vector $\mathbf{z}_i$ for $i = 1, \ldots, n$ are indexed from $j = 1, \ldots, |\mathbf{z}_i|$, where $|\mathbf{z}_i|$ is the number of variables in the data vector $\mathbf{z}_i$. We define two sets of indices $\mathcal{H}$ and $\mathcal{V}$ such that those $j \in \mathcal{H}$ are the hidden variables and those $j \in \mathcal{V}$ are observed variables, i.e. $\mathbf{s}_i = \{\mathbf{z}_{ij} : j \in \mathcal{H}\}$ and $\mathbf{y}_i = \{\mathbf{z}_{ij} : j \in \mathcal{V}\}$.

Note that $\mathbf{z}_i = \{\mathbf{s}_i, \mathbf{y}_i\}$ contains both hidden and observed variables — we refer to this as the *complete-data* for data point i. The *incomplete-data*, $\mathbf{y}_i$, is that which constitutes the observed data. Note that the meaning of $|\cdot|$ will vary depending on the type of its argument, for example: $|\mathbf{z}| = |\mathbf{s}| = |\mathbf{y}|$ is the number of data points, n; $|\mathbf{s}_i|$ is the number of hidden variables (for the ith data point); $|\mathbf{s}_{ij}|$ is the cardinality (or the number of possible settings) of the jth hidden variable (for the ith data point).

In a DAG, the complete-data likelihood factorises into a product of local probabilities on each variable

$$p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \;, \tag{1}$$

where $\mathbf{pa}(j)$ denotes the vector of indices of the parents of the jth variable. Each variable in the graph is multinomial, and the parameters for the graph are the collection of vectors of probabilities on each variable given each configuration of its parents. For example, the parameter for a binary variable which has two ternary parents is a matrix of size $(3^2 \times 2)$ with each row summing to one. For a variable j without any parents ($\mathbf{pa}(j) = \emptyset$), then the parameter is simply a vector of its prior probabilities. Using θ_{jlk} to denote the probability that variable j takes on value k when its parents are in configuration l, then the complete-data likelihood can be written out as a product of terms of the form

$$p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) = \prod_{l=1}^{\left|\mathbf{z}_{i\mathbf{pa}(j)}\right|} \prod_{k=1}^{|\mathbf{z}_{ij}|} \theta_{jlk}^{\delta(\mathbf{z}_{ij},k)\delta(\mathbf{z}_{i\mathbf{pa}(j)},l)} \tag{2}$$

with $\sum_k \theta_{jlk} = 1 \;\forall\; \{j, l\}$. Here we use $\left|\mathbf{z}_{i\mathbf{pa}(j)}\right|$ to denote the number of joint settings of the parents of variable j. We use Kronecker-δ notation: $\delta(\cdot, \cdot)$ is 1 if its arguments are identical and zero otherwise. The parameters are given independent Dirichlet priors, which are conjugate to the complete-data likelihood above (thereby satisfying Condition 1 for conjugate-exponential models (42), which is required later). The prior is factorised over variables and parent configurations; these choices then satisfy the *global* and *local* independence assumptions of Heckerman et al. (1995). For each parameter $\boldsymbol{\theta}_{jl} = \{\theta_{jl1}, \ldots, \theta_{jl|\mathbf{z}_{ij}|}\}$, the Dirichlet prior is

$$p(\boldsymbol{\theta}_{jl} \,|\, \boldsymbol{\lambda}_{jl}, m) = \frac{\Gamma(\lambda_{jl}^0)}{\prod_k \Gamma(\lambda_{jlk})} \prod_k \theta_{jlk}^{\lambda_{jlk}-1} \;, \tag{3}$$

where $\boldsymbol{\lambda}$ are hyperparameters, $\boldsymbol{\lambda}_{jl} = \{\lambda_{jl1}, \ldots, \lambda_{jl|\mathbf{z}_{ij}|}\}$, and $\lambda_{jlk} > 0 \;\forall\; k$, $\lambda_{jl}^0 = \sum_k \lambda_{jlk}$, $\Gamma(\cdot)$ is the gamma function, and the domain of $\boldsymbol{\theta}$ is confined to the simplex of probabilities that sum to 1. This form of prior is assumed throughout this article. Since we do not focus on inferring the hyperparameters we use the shorthand $p(\boldsymbol{\theta} \,|\, m)$ to denote the prior from here on. In the discrete-variable case we are considering, the complete-data

marginal likelihood is tractable to compute:

$$p(\mathbf{z} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \tag{4}$$

$$= \prod_{j=1}^{|\mathbf{z}_i|} \prod_{l=1}^{|\mathbf{z}_{i\mathbf{pa}(j)}|} \frac{\Gamma(\lambda^0_{jl})}{\Gamma(\lambda^0_{jl} + N_{jl})} \prod_{k=1}^{|\mathbf{z}_{ij}|} \frac{\Gamma(\lambda_{jlk} + N_{jlk})}{\Gamma(\lambda_{jlk})} \;, \tag{5}$$

where N_{jlk} is defined as the count in the data for the number of instances of variable j being in configuration k with parental configuration l:

$$N_{jlk} = \sum_{i=1}^{n} \delta(\mathbf{z}_{ij}, k) \delta(\mathbf{z}_{i\mathbf{pa}(j)}, l), \qquad \text{and} \quad N_{jl} = \sum_{k=1}^{|\mathbf{z}_{ij}|} N_{jlk} \;. \tag{6}$$

Note that if the data set is complete — that is to say there are no hidden variables — then $\mathbf{s} = \emptyset$ and so $\mathbf{z} = \mathbf{y}$, and the quantities N_{jlk} can be computed directly from the data.

The incomplete-data likelihood results from summing over all settings of the hidden variables and taking the product over i.i.d. presentations of the data:

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \;. \tag{7}$$

Now the incomplete-data *marginal* likelihood for n cases follows from marginalising out the parameters of the model with respect to their prior distribution:

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \tag{8}$$

$$= \sum_{\{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}\}_{i=1}^{n}} \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \;. \tag{9}$$

The expression (8) is computationally intractable due to the expectation (integral) over the real-valued conditional probabilities $\boldsymbol{\theta}$, which couples the hidden variables across i.i.d. data instances. Put another way (9), pulling the summation to the left of the product over n instances results in a summation with a number of summands exponential in the number of data n. In the worst case, (8) or (9) can be evaluated as the sum of $\left(\prod_{j \in \mathcal{H}} |\mathbf{z}_{ij}|\right)^n$ Dirichlet integrals. To take an example, a model with just $|\mathbf{s}_i| = 2$ hidden variables and $n = 100$ data points requires the evaluation of 2^{100} Dirichlet integrals. This means that a linear increase in the amount of observed data results in an exponential increase in the cost of inference.

Our goal is to learn the conditional independence structure of the model — that is, which variables are parents of each variable. Ideally, we should compare structures based on their posterior probabilities, and to compute this posterior we need to first compute the marginal likelihood (8).

The next section examines several methods that attempt to approximate the marginal likelihood (8). We focus on a variational Bayesian algorithm, which we compare to asymptotic criteria and also to sampling-based estimates. For the moment we assume that the cardinalities of the variables — in particular the hidden variables — are fixed beforehand; our wish is to discover how many hidden variables there are and what their connectivity is to other variables in the graph. The related problem of determining the cardinality of the variables from data can also be addressed in the variational Bayesian framework, as for example has been recently demonstrated for Hidden Markov Models (Beal 2003).

3 Estimating the marginal likelihood

In this section we look at some approximations to the marginal likelihood for a model m, which we refer to henceforth as the *scores* for m. In Section 3.1 we first review ML and MAP parameter learning and briefly present the EM algorithm for a general discrete-variable directed graphical model with hidden variables. Using the final parameters obtained from an EM optimisation, we can then construct various asymptotic approximations to the marginal likelihood, and so derive the BIC and Cheeseman-Stutz criteria, described in Sections 3.2 and 3.3, respectively. An alternative approach is provided by the variational Bayesian framework, which we review in some detail in Section 3.4. In the case of discrete directed acyclic graphs with Dirichlet priors, the model is *conjugate-exponential* (defined below), and the VB framework produces a very simple VBEM algorithm. This algorithm is a generalisation of the EM algorithm, and as such be cast in a way that resembles a direct extension of the EM algorithm for MAP parameter learning; the algorithm for VB learning for these models is presented in Section 3.5. In Section 3.6 we derive an *annealed importance sampling* method (AIS) for this class of graphical model, which is considered to be the current state-of-the-art technique for estimating the marginal likelihood of these models using sampling. Armed with these various approximations we pit them against each other in a model selection task, described in Section 4.

3.1 ML and MAP parameter estimation for DAGs

We begin by deriving the EM algorithm for ML/MAP estimation via a lower bound interpretation (see Neal and Hinton 1998). We start with the incomplete-data log likelihood, and lower bound it by a functional $\mathcal{F}(q_{\mathbf{s}}(\mathbf{s}), \boldsymbol{\theta})$ by appealing to Jensen's

inequality as follows

$$\ln p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \ln \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j\in\mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \tag{10}$$

$$= \sum_{i=1}^{n} \ln \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \frac{\prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \tag{11}$$

$$\geq \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{\prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \tag{12}$$

$$= \mathcal{F}(\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}, \boldsymbol{\theta}) \ . \tag{13}$$

The first line is simply the logarithm of equation (7); in the second line we have used the shorthand $\mathbf{s}_i = \{\mathbf{z}_{ij}\}_{j\in\mathcal{H}}$ to denote all hidden variables corresponding to the ith data point, and have multiplied and divided the inner summand (over $\mathbf{s}_i$) by a *variational distribution* $q_{\mathbf{s}_i}(\mathbf{s}_i)$ — one for each data point $\mathbf{y}_i$. The inequality that folllows results from the concavity of the logarithm function and results in an expression that is a strict lower bound on the log complete data likelihood, denoted $\mathcal{F}(\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}, \boldsymbol{\theta})$. This expression depends on the parameters of the model, $\boldsymbol{\theta}$, and is a functional of the variational distributions $\{q_{\mathbf{s}_i}\}_{i=1}^{n}$.

Since $\mathcal{F}(\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}, \boldsymbol{\theta})$ is a lower bound on a quantity we wish to maximise, we maximise the bound by taking functional derivatives with respect to each $q_{\mathbf{s}_i}(\mathbf{s}_i)$ while keeping the remaining $\{q_{\mathbf{s}_{i'}}(\mathbf{s}_{i'})\}_{i'\neq i}$ fixed, and set these to zero yielding

$$q_{\mathbf{s}_i}(\mathbf{s}_i) = p(\mathbf{s}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}) \quad \forall \ i \ . \tag{14}$$

Thus the optimal setting of each variational distribution is in fact the exact posterior distribution for the hidden variable for that data point. This is the E step of the celebrated EM algorithm; with these settings of the distributions $\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}$, it can easily be shown that the bound is tight — that is to say, the difference between $\ln p(\mathbf{y} \,|\, \boldsymbol{\theta})$ and $\mathcal{F}(\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}, \boldsymbol{\theta})$ is exactly zero.

The M step of the EM algorithm is obtained by taking derivatives of the bound with respect to the parameters $\boldsymbol{\theta}$, while holding fixed the distributions $q_{\mathbf{s}_i}(\mathbf{s}_i) \ \forall \ i$. Each $\boldsymbol{\theta}_{jl}$ is constrained to sum to one, and so we enforce this with Lagrange multipliers c_{jl},

$$\frac{\partial}{\partial \theta_{jlk}} \mathcal{F}(q_{\mathbf{s}}(\mathbf{s}), \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \frac{\partial}{\partial \theta_{jlk}} \ln p(\mathbf{z}_{ij} \,|\, \mathbf{x}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}_j) + c_{jl} \tag{15}$$

$$= \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \delta(\mathbf{z}_{ij}, k) \delta(\mathbf{z}_{i\mathbf{pa}(j)}, l) \frac{\partial}{\partial \theta_{jlk}} \ln \theta_{jlk} + c_{jl} = 0 \ , \tag{16}$$

which upon rearrangement gives

$$\theta_{jlk} \propto \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \delta(\mathbf{z}_{ij}, k) \delta(\mathbf{z}_{i\mathbf{pa}(j)}, l) \ . \tag{17}$$

Due to the normalisation constraint on $\boldsymbol{\theta}_{jl}$ the M step can be written

$$\textbf{M step (ML):} \qquad \theta_{jlk} = \frac{N_{jlk}}{\sum_{k'=1}^{|\mathbf{z}_{ij}|} N_{jlk'}} \ , \tag{18}$$

where the N_{jlk} are defined as

$$N_{jlk} = \sum_{i=1}^{n} \left\langle \delta(\mathbf{z}_{ij}, k)\delta(\mathbf{z}_{i\mathbf{pa}(j)}, l) \right\rangle_{q_{\mathbf{s}_i}(\mathbf{s}_i)} \ , \tag{19}$$

where angled-brackets $\langle\cdot\rangle_{q_{\mathbf{s}_i}(\mathbf{s}_i)}$ are used to denote expectation with respect to the hidden variable posterior $q_{\mathbf{s}_i}(\mathbf{s}_i)$ found in the preceding E step. The N_{jlk} are interpreted as the expected number of counts for observing settings of children and parent configurations over observed and hidden variables. In the cases where both j and $\mathbf{pa}(j)$ are observed variables, N_{jlk} reduces to the simple empirical count, as in (6). Otherwise, if j or its parents are hidden, then expectations need be taken over the posterior $q_{\mathbf{s}_i}(\mathbf{s}_i)$ obtained in the E step.

If we require the MAP EM algorithm, we instead lower bound $\ln p(\boldsymbol{\theta})p(\mathbf{y} \,|\, \boldsymbol{\theta})$. The E step remains the same, but the M step uses augmented counts from the prior of the form in (3) to give the following update:

$$\textbf{M step (MAP):} \qquad \theta_{jlk} = \frac{\lambda_{jlk} - 1 + N_{jlk}}{\sum_{k'=1}^{|\mathbf{z}_{ij}|} \lambda_{jlk'} - 1 + N_{jlk'}} \ . \tag{20}$$

Repeated applications of the E step (14) and the M step (18, 20) are guaranteed to increase the log likelihood (with equation (18)) or the log posterior (with equation (20)) of the parameters at every iteration, and converge to a local maximum. We note that MAP estimation is inherently basis-dependent: for any particular $\boldsymbol{\theta}^*$ having non-zero prior probability, it is possible to find a (one-to-one) reparameterisation $\boldsymbol{\phi}(\boldsymbol{\theta})$ such that the MAP estimate for $\boldsymbol{\phi}$ is at $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$. This is an obvious drawback of MAP parameter estimation. Moreover, the use of (20) can produce erroneous results in the case of $\lambda_{jlk} < 1$, in the form of negative probabilities. Conventionally, researchers have limited themselves to Dirichlet priors in which every $\lambda_{jlk} \geq 1$, although in MacKay (1998) it is shown how a reparameterisation of $\boldsymbol{\theta}$ into the softmax basis results in MAP updates which do not suffer from this problem (which look identical to (20), but without the -1 in numerator and denominator). Note that our EM algorithms were indeed carried out in the softmax basis, which avoids such effects and problems with parameters lying near their domain boundaries.

3.2 The BIC

The Bayesian Information Criterion approximation (BIC; Schwarz 1978) is the asymptotic limit to large data sets of the Laplace approximation (Kass and Raftery 1995; MacKay 1995). The Laplace approximation makes a local quadratic approximation to

the log posterior around a MAP parameter estimate, $\hat{\boldsymbol{\theta}}$,

$$\ln p(\mathbf{y} \,|\, m) = \ln \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m)\, p(\mathbf{y} \,|\, \boldsymbol{\theta}) \approx \ln p(\hat{\boldsymbol{\theta}} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) + \frac{1}{2} \ln \left| -2\pi H^{-1} \right| \;, \quad (21)$$

where $H(\hat{\boldsymbol{\theta}})$ is the Hessian defined as $\left. \frac{\partial^2 \ln p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. The BIC approximation is the asymptotic limit of the above expression, retaining only terms that grow with the number of data n; since the Hessian grows linearly with n, the BIC is given by

$$\ln p(\mathbf{y} \,|\, m)_{\mathrm{BIC}} = \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \frac{d(m)}{2} \ln n \;, \quad (22)$$

where $d(m)$ is the number of parameters in model m. The BIC is interesting because it does not depend on the prior over parameters, and is attractive because it does not involve the burdensome computation of the Hessian of the log likelihood and its determinant. However in general the Laplace approximation, and therefore its BIC limit, have several shortcomings which are outlined below.

The Gaussian assumption is based on the large data limit, and will represent the posterior poorly for small data sets for which, in principle, the advantages of Bayesian integration over ML or MAP are largest. The Gaussian approximation is also poorly suited to bounded, constrained, or positive parameters, since it assigns non-zero probability mass outside of the parameter domain. Moreover, the posterior may not be unimodal for likelihoods with hidden variables, due to problems of identifiability; in these cases the regularity conditions required for convergence do not hold. Even if the exact posterior is unimodal, the resulting approximation may well be a poor representation of the nearby probability *mass*, as the approximation is made about a locally maximum probability *density*. In large models the approximation may become unwieldy to compute, taking $\mathcal{O}(nd^2)$ operations to compute the derivatives in the Hessian, and then a further $\mathcal{O}(d^3)$ operations to calculate its determinant (d is the number of parameters in the model) — further approximations would become necessary, such as those ignoring off-diagonal elements or assuming a block-diagonal structure for the Hessian, which correspond to neglecting dependencies between parameters.

For BIC, we require the number of free parameters in each structure. In these experiments we use a simple counting argument and apply the following counting scheme. If a variable j has no parents in the DAG, then it contributes $(|\mathbf{z}_{ij}| - 1)$ free parameters, corresponding to the degrees of freedom in its vector of prior probabilities (constrained to lie on the simplex $\sum_k p_k = 1$). Each variable that has parents contributes $(|\mathbf{z}_{ij}| - 1)$ parameters for each configuration of its parents. Thus in model m the total number of parameters $d(m)$ is given by

$$d(m) = \sum_{j=1}^{|\mathbf{z}_i|} (|\mathbf{z}_{ij}| - 1) \prod_{l=1}^{|\mathbf{z}_{i\mathbf{pa}(j)}|} \left| \mathbf{z}_{i\mathbf{pa}(j)l} \right| \;, \quad (23)$$

where $\left| \mathbf{z}_{i\mathbf{pa}(j)l} \right|$ denotes the cardinality (number of settings) of the lth parent of the jth variable. We have used the convention that the product over zero factors has a value of

one to account for the case in which the jth variable has no parents — that is to say $\prod_{l=1}^{\left|\mathbf{z}_{i\mathbf{pa}(j)}\right|} \left|\mathbf{z}_{i\mathbf{pa}(j)l}\right| = 1$, if the number of parents $\left|\mathbf{z}_{i\mathbf{pa}(j)l}\right|$ is 0.

The BIC approximation needs to take into account aliasing in the parameter posterior. In discrete-variable DAGs, parameter aliasing occurs from two symmetries: first, a priori identical hidden variables can be permuted, and second, the labellings of the states of each hidden variable can be permuted. As an example, let us imagine the parents of a single observed variable are 3 hidden variables having cardinalities $(3, 3, 4)$. In this case the number of aliases is 1728 $(= 2! \times 3! \times 3! \times 4!)$. If we assume that the aliases of the posterior distribution are well separated then the score is given by

$$\ln p(\mathbf{y} \,|\, m)_{\mathrm{BIC}} = \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \frac{d(m)}{2} \ln n + \ln S \tag{24}$$

where S is the number of aliases, and $\hat{\boldsymbol{\theta}}$ is the MAP estimate as described in the previous section. This correction is accurate only if the modes of the posterior distribution are well separated, which should be the case in the large data set size limit for which BIC is useful. However, since BIC is correct only up to an indeterminate missing factor, we might think that this correction is not necessary. In the experiments we examine the BIC score with and without this correction, and also with and without the inclusion of the prior term $\ln p(\hat{\boldsymbol{\theta}} \,|\, m)$.

3.3 The Cheeseman-Stutz approximation

The Cheeseman-Stutz (CS) approximation makes use of the following identity for the incomplete-data marginal likelihood:

$$p(\mathbf{y} \,|\, m) = p(\hat{\mathbf{z}} \,|\, m) \frac{p(\mathbf{y} \,|\, m)}{p(\hat{\mathbf{z}} \,|\, m)} = p(\hat{\mathbf{z}} \,|\, m) \frac{\int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)}{\int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}' \,|\, m) p(\hat{\mathbf{z}} \,|\, \boldsymbol{\theta}', m)} \;, \tag{25}$$

which is true for any completion $\hat{\mathbf{z}} = \{\hat{\mathbf{s}}, \mathbf{y}\}$ of the data. This form is useful because the complete-data marginal likelihood, $p(\hat{\mathbf{z}} \,|\, m)$, is tractable to compute for discrete DAGs with independent Dirichlet priors: it is just a product of Dirichlet integrals, as given in (5). By applying Laplace approximations to the integrals in both the numerator and denominator, about points $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}'$ in parameter space respectively, and then assuming the limit of an infinite amount of data in order to recover BIC-type forms for both integrals, we immediately obtain the following estimate of the marginal (incomplete) likelihood

$$\begin{aligned} \ln p(\mathbf{y} \,|\, m) \approx \ln p(\mathbf{y} \,|\, m)_{\mathrm{CS}} \equiv \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) + \ln p(\hat{\boldsymbol{\theta}} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \frac{d}{2} \ln n \\ - \ln p(\hat{\boldsymbol{\theta}}' \,|\, m) - \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) + \frac{d'}{2} \ln n \qquad (26) \\ = \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) + \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) - \ln p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) \;. \qquad (27) \end{aligned}$$

The last line follows if we choose $\hat{\boldsymbol{\theta}}'$ to be identical to $\hat{\boldsymbol{\theta}}$ and further assume that the number of parameters in the models for complete and incomplete data are the same, i.e.

$d = d'$ (Cheeseman and Stutz 1996). In the case of the models examined in this article, we can ensure that the mode of the posterior in the complete setting is at locations $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}}$ by completing the hidden data $\{\mathbf{s}_i\}_{i=1}^n$ with their expectations under their posterior distributions $p(\mathbf{s}_i \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}})$, or simply: $\hat{\mathbf{s}}_{ijk} = \langle \delta(\mathbf{s}_{ij}, k) \rangle_{q_{\mathbf{s}_i}(\mathbf{s}_i)}$. This procedure will generally result in non-integer counts N_{jlk} on application of (19). Upon parameter re-estimation using equation (20), we note that $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}}$ remains invariant. The most important aspect of the CS approximation is that each term of (27) can be tractably evaluated as follows:

$$\text{from (5)} \qquad p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) = \prod_{j=1}^{|\mathbf{z}_i|} \prod_{l=1}^{|\mathbf{z}_{i\mathbf{pa}(j)}|} \frac{\Gamma(\lambda_{jl}^0)}{\Gamma(\lambda_{jl} + \hat{N}_{jl})} \prod_{k=1}^{|\mathbf{z}_{ij}|} \frac{\Gamma(\lambda_{jlk} + \hat{N}_{jlk})}{\Gamma(\lambda_{jlk})} \;; \tag{28}$$

$$\text{from (7)} \qquad p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} \prod_{l=1}^{|\mathbf{z}_{i\mathbf{pa}(j)}|} \prod_{k=1}^{|\mathbf{z}_{ij}|} \hat{\theta}_{jlk}^{\delta(\mathbf{z}_{ij},k)\delta(\mathbf{z}_{i\mathbf{pa}(j)},l)} \;; \tag{29}$$

$$\text{from (1)} \qquad p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \prod_{j=1}^{|\mathbf{z}_i|} \prod_{l=1}^{|\mathbf{z}_{i\mathbf{pa}(j)}|} \prod_{k=1}^{|\mathbf{z}_{ij}|} \hat{\theta}_{jlk}^{\hat{N}_{jlk}} \;, \tag{30}$$

where the $\hat{N}_{jlk}$ are identical to the N_{jlk} of equation (19) if the completion of the data with $\hat{\mathbf{s}}$ is done with the posterior found in the M step of the MAP EM algorithm used to find $\hat{\boldsymbol{\theta}}$. Equation (29) is simply the likelihood output by the EM algorithm, equation (28) is a function of the counts obtained in the EM algorithm, and equation (30) is a simple computation again.

As with BIC, the Cheeseman-Stutz score also needs to be corrected for aliases in the parameter posterior, and is subject to the same caveat that these corrections are only accurate if the aliases in the posterior are well separated. Finally, we note that CS is in fact a lower bound on the marginal likelihood, and is intricately related to our proposed method that is described next. In Section 6 we revisit the CS approximation and derive a key result on the tightness of its bound.

3.4 Estimating marginal likelihood using Variational Bayes

Here we briefly review a method of lower bounding the marginal likelihood, and the corresponding deterministic iterative algorithm for optimising this bound that has come to be known as *variational Bayes* (VB). Variational methods have been used in the past to tackle intractable posterior distributions over hidden variables (Neal 1992; Hinton and Zemel 1994; Saul and Jordan 1996; Jaakkola 1997; Ghahramani and Jordan 1997; Ghahramani and Hinton 2000), and more recently have tackled Bayesian learning in specific models (Hinton and van Camp 1993; Waterhouse et al. 1996; MacKay 1997; Bishop 1999; Ghahramani and Beal 2000). Inspired by MacKay (1997), Attias (2000) first described the general form of variational Bayes and showed that it is a generalisation of the celebrated EM algorithm of Dempster et al. (1977). Ghahramani and Beal (2001) and Beal (2003) built upon this work, applying it to the large class of conjugate-

exponential models (described below). Just as in the standard E step of EM, we obtain a posterior distribution over the hidden variables, and we now also treat the parameters of the model as uncertain quantities and infer their posterior distribution as well. Since the hidden variables and parameters are coupled, computing the exact posterior distribution over both is intractable and we use the variational methodology to instead work in the space of simpler distributions — those that are factorised between hidden variables and parameters.

As before, let $\mathbf{y}$ denote the observed variables, $\mathbf{x}$ denote the hidden variables, and $\boldsymbol{\theta}$ denote the parameters. We assume a prior distribution over parameters $p(\boldsymbol{\theta} \,|\, m)$, conditional on the model m. The marginal likelihood of a model, $p(\mathbf{y} \,|\, m)$, can be lower bounded by introducing any distribution over both latent variables and parameters which has support where $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ does, by appealing to Jensen's inequality:

$$\ln p(\mathbf{y} \,|\, m) = \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m) = \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} \tag{31}$$

$$\geq \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})} \;. \tag{32}$$

Maximising this lower bound with respect to the free distribution $q(\mathbf{x}, \boldsymbol{\theta})$ results in $q(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$, which when substituted above turns the inequality into an equality. This does not simplify the problem, since evaluating the exact posterior distribution $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ requires knowing its normalising constant, the marginal likelihood. Instead we constrain the posterior to be a simpler, factorised (separable) approximation $q(\mathbf{x}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, which we are at liberty to do as (32) is true for any $q(\mathbf{x}, \boldsymbol{\theta})$:

$$\ln p(\mathbf{y} \,|\, m) \geq \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{33}$$

$$= \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\int d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)}{q_{\mathbf{x}}(\mathbf{x})} + \ln \frac{p(\boldsymbol{\theta} \,|\, m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right] \tag{34}$$

$$= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \tag{35}$$

$$= \mathcal{F}_m(q_{\mathbf{x}_1}(\mathbf{x}_1), \ldots, q_{\mathbf{x}_n}(\mathbf{x}_n), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \;. \tag{36}$$

The last equality is a consequence of the data $\mathbf{y}$ being i.i.d. and is explained below. The quantity $\mathcal{F}_m$ is a functional of the free distributions, $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$; for brevity we omit the implicit dependence of $\mathcal{F}_m$ on the fixed data set $\mathbf{y}$.

The variational Bayesian algorithm iteratively maximises $\mathcal{F}_m$ in (35) with respect to the free distributions, $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, which is essentially coordinate ascent in the function space of variational distributions. It is not difficult to show that by taking functional derivatives of (35), we obtain the VBE and VBM update equations shown below. Each application of the VBE and VBM steps is guaranteed to increase or leave unchanged the lower bound on the marginal likelihood, and successive applications are guaranteed to converge to a local maximum of $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$. The VBE step is

$$\textbf{VBE step:} \quad q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = \frac{1}{\mathcal{Z}_{\mathbf{x}_i}} \exp \left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\; \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m) \right] \quad \forall\; i \;, \tag{37}$$

giving the hidden variable variational posterior

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) \ , \tag{38}$$

and the VBM step by

$$\textbf{VBM step:} \quad q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} p(\boldsymbol{\theta} \,|\, m) \, \exp\left[\int d\mathbf{x} \; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \, \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m) \right] \ . \tag{39}$$

Here t indexes the iteration number and $\mathcal{Z}_{\boldsymbol{\theta}}$ and $\mathcal{Z}_{\mathbf{x}_i}$ are (readily computable) normalisation constants. The factorisation of the distribution over hidden variables between different data points in (38) is a consequence of the i.i.d. data assumption, and falls out of the VB optimisation only because we have decoupled the distributions over hidden variables and parameters. At this point it is well worth noting the symmetry between the hidden variables and the parameters. The only distinguishing feature between hidden variables and parameters is that the number of hidden variables increases with data set size, whereas the number of parameters is assumed fixed.

Re-writing (33), it is easy to see that maximising $\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$ is equivalent to minimising the Kullback-Leibler (KL) divergence between $q_{\mathbf{x}}(\mathbf{x}) \, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and the joint posterior over hidden states and parameters $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$:

$$\ln p(\mathbf{y} \,|\, m) - \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta} \, d\mathbf{x} \; q_{\mathbf{x}}(\mathbf{x}) \; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \; \ln \frac{q_{\mathbf{x}}(\mathbf{x}) \; q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)} \tag{40}$$

$$= \mathrm{KL}\left[q_{\mathbf{x}}(\mathbf{x}) \, q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \,\|\, p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m) \right] \geq 0 \ . \tag{41}$$

The variational Bayesian EM algorithm reduces to the ordinary EM algorithm for ML estimation if we restrict the parameter distribution to a point estimate, i.e. a Dirac delta function, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, in which case the M step simply involves re-estimating $\boldsymbol{\theta}^*$. Note that the same cannot be said in the case of MAP estimation, which is inherently basis dependent, unlike both VB and ML algorithms. By construction, the VBEM algorithm is guaranteed to monotonically increase an objective function $\mathcal{F}_m$, as a function of distributions over parameters and hidden variables. Since we integrate over model parameters there is a naturally incorporated model complexity penalty. It turns out that, for a large class of models that we will examine next, the VBE step has approximately the same computational complexity as the standard E step in the ML framework, which makes it viable as a Bayesian replacement for the EM algorithm. Moreover, for a large class of models $p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$ that we call *conjugate-exponential* (CE) models, the VBE and VBM steps have very simple and intuitively appealing forms. We examine these points next. CE models satisfy two conditions:

Condition (1). *The complete-data likelihood is in the exponential family:*

$$p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) = g(\boldsymbol{\theta}) \, f(\mathbf{x}_i, \mathbf{y}_i) \, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} \ , \tag{42}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of natural parameters, $\mathbf{u}$ and f are the functions that define the exponential family, and g is a normalisation constant:

$$g(\boldsymbol{\theta})^{-1} = \int d\mathbf{x}_i \, d\mathbf{y}_i \; f(\mathbf{x}_i, \mathbf{y}_i) \, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} \; . \tag{43}$$

Condition (2). *The parameter prior is conjugate to the complete-data likelihood:*

$$p(\boldsymbol{\theta} \,|\, \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) \, g(\boldsymbol{\theta})^\eta \, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}} \; , \tag{44}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior, and h is a normalisation constant:

$$h(\eta, \boldsymbol{\nu})^{-1} = \int d\boldsymbol{\theta} \; g(\boldsymbol{\theta})^\eta \, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}} \; . \tag{45}$$

From the definition of conjugacy, we see that the hyperparameters of a conjugate prior can be interpreted as the number (η) and values ($\boldsymbol{\nu}$) of pseudo-observations under the corresponding likelihood. The list of latent-variable models of practical interest with complete-data likelihoods in the exponential family is very long, for example: Gaussian mixtures, factor analysis, principal components analysis, hidden Markov models and extensions, switching state-space models, discrete-variable belief networks. Of course there are also many as yet undreamt-of models combining Gaussian, gamma, Poisson, Dirichlet, Wishart, multinomial, and other distributions in the exponential family. Models whose complete-data likelihood is not in the exponential family can often be approximated by models which are in the exponential family and have been given additional hidden variables (for example, see Attias 1999a).

In Bayesian inference we want to determine the posterior over parameters and hidden variables $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, \eta, \boldsymbol{\nu})$. In general this posterior is *neither* conjugate nor in the exponential family, and is intractable to compute. We can use the variational Bayesian VBE (37) and VBM (39) update steps, but we have no guarantee that we will be able to represent the results of the integration and exponentiation steps analytically. Here we see how models with CE properties are especially amenable to the VB approximation, and derive the VBEM algorithm for CE models.

Given an i.i.d. data set $\mathbf{y} = \{\mathbf{y}_1, \ldots \mathbf{y}_n\}$, if the model satisfies conditions (1) and (2), then the following results (a), (b) and (c) hold.

(a) The VBE step yields:

$$q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}(\mathbf{x}_i) \; , \tag{46}$$

and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ is in the exponential family:

$$q_{\mathbf{x}_i}(\mathbf{x}_i) \propto f(\mathbf{x}_i, \mathbf{y}_i) \, e^{\overline{\boldsymbol{\phi}}^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)} = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \overline{\boldsymbol{\phi}}) \; , \tag{47}$$

with a natural parameter vector

$$\overline{\boldsymbol{\phi}} = \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\boldsymbol{\phi}(\boldsymbol{\theta}) \equiv \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{48}$$

obtained by taking the expectation of $\boldsymbol{\phi}(\boldsymbol{\theta})$ under $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ (denoted using angle-brackets $\langle\cdot\rangle$). For invertible $\boldsymbol{\phi}$, defining $\tilde{\boldsymbol{\theta}}$ such that $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}$, we can rewrite the approximate posterior as

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i \,|\, \mathbf{y}_i, \tilde{\boldsymbol{\theta}}) \;. \tag{49}$$

(b) The VBM step yields that $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is conjugate and of the form:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}})\, g(\boldsymbol{\theta})^{\tilde{\eta}}\, e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}} \;, \tag{50}$$

where $\tilde{\eta} = \eta + n$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^{n} \overline{\mathbf{u}}(\mathbf{y}_i)$, and $\overline{\mathbf{u}}(\mathbf{y}_i) = \langle \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \rangle_{q_{\mathbf{x}_i}(\mathbf{x}_i)}$ is the expectation of the sufficient statistic $\mathbf{u}$. We have used $\langle\cdot\rangle_{q_{\mathbf{x}_i}(\mathbf{x}_i)}$ to denote expectation under the variational posterior over the latent variable(s) associated with the ith datum.

(c) Results (a) and (b) hold for every iteration of variational Bayesian EM — i.e. the forms in (47) and (50) are *closed* under VBEM.

Results (a) and (b) follow from direct substitution of the forms in (42) and (44) into the VB update equations (37) and (39). Furthermore, if $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ are initialised according to (47) and (50), respectively, and conditions (42) and (44) are met, then result (c) follows by induction.

As before, since $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ are coupled, (50) and (47) do not provide an analytic solution to the minimisation problem, so the optimisation problem is solved numerically by iterating between these equations. To summarise, for CE models:

> **VBE Step**: *Compute the expected sufficient statistics* $\{\overline{\mathbf{u}}(\mathbf{y}_i)\}_{i=1}^{n}$ *under the hidden variable distributions* $q_{\mathbf{x}_i}(\mathbf{x}_i)$*, for all* i*.*
>
> **VBM Step**: *Compute the expected natural parameters* $\overline{\boldsymbol{\phi}} = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ *under the parameter distribution given by* $\tilde{\eta}$ *and* $\tilde{\boldsymbol{\nu}}$*.*

A major implication of these results for CE models is that, if there exists such a $\tilde{\boldsymbol{\theta}}$ satisfying $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}$, the posterior over hidden variables calculated in the VBE step is exactly the posterior that would be calculated had we been performing a standard E step using $\tilde{\boldsymbol{\theta}}$. That is, the inferences using an ensemble of models $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ can be represented by the effect of a point parameter, $\tilde{\boldsymbol{\theta}}$. The task of performing many inferences, each of which corresponds to a different parameter setting, can be replaced with a single inference step tractably.

We can draw a tight parallel between the EM algorithm for ML/MAP estimation, and our VBEM algorithm applied specifically to conjugate-exponential models. These

EM for MAP estimation	**Variational Bayesian EM**
Goal: maximise $p(\boldsymbol{\theta} \,|\, \mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$	**Goal:** lower bound $p(\mathbf{y} \,|\, m)$
E Step: compute $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}^{(t)})$	**VBE Step:** compute $q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x} \,|\, \mathbf{y}, \overline{\boldsymbol{\phi}}^{(t)})$
M Step: $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$	**VBM Step:** $q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp \int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$

Table 1: Comparison of EM for ML/MAP estimation with VBEM for CE models.

are summarised in table 1. This general result of VBEM for CE models was reported in Ghahramani and Beal (2001), and generalises the well known EM algorithm for ML estimation (Dempster et al. 1977). It is a special case of the variational Bayesian algorithm (equations (37) and (39)) used in Ghahramani and Beal (2000) and in Attias (2000), yet encompasses many of the models that have been so far subjected to the variational treatment. Its particular usefulness is as a guide for the design of models, to make them amenable to efficient approximate Bayesian inference.

The VBE step has about the same time complexity as the E step, and is in all ways identical except that it is re-written in terms of the expected natural parameters. The VBM step computes a *distribution* over parameters (in the conjugate family) rather than a point estimate. Both ML/MAP EM and VBEM algorithms monotonically increase an objective function, but the latter also incorporates a model complexity penalty by integrating over parameters and thereby embodying an Occam's razor effect.

3.5 The variational Bayesian lower bound for discrete-valued DAGs

We wish to approximate the incomplete-data log marginal likelihood (8) given by

$$\ln p(\mathbf{y} \,|\, m) = \ln \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta} \,|\, m) \prod_{i=1}^{n} \sum_{\{\mathbf{z}_{ij}\}_{j \in \mathcal{H}}} \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}) \;. \tag{51}$$

We can form the lower bound using (33), introducing variational distributions $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $\{q_{\mathbf{s}_i}(\mathbf{s}_i)\}_{i=1}^{n}$ to yield

$$\begin{aligned} \ln p(\mathbf{y} \,|\, m) &\geq \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta} \,|\, m)}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} + \sum_{i=1}^{n} \int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{z}_i \,|\, \boldsymbol{\theta}, m)}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \\ &= \mathcal{F}_m(q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), q(\mathbf{s})) \;. \end{aligned} \tag{52}$$

We now take functional derivatives to write down the variational Bayesian EM algorithm. The VBM step is straightforward:

$$\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta} \,|\, m) + \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{z}_i \,|\, \boldsymbol{\theta}, m) + \mathrm{c} \ , \tag{53}$$

with c a constant required for normalisation. Given that the prior over parameters factorises over variables, as in (3), and the complete-data likelihood factorises over the variables in a DAG, as in (1), equation (53) can be broken down into individual terms:

$$\ln q_{\boldsymbol{\theta}_{jl}}(\boldsymbol{\theta}_{jl}) = \ln p(\boldsymbol{\theta}_{jl} \,|\, \boldsymbol{\lambda}_{jl}, m) + \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \boldsymbol{\theta}, m) + c_{jl} \ , \tag{54}$$

where $\mathbf{z}_{ij}$ may be either a hidden or observed variable, and each c_{jl} is a Lagrange multiplier from which a normalisation constant is obtained. Since the prior is Dirichlet, it is easy to show that equation (54) has the form of the Dirichlet distribution — thus conforming to result (b) in (50). We define the expected counts under the hidden variable variational posterior distribution

$$N_{jlk} = \sum_{i=1}^{n} \left\langle \delta(\mathbf{z}_{ij}, k) \delta(\mathbf{z}_{i\mathbf{pa}(j)}, l) \right\rangle_{q_{\mathbf{s}_i}(\mathbf{s}_i)} \ . \tag{55}$$

That is, N_{jlk} is the expected total number of times the jth variable (hidden or observed) is in state k when its parents (hidden or observed) are in state l, where the expectation is taken with respect to the variational distribution $q_{\mathbf{s}_i}(\mathbf{s}_i)$ over the hidden variables. Then the variational posterior for the parameters is given simply by

$$q_{\boldsymbol{\theta}_{jl}}(\boldsymbol{\theta}_{jl}) = \mathrm{Dir}\left(\lambda_{jlk} + N_{jlk} \ : \ k = 1, \ldots, |\mathbf{z}_{ij}|\right) \ . \tag{56}$$

For the VBE step, taking derivatives of (52) with respect to each $q_{\mathbf{s}_i}(\mathbf{s}_i)$ yields

$$\ln q_{\mathbf{s}_i}(\mathbf{s}_i) = \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln p(\mathbf{z}_i \,|\, \boldsymbol{\theta}, m) + c'_i = \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m) + c'_i \ , \tag{57}$$

where each c'_i is a Lagrange multiplier for normalisation of the posterior. Since the complete-data likelihood $p(\mathbf{z}_i \,|\, \boldsymbol{\theta}, m)$ is in the exponential family and we have placed conjugate Dirichlet priors on the parameters, we can immediately utilise the result in (49) which gives simple forms for the VBE step:

$$q_{\mathbf{s}_i}(\mathbf{s}_i) \propto q_{\mathbf{z}_i}(\mathbf{z}_i) = \prod_{j=1}^{|\mathbf{z}_i|} p(\mathbf{z}_{ij} \,|\, \mathbf{z}_{i\mathbf{pa}(j)}, \tilde{\boldsymbol{\theta}}) \ . \tag{58}$$

Thus the approximate posterior over the hidden variables $\mathbf{s}_i$ resulting from a variational Bayesian approximation is identical to that resulting from exact inference in a model with known point parameters $\tilde{\boldsymbol{\theta}}$ — the choice of $\tilde{\boldsymbol{\theta}}$ must satisfy $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}$. The natural parameters for this model are the log probabilities $\{\ln \boldsymbol{\theta}_{jlk}\}$, where j specifies which

variable, l indexes the possible configurations of its parents, and k the possible settings of the variable. Thus

$$\ln \tilde{\boldsymbol{\theta}}_{jlk} = \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}_{jlk}) = \overline{\boldsymbol{\phi}}_{jlk} = \int d\boldsymbol{\theta}_{jl}\; q_{\boldsymbol{\theta}_{jl}}(\boldsymbol{\theta}_{jl}) \ln \boldsymbol{\theta}_{jlk} \;. \tag{59}$$

Under a Dirichlet distribution, the expectations are differences of digamma functions

$$\ln \tilde{\boldsymbol{\theta}}_{jlk} = \psi(\lambda_{jlk} + N_{jlk}) - \psi(\sum_{k=1}^{|\mathbf{z}_{ij}|} \lambda_{jlk} + N_{jlk}) \quad \forall \; \{j, l, k\} \;, \tag{60}$$

where the N_{jlk} are defined in (55), and $\psi(\cdot)$ is the digamma function. Since this expectation operation takes the geometric mean of the probabilities, the propagation algorithm in the VBE step is now passed sub-normalised probabilities as parameters: $\sum_{k=1}^{|\mathbf{z}_{ij}|} \tilde{\boldsymbol{\theta}}_{jlk} \leq 1 \;\forall\; \{j, l\}$. This use of sub-normalised probabilities also occurred in MacKay (1997) in the context of variational Bayesian Hidden Markov Models.

The expected natural parameters become normalised only if the distribution over parameters is a delta function, in which case this reduces to the MAP inference scenario of Section 3.1. In fact, using the property of the digamma function for large arguments, $\lim_{x\to\infty} \psi(x) = \ln x$, we find that equation (60) becomes

$$\lim_{n\to\infty} \; \ln \tilde{\boldsymbol{\theta}}_{jlk} = \ln(\lambda_{jlk} + N_{jlk}) - \ln(\sum_{k=1}^{|\mathbf{z}_{ij}|} \lambda_{jlk} + N_{jlk}) \;, \tag{61}$$

which has recovered the MAP estimator for $\boldsymbol{\theta}$ (20), up to the -1 entries in numerator and denominator which become vanishingly small for large data, and vanish completely if MAP is performed in the softmax parameterisation (see MacKay 1998). Thus in the limit of large data VB recovers the MAP parameter estimate.

To summarise, the VBEM implementation for discrete DAGs consists of iterating between the VBE step (58) which infers distributions over the hidden variables given a distribution over the parameters, and a VBM step (56) which finds a variational posterior distribution over parameters based on the hidden variables' sufficient statistics from the VBE step. Each step monotonically increases or leaves unchanged a lower bound on the marginal likelihood of the data, and the algorithm is guaranteed to converge to a local maximum of the lower bound. The VBEM algorithm uses as a subroutine the algorithm used in the E step of the corresponding EM algorithm, and so the VBE step's computational complexity is the same as for EM — there is some overhead in calculating differences of digamma functions instead of ratios of expected counts, but this is presumed to be minimal and fixed. As with BIC and Cheeseman-Stutz, the lower bound does not take into account aliasing in the parameter posterior, and needs to be corrected as described in Section 3.2.

3.6 Annealed Importance Sampling (AIS)

AIS (Neal 2001) is a state-of-the-art technique for estimating marginal likelihoods, which breaks a difficult integral into a series of easier ones. It combines techniques from importance sampling, Markov chain Monte Carlo, and simulated annealing (Kirkpatrick et al. 1983). AIS builds on work in the Physics community for estimating the free energy of systems at different temperatures, for example: thermodynamic integration (Neal 1993), *tempered transitions* (Neal 1996), and the similarly inspired *umbrella sampling* (Torrie and Valleau 1977). Most of these, as well as other related methods, are reviewed in Gelman and Meng (1998).

Obtaining samples from the posterior distribution over parameters, with a view to forming a Monte Carlo estimate of the marginal likelihood of the model, is usually a very challenging problem. This is because, even with small data sets and models with just a few parameters, the distribution is likely to be very peaky and have its mass concentrated in tiny volumes of space. This makes simple approaches such as sampling parameters directly from the prior or using simple importance sampling infeasible. The basic idea behind annealed importance sampling is to move in a *chain* from an easy-to-sample-from distribution, via a series of intermediate distributions, through to the complicated posterior distribution. By annealing the distributions in this way the parameter samples should hopefully come from representative areas of probability mass in the posterior. The key to the annealed importance sampling procedure is to make use of the importance weights gathered at all the distributions up to and including the final posterior distribution, in such a way that the final estimate of the marginal likelihood is unbiased. A brief description of the AIS procedure follows. We define a series of inverse-temperatures $\{\tau(k)\}_{k=0}^{K}$ satisfying

$$0 = \tau(0) < \tau(1) < \cdots < \tau(K-1) < \tau(K) = 1 \ . \tag{62}$$

We refer to temperatures and inverse-temperatures interchangeably throughout this section. We define the function:

$$f_k(\boldsymbol{\theta}) \equiv p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)^{\tau(k)} \ , \qquad k \in \{0, \ldots, K\} \ . \tag{63}$$

Thus the set of functions $\{f_k(\boldsymbol{\theta})\}_{k=0}^{K}$ form a series of unnormalised distributions which *interpolate* between the prior and posterior, parameterised by τ. We also define the normalisation constants $\mathcal{Z}_k \equiv \int d\boldsymbol{\theta} \ f_k(\boldsymbol{\theta})$, and note that $\mathcal{Z}_0 = \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta} \,|\, m) = 1$ from normalisation of the prior and that $\mathcal{Z}_K = \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m) = p(\mathbf{y} \,|\, m)$, which is exactly the marginal likelihood we wish to estimate. We can estimate $\mathcal{Z}_K$, or equivalently $\frac{\mathcal{Z}_K}{\mathcal{Z}_0}$, using the identity

$$p(\mathbf{y} \,|\, m) = \frac{\mathcal{Z}_K}{\mathcal{Z}_0} \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_0} \ \frac{\mathcal{Z}_2}{\mathcal{Z}_1} \ \cdots \ \frac{\mathcal{Z}_K}{\mathcal{Z}_{K-1}} = \prod_{k=1}^{K} \mathcal{R}_k \ . \tag{64}$$

Each of the K ratios in this expression can be individually estimated without bias using importance sampling: the kth ratio, denoted $\mathcal{R}_k$, can be estimated from a set of (not necessarily independent) samples of parameters $\{\boldsymbol{\theta}^{(k,c)}\}_{c \in \mathcal{C}_k}$, which are drawn from

the higher temperature $\tau(k-1)$ distribution (the importance distribution) — i.e. each $\boldsymbol{\theta}^{(k,c)} \sim f_{k-1}(\boldsymbol{\theta})$, and the importance weights are computed at the lower temperature $\tau(k)$. These samples are used to construct the Monte Carlo estimate for $\mathcal{R}_k$:

$$\mathcal{R}_k \equiv \frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}} = \int d\boldsymbol{\theta} \; \frac{f_k(\boldsymbol{\theta})}{f_{k-1}(\boldsymbol{\theta})} \frac{f_{k-1}(\boldsymbol{\theta})}{\mathcal{Z}_{k-1}} \approx \frac{1}{C_k} \sum_{c \in \mathcal{C}_k} \frac{f_k(\boldsymbol{\theta}^{(k,c)})}{f_{k-1}(\boldsymbol{\theta}^{(k,c)})} \;, \quad \text{with } \boldsymbol{\theta}^{(k,c)} \sim f_{k-1}(\boldsymbol{\theta}) \tag{65}$$

$$\hat{\mathcal{R}}_k = \frac{1}{C_k} \sum_{c \in \mathcal{C}_k} p(\mathbf{y} \,|\, \boldsymbol{\theta}^{(k,c)}, m)^{\tau(k)-\tau(k-1)} \;. \tag{66}$$

The variance of the estimate of each $\mathcal{R}_k$ depends on the constituent distributions $\{f_k(\boldsymbol{\theta}), f_{k-1}(\boldsymbol{\theta})\}$ being sufficiently close so as to produce low-variance weights (the summands in (65)). Neal (2001) shows that it is a sufficient condition that the C_k each be chosen to be exactly one for the product of the $c\hat{R}_k$ estimators to be an unbiased estimate of the marginal likelihood, $p(\mathbf{y} \,|\, m)$, in (64). It is an open research question as to whether values of $c\hat{R}_k$ can be shown to lead to an unbiased estimate. In our experiments (Section 4) we use $C_k = 1$ and so remain in the realm of an unbiased estimator.

Metropolis-Hastings for discrete-variable models

In general we expect it to be difficult to sample directly from the forms $f_k(\boldsymbol{\theta})$ in (63), and so Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970) steps are used at each temperature to generate the set of C_k samples required for each importance calculation in (66). In the discrete-variable graphical models covered in this article, the parameters are multinomial probabilities, hence the support of the Metropolis proposal distributions is restricted to the simplex of probabilities summing to 1. One might suggest using a Gaussian proposal distribution in the softmax parameterisation, $\mathbf{b}$, of the current parameters $\boldsymbol{\theta}$, like so: $\theta_i \equiv e^{b_i} / \sum_j^{|\boldsymbol{\theta}|} e^{b_j}$. However the Jacobian of the transformation from this vector $\mathbf{b}$ back to the vector $\boldsymbol{\theta}$ is zero, and it is hard to construct a reversible Markov chain.

A different and intuitively appealing idea is to use a Dirichlet distribution as the proposal distribution, with its mean positioned at the current parameter. The precision of the Dirichlet proposal distribution at inverse-temperature $\tau(k)$ is governed by its *strength*, $\alpha(k)$, which is a free variable to be set as we wish, provided it is not in any way a function of the sampled parameters. An MH acceptance function is required to maintain detailed balance: if $\boldsymbol{\theta}'$ is the sample under the proposal distribution centered around the current parameter $\boldsymbol{\theta}^{(k,c)}$, then the acceptance function is:

$$a(\boldsymbol{\theta}', \boldsymbol{\theta}^{(k,c)}) = \min \left(\frac{f_k(\boldsymbol{\theta}')}{f_k(\boldsymbol{\theta}^{(k,c)})} \frac{\mathrm{Dir}(\boldsymbol{\theta}^{(k,c)} \,|\, \boldsymbol{\theta}', \alpha(k))}{\mathrm{Dir}(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}^{(k,c)}, \alpha(k))} \;,\; 1 \right) \;, \tag{67}$$

where $\mathrm{Dir}(\boldsymbol{\theta} \,|\, \overline{\boldsymbol{\theta}}, \alpha)$ is the probability density of a Dirichlet distribution with mean $\overline{\boldsymbol{\theta}}$ and

strength α, evaluated at $\boldsymbol{\theta}$. The next sample is instantiated as follows:

$$\boldsymbol{\theta}^{(k,c+1)} = \begin{cases} \boldsymbol{\theta}' & \text{if } w < a(\boldsymbol{\theta}', \boldsymbol{\theta}^{(k,c)}) \quad \text{(accept)} \\ \boldsymbol{\theta}^{(k,c)} & \text{otherwise} \quad \text{(reject)} \ , \end{cases} \tag{68}$$

where $w \sim \mathrm{U}(0,1)$ is a random variable sampled from a uniform distribution on $[0,1]$. By repeating this procedure of accepting or rejecting $C'_k \geq C_k$ times at the temperature $\tau(k)$, the MCMC sampler generates a set of (dependent) samples $\{\boldsymbol{\theta}^{(k,c)}\}_{c=1}^{C'_k}$. A subset of these $\{\boldsymbol{\theta}^{(k,c)}\}_{c\in\mathcal{C}_k}$, with $|\mathcal{C}_k| = C_k \leq C'_k$, is then used as the importance samples in the computation above (66). This subset will generally not include the first few samples, as these samples are likely not yet samples from the equilibrium distribution at that temperature, and in the case of $C_k = 1$ will contain only the most recent sample.

An algorithm to compute all ratios

The entire algorithm for computing all K marginal likelihood ratios is given in Algorithm 3.1. It has several parameters, in particular: the number of annealing steps, K; their

1. Initialise $\boldsymbol{\theta}_{\mathrm{ini}} \sim f_0(\boldsymbol{\theta})$ i.e. from the prior $p(\boldsymbol{\theta} \,|\, m)$
2. For $k = 1$ to K annealing steps
   (a) Run MCMC at temperature $\tau(k-1)$ as follows:
      i. Initialise $\boldsymbol{\theta}^{(k,0)} \leftarrow \boldsymbol{\theta}_{\mathrm{ini}}$ from previous temp.
      ii. Generate the set $\{\boldsymbol{\theta}^{(k,c)}\}_{c=1}^{C'_k} \sim f_{k-1}(\boldsymbol{\theta})$ as follows:
         A. For $c = 1$ to C'_k
            Propose $\boldsymbol{\theta}' \sim \mathrm{Dir}(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}^{(k,c-1)}, \alpha(k))$
            Accept $\boldsymbol{\theta}^{(k,c)} \leftarrow \boldsymbol{\theta}'$ according to (67) and (68)
         End For
         B. Store $\boldsymbol{\theta}_{\mathrm{ini}} \leftarrow \boldsymbol{\theta}^{(k,C'_k)}$
      iii. Store a subset of these $\{\boldsymbol{\theta}^{(k,c)}\}_{c\in\mathcal{C}_k}$ with $|\mathcal{C}_k| = C_k \leq C'_k$
   (b) Calculate $\mathcal{R}_k \equiv \frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}} \approxeq \frac{1}{C_k} \sum_{c=1}^{C_k} \frac{f_k(\boldsymbol{\theta}^{(k,c)})}{f_{k-1}(\boldsymbol{\theta}^{(k,c)})}$
   End For
3. Output $\{\ln \mathcal{R}_k\}_{k=1}^K$ and $\ln \hat{\mathcal{Z}}_K = \sum_{k=1}^K \ln \mathcal{R}_k$ as the approximation to $\ln \mathcal{Z}_K$

Algorithm 3.1: AIS. Computes ratios $\{\mathcal{R}_k\}_{k=1}^K$ for the marginal likelihood estimate.

inverse-temperatures (the annealing schedule), $\{\tau(k)\}_{k=1}^K$; the parameters of the MCMC importance sampler at each temperature $\{C'_k, C_k, \alpha(k)\}_{k=1}^K$, which are the number of proposed samples, the number used for the importance estimate, and the precision of

the proposal distribution, respectively. We remind the reader that the estimate has only been proven to be unbiased in the case of $C_k = 1$.

Algorithm 3.1 produces only a single estimate of the marginal likelihood; a particular attraction of AIS is that one can take averages of estimates from a number G of AIS runs to form another unbiased estimate of the marginal likelihood with lower variance: $[\mathcal{Z}_K/\mathcal{Z}_0]^{(G)} = G^{-1}\sum_{g=1}^{G}\prod_{k=1}^{K^{(g)}}\mathcal{R}_k^{(g)}$. In Section 5 we discuss the performance of AIS for estimating the marginal likelihood of the graphical models used in this article, addressing the specific choices of proposal widths, number of samples, and annealing schedules used in the experiments.

4 Experiments

In this section we experimentally examine the accuracy of each of the scoring methods described in the previous section. To this end, we first describe the class defining our space of hypothesised structures, then choose a particular member of the class as the "true" structure; we generate a set of parameters for that structure, and then generate varying-sized data sets from that structure with those parameters. Each score is then used to estimate the marginal likelihoods of each structure in the class, for each possible data set size. From these estimates, a posterior distribution over the structures can be computed for each data set size. Our goal is to assess how closely these approximate distributions reflect the true posterior distributions. We use three metrics: i) the rank given to the true structure (i.e. the modal structure has rank 1); ii) the difference between the estimated marginal likelihoods of the top-ranked and true structures; iii) The Kullback-Leibler divergence of the approximate to the true posterior distributions.

A specific class of graphical model: We examine discrete directed bipartite graphical models, i.e. those graphs in which only hidden variables can be parents of observed variables, and the hidden variables themselves have no parents. For our in depth study we restrict ourselves to graphs which have just $k = |\mathcal{H}| = 2$ hidden variables, and $p = |\mathcal{V}| = 4$ observed variables; both hidden variables are binary i.e. $|\mathbf{s}_{ij}| = 2$ for $j \in \mathcal{H}$, and each observed variable has cardinality $|\mathbf{y}_{ij}| = 5$ for $j \in \mathcal{V}$.

The number of distinct graphs: In the class of graphs described above, with k distinct hidden variables and p observed variables, there are 2^{kp} possible structures, corresponding to the presence or absence of a directed link between each hidden and each conditionally independent observed variable. If the hidden variables are unidentifiable, which is the case in our example model where they have the same cardinality, then the number of possible graphs is reduced due to permutation symmetries. It is straightforward to show in this example that the number of distinct graphs is reduced from $2^{2\times 4} = 256$ down to 136.

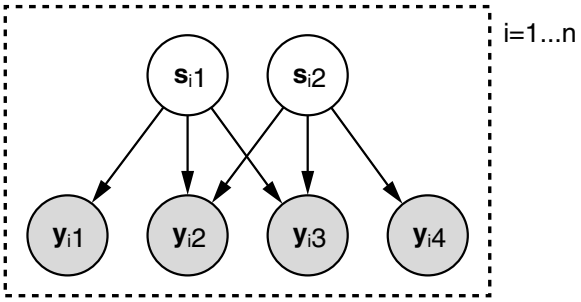


Figure 1: The true structure that was used to generate all the data sets used in the experiments. The hidden variables (top) are each binary, and the observed variables (bottom) are each five-valued. This structure has 50 parameters, and is two links away from the fully-connected structure. In total, there are 136 possible distinct structures with two (identical) hidden variables and four observed variables.

The specific model and generating data: We chose the particular structure shown in Figure 1, which we call the "true" structure. This structure contains enough links to induce non-trivial correlations amongst the observed variables, whilst the class as a whole has few enough nodes to allow us to examine exhaustively every possible structure of the class. There are only three other structures in the class which have more parameters than our chosen structure: These are: two structures in which either the left- or right-most visible node has both hidden variables as parents instead of just one, and one structure which is fully connected. One should note that our chosen true structure is at the higher end of complexity in this class, and so we might find that scoring methods that do not penalise complexity do seemingly better than naively expected.

Evaluation of the marginal likelihood of all possible alternative structures in the class is done for academic interest only, since for non-trivial numbers of variables the number of structures is huge. In practice one can embed different structure scoring methods in a greedy model search outer loop (for example, see Friedman 1998) to find probable structures. Here, we are not so much concerned with structure *search* per se, since a prerequisite for a good structure search algorithm is an efficient and accurate method for evaluating any particular structure. Our aim in these experiments is to establish the reliability of the variational bound as a score, compared to annealed importance sampling, and the currently employed asymptotic scores such as the BIC and Cheeseman-Stutz criteria.

The parameters of the true model: Conjugate uniform symmetric Dirichlet priors were placed over all the parameters of the model — that is to say $\lambda_{jlk} = 1 \ \forall\{jlk\}$ in (3). This particular prior was arbitrarily chosen for the purposes of the experiments; we do not expect it to influence the trends in our conclusions. For the network shown in Figure 1, parameters were sampled from the prior, once and for all, to instantiate a true underlying model, from which data was then generated. The sampled parameters are

shown below (their sizes are functions of each node's and its parents' cardinalities):

$$\boldsymbol{\theta}_1 = \begin{bmatrix} .12 & .88 \end{bmatrix} \quad \boldsymbol{\theta}_3 = \begin{bmatrix} .03 & .03. & .64 & .02 & .27 \\ .18 & .15 & .22 & .19 & .27 \end{bmatrix} \quad \boldsymbol{\theta}_6 = \begin{bmatrix} .10 & .08 & .43 & .03 & .36 \\ .30 & .14 & .07 & .04 & .45 \end{bmatrix}$$

$$\boldsymbol{\theta}_2 = \begin{bmatrix} .08 & .92 \end{bmatrix} \quad \boldsymbol{\theta}_4 = \begin{bmatrix} .10 & .54 & .07 & .14 & .15 \\ .04 & .15 & .59 & .05 & .16 \\ .20 & .08 & .36 & .17 & .18 \\ .19 & .45 & .10 & .09 & .17 \end{bmatrix} \quad \boldsymbol{\theta}_5 = \begin{bmatrix} .11 & .47 & .12 & .30 & .01 \\ .27 & .07 & .16 & .25 & .25 \\ .52 & .14 & .15 & .02 & .17 \\ .04 & .00 & .37 & .33 & .25 \end{bmatrix},$$

where $\{\boldsymbol{\theta}_j\}_{j=1}^{2}$ are the parameters for the hidden variables, and $\{\boldsymbol{\theta}_j\}_{j=3}^{6}$ are the parameters for the remaining four observed variables, $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{i4}$. Each row of each matrix denotes the probability of each multinomial setting for a particular configuration of the parents, and sums to one (up to rounding error). Note that there are only two rows for $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_6$, as both these observed variables have just a single binary parent. For variables $\mathbf{y}i2$ and $\mathbf{y}i3$, the four rows correspond to the parent configurations: {[1 1], [1 2], [2 1], [2 2]} (with parameters $\boldsymbol{\theta}_5$ and $\boldsymbol{\theta}_6$ respectively). In this particular instantiation of the parameters, both the hidden variable priors are close to deterministic, causing approximately 80% of the data to originate from the [2 2] setting of the hidden variables. This means that we may need many data points before the evidence for two hidden variables outweighs that for one.

Incrementally larger nested data sets were generated from these parameter settings, with $n \in \{10, 20, 40, 80, 110, 160, 230, 320, 400, 430, 480, 560, 640, 800, 960, 1120, 1280, 2560, 5120, 10240\}$. The items in the $n = 10$ data set are a subset of the $n = 20$ and subsequent data sets, etc. The particular values of n were chosen from an initially exponentially increasing data set size, followed by inclusion of some intermediate data sizes to concentrate on interesting regions of behaviour.

4.1 Comparison of scores to AIS

All 136 possible distinct structures were scored for each of the 20 data set sizes given above, using MAP, BIC, CS, VB and AIS scores. We ran EM on each structure to compute the MAP estimate of the parameters (Section 3.1), and from it computed the BIC score (Section 3.2). Even though the MAP probability of the data is not strictly an approximation to the marginal likelihood, it can be shown to be an upper bound and so we include it for comparison. We also computed the BIC score including the parameter prior, denoted BICp, which was obtained by including a term $\ln p(\hat{\boldsymbol{\theta}} \,|\, m)$ in equation (24). From the same EM optimisation we computed the CS score (Section 3.3). We then ran the variational Bayesian EM algorithm with the same initial conditions to give a lower bound on the marginal likelihood (Section 3.5). To avoid local optima, several optimisations were carried out with different parameter initialisations, drawn each time from the prior over parameters. The same initial parameters were used for both EM and VBEM; in the case of VBEM the following protocol was used to obtain a parameter distribution: a conventional E-step was performed at the initial parameter to obtain

$\{p(\mathbf{s}_i|\mathbf{y}_i, \boldsymbol{\theta})\}_{i=1}^n$, which was then used in place of $q_{\mathbf{s}}(\mathbf{s})$ for input to the VBM step, which was thereafter followed by VBE and VBM iterations until convergence. The highest score over three random initialisations was taken for each algorithm; empirically this heuristic appeared to avoid local maxima problems. The EM and VBEM algorithms were terminated after either 1000 iterations had been reached, or the change in log likelihood (or lower bound on the log marginal likelihood, in the case of VBEM) became less than 10^{-6} per datum.

For comparison, the AIS sampler was used to estimate the marginal likelihood (see Section 3.6), annealing from the prior to the posterior in $K = 16384$ steps. A nonlinear annealing schedule was employed, tuned to reduce the variance in the estimate, and the Metropolis proposal width was tuned to give reasonable acceptance rates. We chose to have just a single sampling step at each temperature (i.e. $C'_k = C_k = 1$), for which AIS has been proven to give unbiased estimates, and initialised the sampler at each temperature with the parameter sample from the previous temperature. These particular choices are explained and discussed in detail in Section 5. Initial marginal likelihood estimates from single runs of AIS were quite variable, and for this reason several more batches of AIS runs were undertaken, each using a different random initialisation (and random numbers thereafter); the total of G batches of estimates were averaged according to the procedure given at the end of Section 3.6 to give the $\mathrm{AIS}^{(G)}$ score. In total, $G = 26$ batches of AIS runs were carried out.

Scoring all possible structures

Figure 2 shows the MAP, BIC, BICp, CS, VB and $\mathrm{AIS}^{(26)}$ scores obtained for each of the 136 possible structures against the number of parameters in the structure. Score is measured on the vertical axis, with each scoring method (columns) sharing the same vertical axis range for a particular data set size (rows). The horizontal axis of each plot corresponds to the number of parameters in the structure (as described in Section 3.2). For example, at the extremes there is one structure with 66 parameters (the fully connected structure) and one structure with 18 parameters (the fully unconnected structure). The structure that generated the data has exactly 50 parameters. In each plot we can see that several structures can occupy the same column, having the same number of parameters. This means that, at least visually, it is not always possible to unambiguously assign each point in the column to a particular structure.

The scores shown are those corrected for aliases (see equation (24)). Plots for uncorrected scores are almost identical. In each plot, the true structure is highlighted by a '$\circ$' symbol, and the structure currently ranked highest by that scoring method is marked with a '$\times$'. We can see the general upward trend for the MAP score, which prefers more complicated structures, and the pronounced downward trend for the BIC and BICp scores, which (over-)penalise structure complexity. In addition, one can see that neither upward or downward trends are apparent for VB or AIS scores. The CS score tends to show a downward trend similar to BIC and BICp, and while this trend weakens with increasing data, it is still present at $n = 10240$ (bottom row). Although not verifiable from these plots, the vast majority of the scored structures and data set

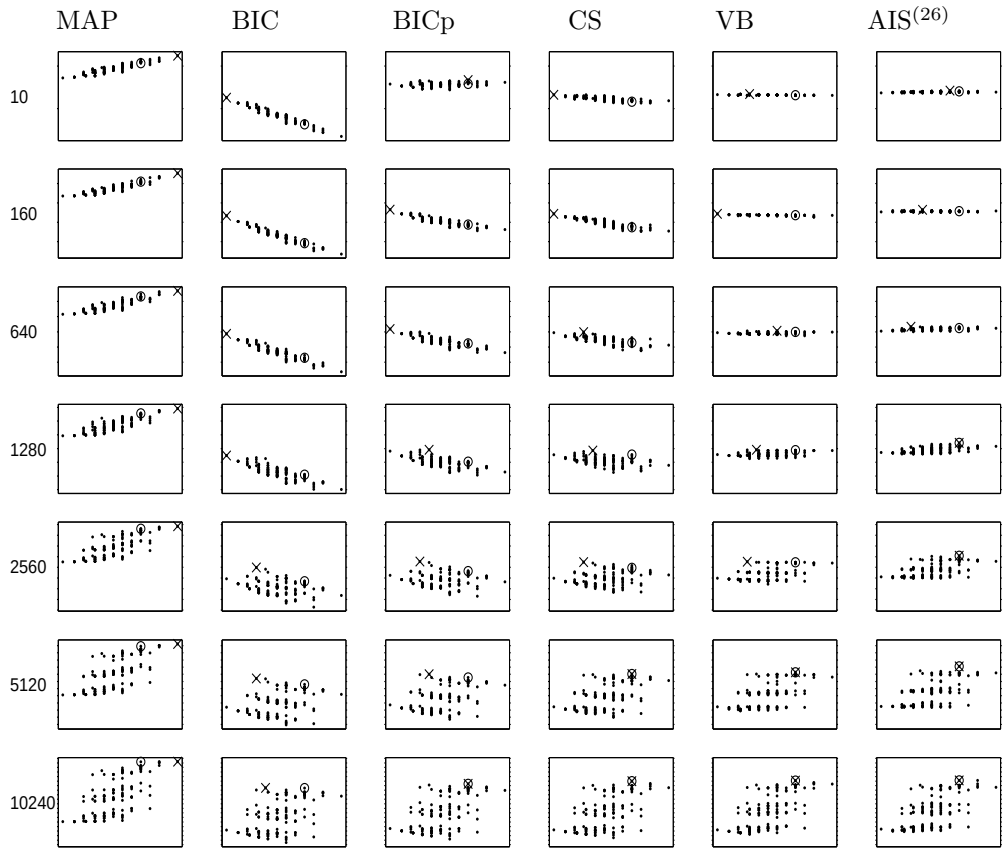


Figure 2: Scores for all 136 of the structures in the model class, by each of six scoring methods. Each plot has the score (approximation to the log marginal likelihood) on the vertical axis, with tick marks every 40 nats, and the number of parameters on the horizontal axis (ranging from 18 to 66). The middle four scores have been corrected for aliases (see Section 3.2). Each row corresponds to a data set of a different size, n: from top to bottom we have $n = 10, 160, 640, 1280, 2560, 5120, 10240$. The true structure is denoted with a '$\circ$' symbol, and the highest scoring structure in each plot marked by the '$\times$' symbol. Every plot in the same row has the same scaling for the vertical score axis, set to encapsulate every structure for all scores.

sizes, the AIS$^{(26)}$ score is higher than the VB lower bound, as we would expect.

The plots for large n show a distinct horizontal banding of the scores into three levels; this is an interesting artifact of the particular model used to generate the data. For example, we find on closer inspection some strictly followed trends: all those model structures residing in the upper band have the first three observable variables ($j = 3, 4, 5$) governed by at least one of the hidden variables; all those structures in the middle band have the third observable ($j = 5$) connected to at least one hidden variable.

In this particular example, AIS finds the correct structure at $n = 960$ data points, but unfortunately does not retain this result reliably until $n = 2560$. At $n = 10240$ data points, BICp, CS, VB and AIS all report the true structure as being the one with the highest score amongst the other contending structures. Interestingly, BIC still does not select the correct structure, and MAP has given a structure with sub-maximal parameters the highest score, which may well be due to local maxima in the EM optimisation.

Ranking of the true structure

Table 2 shows the ranking of the true structure, as it sits amongst all the possible 136 structures, as measured by each of the scoring methods MAP, BIC, BICp, CS, VB and AIS$^{(26)}$; this is also plotted in Figure 3(a), where the MAP ranking is not included for clarity. Higher positions in the plot correspond to better rankings: a ranking of 1 means that the scoring method has given the highest marginal likelihood to the true structure. We should keep in mind that, at least for small data set sizes, there is no reason to assume that the true posterior over structures has the true structure at its mode. Therefore we should not expect high rankings at small data set sizes.

For small n, for the most part the AIS score produces different (higher) rankings for the true structure than do the other scoring methods. We expect AIS to perform accurately with small data set sizes, for which the posterior distribution over parameters is not at all peaky, and so this suggests that the other approximations are performing poorly in comparison. However, for almost all n, VB outperforms BIC, BICp and CS, consistently giving a higher ranking to the true structure. Of particular note is the stability of the VB score ranking with respect to increasing amounts of data as compared to AIS (and to some extent CS). Columns in Table 2 with asterisks (*) correspond to scores that are not corrected for aliases, and are omitted from Figure 3(a). These corrections assume that the posterior aliases are well separated, and are valid only for large amounts of data and/or strongly-determined parameters. The correction nowhere degrades the rankings of any score, and in fact improves them very slightly for CS, and especially so for VB.

KL divergence of the methods' posterior distributions from the AIS estimate

In Figure 3(c) we plot the Kullback-Leibler (KL) divergence between the AIS computed posterior and the posterior distribution computed by each of the approximations BIC,

n	MAP	BIC*	BICp*	CS*	VB*	BIC	BICp	CS	VB	AIS$^{(26)}$
10	21	127	55	129	122	127	50	129	115	20
20	12	118	64	111	124	118	64	111	124	92
40	28	127	124	107	113	127	124	107	113	17
80	8	114	99	78	116	114	99	78	116	28
110	8	109	103	98	114	109	103	98	113	6
160	13	119	111	114	83	119	111	114	81	49
230	8	105	93	88	54	105	93	88	54	85
320	8	111	101	90	44	111	101	90	33	32
400	6	101	72	77	15	101	72	77	15	22
430	7	104	78	68	15	104	78	68	14	14
480	7	102	92	80	55	102	92	80	44	12
560	9	108	98	96	34	108	98	96	31	5
640	7	104	97	105	19	104	97	105	17	28
800	9	107	102	108	35	107	102	108	26	49
960	13	112	107	76	16	112	107	76	13	1
1120	8	105	96	103	12	105	96	103	12	1
1280	7	90	59	8	3	90	59	6	3	1
2560	6	25	17	11	11	25	15	11	11	1
5120	5	6	5	1	1	6	5	1	1	1
10240	3	2	1	1	1	2	1	1	1	1

Table 2: Ranking of the true structure by each of the scoring methods, as the size of the data set is increased. Asterisks (*) denote scores uncorrected for parameter aliasing in the posterior. These results are from data generated from only one instance of parameters under the true structure's prior over parameters.

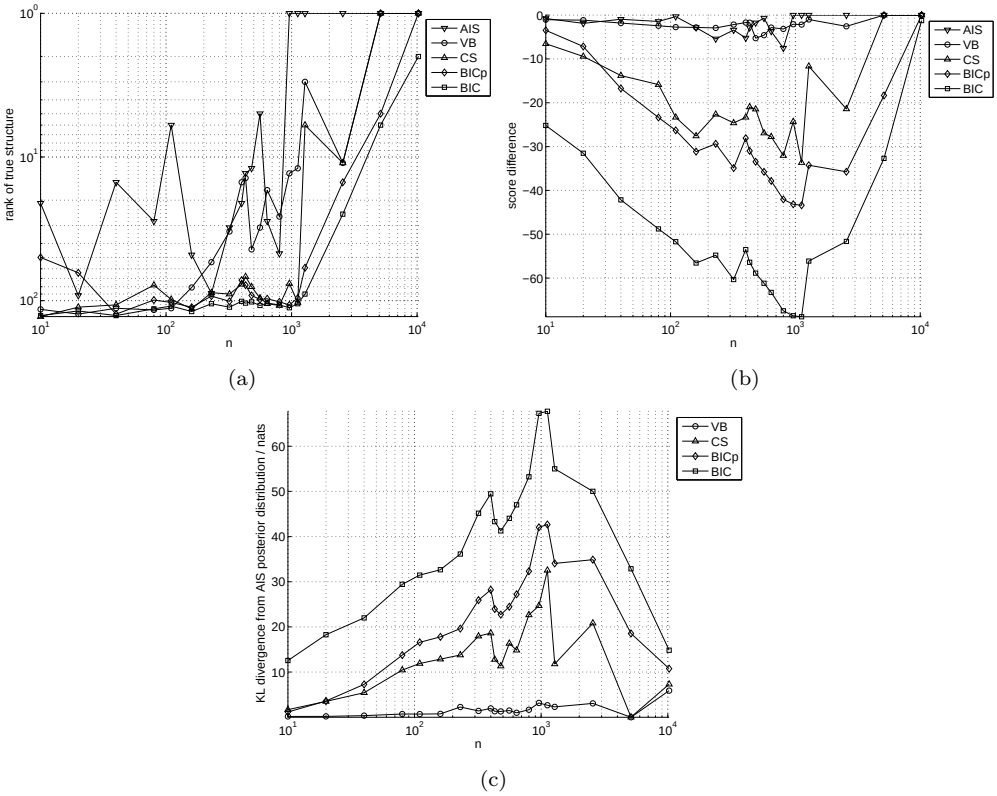


Figure 3: **(a)** Ranking given to the true structure by each scoring method for varying data set sizes (higher in plot is better), by BIC, BICp, CS, VB and AIS$^{(26)}$ methods. **(b)** Differences in log marginal likelihood estimates (scores) between the top-ranked structure and the true structure, as reported by each method. All differences are exactly zero or negative (see text). Note that these score differences are not normalised for the number of data n. **(c)** KL divergences of the approximate posterior distributions from the estimate of the posterior distribution provided by the AIS method; this measure is zero if the distributions are identical.

BICp, CS, and VB. We see quite clearly that VB has the lowest KL by a long way out of all the approximations, over a wide range of data set sizes, suggesting it is remaining most faithful to the true posterior distribution as approximated by AIS. The increase of the KL for the CS and VB methods at $n = 10240$ is almost certainly due to the AIS sampler having difficulty at high n (discussed below in Section 5), and should not be interpreted as a degradation in performance of either the CS or VB methods. It is interesting to note that the BIC, BICp, and CS approximations require a vast amount of data before their KL divergences reduce to the level of VB.

Computation Time

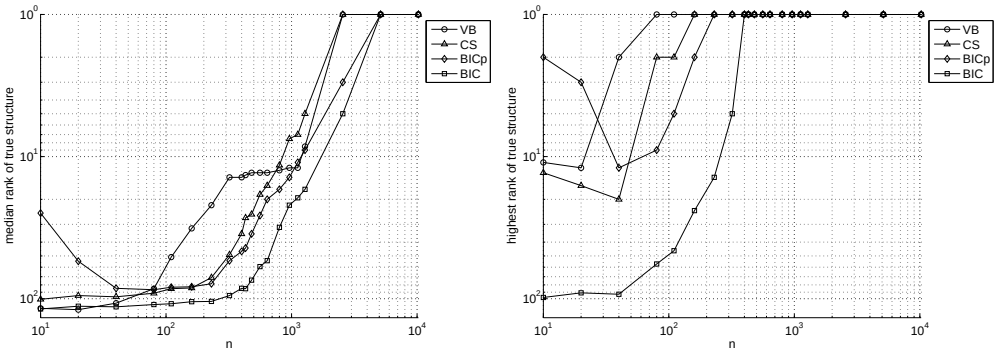
Scoring all 136 structures at 480 data points on a 1GHz Pentium III processor, with an implementation in MATLAB, took: 200 seconds for the MAP EM algorithms required for BIC/BICp/CS, 575 seconds for the VBEM algorithm required for VB, and 55000 seconds (15 hours) for a single set of runs of the AIS algorithm (using 16384 samples as in the main experiments); note the results for AIS here used averages of 26 runs. The massive computational burden of the sampling method (approx 75 hours for just 1 of 26 runs) makes CS and VB attractive alternatives for consideration.

4.2 Performance averaged over the parameter prior

The experiments in the previous section used a single instance of sampled parameters for the true structure, and generated data from this particular model. The reason for this was that, even for a single experiment, computing an exhaustive set of AIS scores covering all data set sizes and possible model structures takes in excess of 15 CPU days.

In this section we compare the performance of the scores over many different sampled parameters of the true structure (shown in Figure 1). 106 parameters were sampled from the prior and incremental data sets generated for each of these instances as the true model. MAP EM and VBEM algorithms were employed to calculate the scores as described in Section 4.1. For each instance of the true model, calculating scores for all data set sizes used and all possible structures, using three random restarts, for BIC/BICp/CS and VB took approximately 2.4 and 4.2 hours respectively on an Athlon 1800 Processor machine, which corresponds to about 1.1 and 1.9 seconds for each individual score.

Figure 4(a) shows the median ranking given to the true structure by each scoring method, computed over the 106 randomly sampled parameter settings. This plot corresponds to a smoothed version of Figure 3(a), but unfortunately cannot contain AIS averages for the computational reasons mentioned above. For the most part VB outperforms the other scores, although there is a region in which VB seems to underperform CS, as measured by this median score. For several cases, the VBEM optimisation reached the maximum number of allowed iterations before it had converged, whereas EM always converged. Allowing longer runs should result in improved VB performance. The VB score of the true structure is generally much closer to that of the top-ranked



(a) (b)

Figure 4: **(a)** Median ranking of the true structure as reported by BIC, BICp, CS and VB methods, against the size of the data set n, taken over 106 instances of the true structure. **(b)** The highest ranking given to the true structure under BIC, BICp, CS and VB methods, against the size of the data set n, taken over 106 instances of the true structure.

structure than is the case for any of the other scores. Figure 4(b) shows the best performance of the BIC, BICp, CS and VB methods over the 106 parameter instances in terms of the rankings. Lastly, we can examine the success rate of each score at picking the correct structure: Figure 5 shows the fraction of times that the true structure is ranked top by the different scoring methods, and other measures of performance.

5 AIS analysis, limitations, and extensions

The technique of annealed importance sampling is currently regarded as a state-of-the-art method for estimating the marginal likelihood in discrete-variable directed acyclic graphical models. In this section the AIS method is critically examined to gauge its reliability as a tool for judging the performance of the BIC, CS and VB scores.

The implementation of AIS has considerable flexibility: the user must specify the length, granularity and shape of the annealing schedules, the form of the Metropolis-Hastings (MH) sampling procedure, the number of samples taken at each temperature, etc. These and other parameters were described in Section 3.6; here we clarify our choices of settings and discuss some further ways in which the sampler could be improved.

How can we be sure that the AIS sampler is reporting the correct answer for the marginal likelihood of each structure? To be sure of a correct answer, one should use as long and gradual an annealing schedule as possible, containing as many samples at each temperature as is computationally viable. In the AIS experiments in this article, we always opted for a single sample at each step of the annealing schedule, initialising

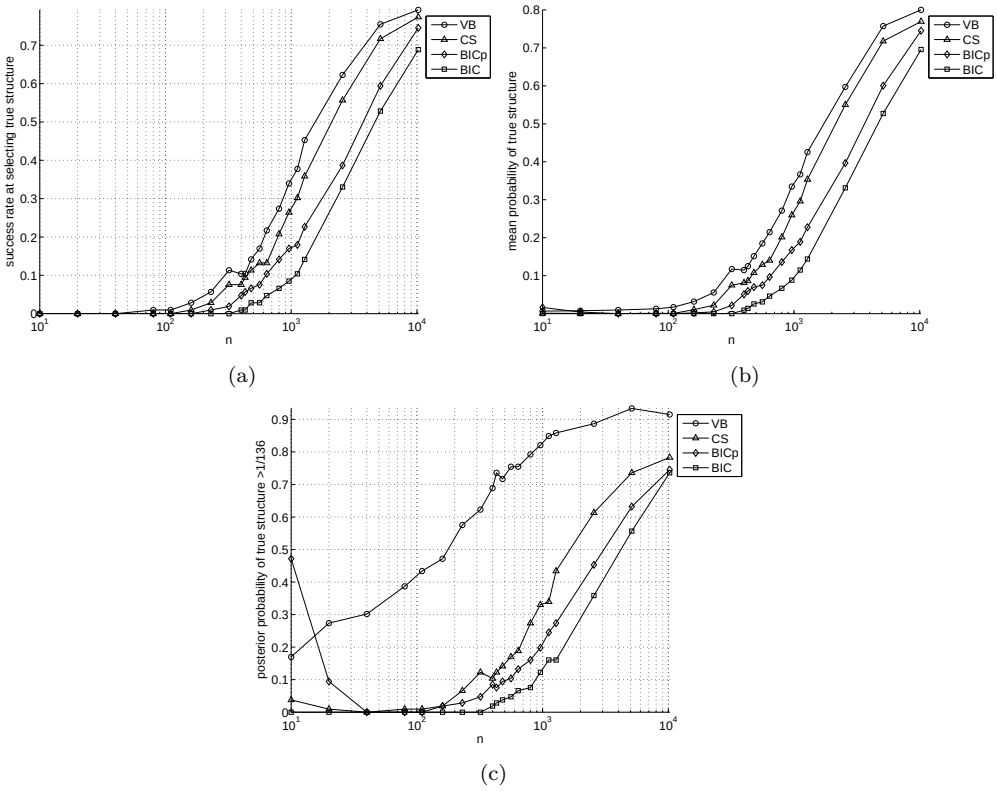


Figure 5: **(a)** The success rate of the scoring methods BIC, BICp, CS and VB, as measured by the fraction of 106 trials in which the true structure was given ranking 1 amongst the 136 candidate structures, plotted as a function of the data set size. **(b)** The posterior probability of the true structure, averaged over 106 trials. Note that for each trial, a posterior probability of greater that .5 is sufficient to guarantee that the true structure ranks top. **(c)** The fraction of trials in which the true structure was given posterior probability $> \frac{1}{136}$, i.e. greater than uniform probability.

the parameter at the next temperature at the previous sample, and ensured that the schedule itself was as finely grained as we could afford. This reduces the variables at our disposal to a single parameter, namely the total number of samples taken in each run of AIS, which is then directly related to the schedule granularity.

We examine the performance of the AIS sampler as a function of the number of samples. Figure 6(a) shows several AIS estimates of the marginal likelihood for the data set of size $n = 480$ under the model having the true structure. Each trace corresponds to a different point of initialisation of the AIS algorithm, obtained by sampling a parameter from the prior using 10 different random seeds. The top-most trace is initialised at the true parameters (which we as the experimenter have access to). Each point on a trace corresponds to a different temperature schedule for the AIS sampler with that initialising seed. Thus, a point at the right of the plot with high K corresponds to a schedule with many small steps in temperature, whereas a point at the left with low K corresponds to a coarser temperature schedule. Also plotted for reference are the VB and BIC estimates of the log marginal likelihood for this data set under the true structure, which are not functions of the annealing schedule. We know that the VB score is a lower bound on the log marginal likelihood, and so those estimates from AIS that consistently fall below this score must be indicative of an inadequate annealing schedule shape, duration and/or MH design.

For short annealing schedules, which are necessarily coarse to satisfy the boundary requirements on τ in equation (62), it is clear that the AIS sampling is badly underestimating the log marginal likelihood. The rapid annealing schedule does not give the sampler time to locate and exploit regions of high posterior probability, forcing it to neglect representative volumes of the posterior mass. Conversely, the AIS run started from the true parameters over-estimates the marginal likelihood, because it is prevented from exploring regions of low probability. Thus, for coarse schedules of less than about $K = 1000$ samples, the AIS estimate of the log marginal likelihood seems biased and has very high variance. Note that the AIS algorithm gives unbiased estimates of the marginal likelihood, but not necessarily the log marginal likelihood.

We see that all runs converge for sufficiently long annealing schedules, with AIS passing the BIC score at about 1000 samples, and the VB lower bound at about 5000 samples. Thus, loosely speaking, where the AIS and VB scores intersect we can consider their estimates to be roughly equally reliable. At $n = 480$ the VB scoring method requires about $1.5s$ to score the structure, whereas AIS at $n = 480$ and $K = 2^{13}$ requires about $100s$. Thus for this scenario, VB is 70 times more efficient at scoring the structures (at its own reliability).

In this article's main experiments, a value of $K = 2^{14} = 16384$ steps was used, and it is clear from Figure 6(a) that we can be fairly sure of the AIS method reporting a reasonably accurate result at this value of K, at least for $n = 480$. However, how would we expect these plots to look for larger data sets in which the posterior over parameters is more peaky and potentially more difficult to navigate during the annealing?

A good indicator of the mobility of the MH sampler is the acceptance rate of proposed samples. Figure 6(b) shows the fraction of accepted proposals during the annealing run,

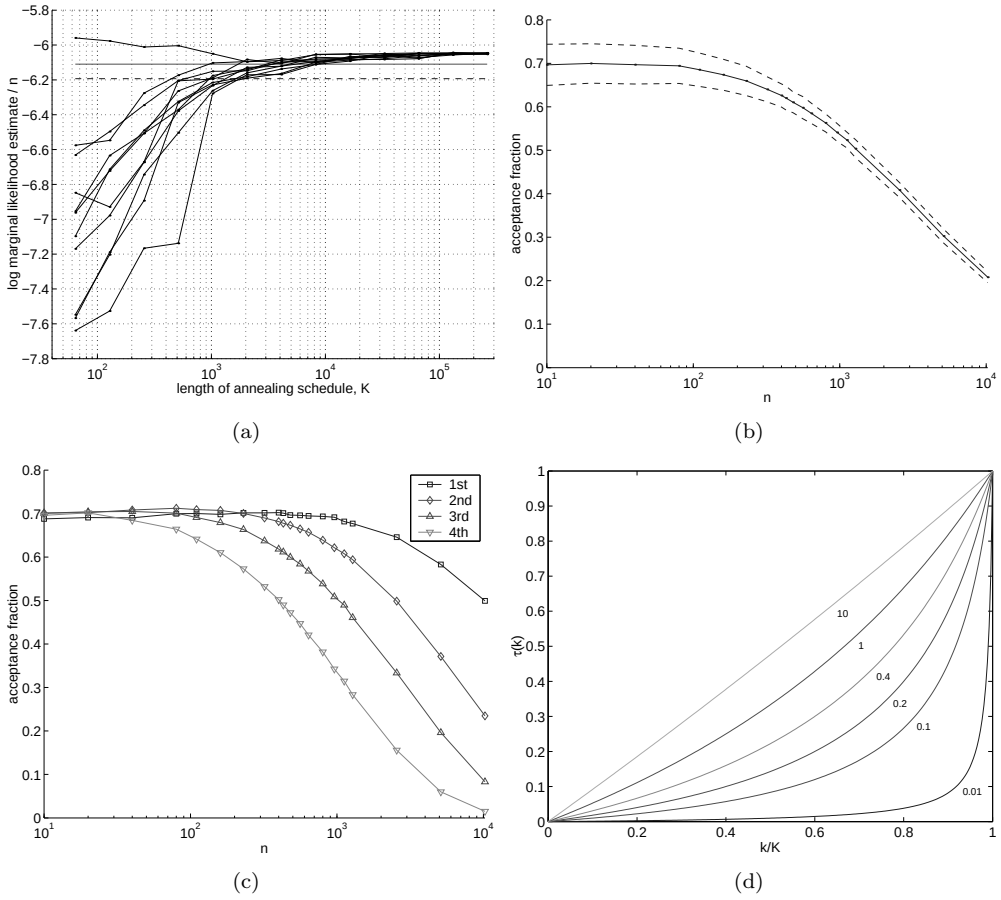


Figure 6: **(a)** Logarithm of AIS estimates (vertical) of the marginal likelihood for different initial conditions of the sampler (different traces) and different duration of annealing schedules (horizontal), for the true structure with $n = 480$ data points. The top-most trace is that corresponding to setting the initial parameters to the true values that generated the data. Shown are also the BIC score (dashed) and the VB lower bound (solid). **(b)** Acceptance rates of the MH proposals along the entire annealing schedule, for one batch of AIS scoring of all structures, against the size of the data set, n. The dotted lines are the sample standard deviations across all structures for each n. **(c)** Acceptance rates of the MH proposals for each of four quarters of the annealing schedule, for one batch of AIS scoring of all structures, against the size of the data set, n. Standard errors of the means are omitted for clarity. **(d)** Non-linear AIS annealing schedules, plotted for six different values of e_τ. In the experiments performed in this article, $e_\tau = 0.2$.

n	10...560	640	800	960	1120	1280	2560	5120	10240
single									
# AIS$^{(1)}$ < VB*	≤5.7	12.3	8.5	12.3	10.4	17.0	25.5	53.8	71.7
# AIS$^{(1)}$ < VB	≤7.5	15.1	9.4	14.2	12.3	20.8	31.1	59.4	74.5
% M-H rej.	<40.3	41.5	43.7	45.9	47.7	49.6	59.2	69.7	79.2
averaged									
# AIS$^{(5)}$ < VB*	0	0.0	0.0	0.0	0.0	0.7	3.7	13.2	50.0
# AIS$^{(5)}$ < VB	≤1.9	0.0	0.0	0.0	1.5	2.2	5.1	19.9	52.9

Table 3: AIS violations: for each size data set, n, we show the percentage of times, over the 136 structures, that a particular *single* AIS run reports marginal likelihoods below the VB lower bound. These are given for the VB scores that are uncorrected (*) and corrected for aliases. Also shown are the average percentage rejection rates of the MH sampler used to gather samples for the AIS estimates. The bottom half of the table shows the similar violations by the AIS score made from averaging the estimates of marginal likelihoods from five separate runs of AIS (see Section 3.6).

averaged over AIS scoring of all 136 possible structures, plotted against the size of the data set, n; the error bars are the standard errors of the mean acceptance rate across scoring all structures. We can see that at $n = 480$, the acceptance rate is rarely below 60%, and so one would indeed expect to see the sort of convergence shown in Figure 6(a). However, for the larger data sets the acceptance rate drops to 20%, implying that the sampler is having difficulty obtaining representative samples from the posterior distributions in the annealing schedule. Fortunately this drop is only linear in the logarithm of the data size.

By examining the reported AIS scores, both for single and pooled runs, over the 136 structures and 20 data set sizes, and comparing them to the VB lower bound, we can see how often AIS violates the lower bound. Table 3 compares the number of times the reported AIS scores AIS$^{(1)}$ and AIS$^{(5)}$ are below the VB lower bound, along with the rejection rates of the MH sampler that were plotted in Figure 6(b) (not a function of G). From the table we see that for small data sets, the AIS method reports "valid" results and the MH sampler is accepting a reasonable proportion of proposed parameter samples. However, at and beyond $n = 560$, the AIS sampler degrades to the point where it reports "invalid" results for more than half the 136 structures it scores. However, since the AIS estimate is noisy and we know that the tightness of the VB lower bound increases with n, this criticism could be considered too harsh — indeed if the bound were tight, we would expect the AIS score to violate the bound on roughly 50% of the runs anyway. The lower half of the table shows that, by combining AIS estimates from separate runs, we obtain an estimate that violates the VB lower bound far less often, and as expected we see the 50% violation rate for large amounts of data. This is a very useful result, and obviates to some extent the MH sampler's deficiency in all five runs. Diagnostically speaking, this analysis is good example of the use of readily-computable VB lower bounds for evaluating the reliability of the AIS method

post hoc.

Let us return to examining why the sampler is troubled for large data set sizes. Figure 6(c) shows the fraction of accepted MH proposals during each of four quarters of the annealing schedule used in the experiments. The rejection rate tends to increase moving from the beginning of the schedule (the prior) to the end (the posterior), the degradation becoming more pronounced for large data sets. This is most probably due to the proposal width remaining unchanged throughout all the AIS implementations; ideally, one would use a predetermined sequence of proposal widths which would be a function of the amount of data, n, and the position along the schedule.

We can use a heuristic argument to roughly predict the optimal proposal width to use for the AIS method. From mathematical arguments the precision of the posterior distribution over parameters is approximately proportional to the size of the data set n. Furthermore, the distribution being sampled from at step k of the AIS schedule is effectively that resulting from a fraction $\tau(k)$ of the data. Therefore, these two factors imply that the width of the MH proposal distribution should be inversely proportional to $\sqrt{n\tau(k)}$. In the case of multinomial variables, since the variance of a Dirichlet distribution is approximately inversely proportional to the strength α, then the optimal strength of the proposal distribution should be $\alpha_{opt} \propto n\tau(k)$ if its precision is to match the posterior precision. Note that we are at liberty to set these proposal precisions arbitrarily beforehand without causing the sampler to become biased.

We have not yet discussed the shape of the annealing schedule: should the inverse-temperatures $\{\tau(k)\}_{k=1}^{K}$ change linearly from 0 to 1, or follow some other function? The particular annealing schedule in these experiments was chosen to be nonlinear, lingering at higher temperatures for longer than at lower temperatures, according to

$$\tau(k) = \frac{e_\tau k/K}{1 - k/K + e_\tau} \qquad k \in \{0, \ldots, K\} \; . \tag{69}$$

For any setting of $e_\tau > 0$, the series of temperatures is monotonic and the initial and final temperatures satisfy (62): $\tau(0) = 0$, and $\tau(K) = 1$. For large e_τ the schedule becomes linear, and is plotted for different values of e_τ in Figure 6(d). A setting of $e_\tau = 0.2$ was found to reduce the degree of hysteresis in the annealing ratios.

6 Comparison to Cheeseman-Stutz (CS) approximation

In this section we present two important theoretical results regarding the approximation of Cheeseman and Stutz (1996), covered in Section 3.3. We briefly review the CS approximation, as used to approximate the marginal likelihood of finite mixture models, and then show that it is in fact a lower bound on the marginal likelihood in the case of mixture models (Minka 2001), and that similar CS constructions can be made for any model containing hidden variables. This observation brings CS into the family of bounding approximations, of which VB is a member. We then conclude the section by presenting a construction that proves that VB can be used to obtain a bound that is *always* tighter than CS.

Let m be a directed acyclic graph with parameters $\boldsymbol{\theta}$ giving rise to an i.i.d. data set denoted by $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, with corresponding discrete hidden variables $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ each of cardinality k. Let $\hat{\boldsymbol{\theta}}$ be a result of an EM algorithm which has converged to a local maximum in the likelihood $p(\mathbf{y} \,|\, \boldsymbol{\theta})$, and let $\hat{\mathbf{s}} = \{\hat{\mathbf{s}}_i\}_{i=1}^n$ be a completion of the hidden variables, chosen according to the posterior distribution over hidden variables given the data and $\hat{\boldsymbol{\theta}}$, such that $\hat{\mathbf{s}}_{ij} = p(\mathbf{s}_{ij} = j \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \;\forall\; i = 1, \ldots, n$.

Since we are completing the hidden variables with real, as opposed to discrete values, this complete data set does not in general correspond to a realisable data set under the generative model. This point raises the question of how its marginal probability $p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m)$ is defined. We will see in the following theorem and proof (Theorem 6) that both the completion required of the hidden variables and the completed data marginal probability are well-defined, and follow from equations (77) and (78) below.

The CS approximation is given by

$$p(\mathbf{y} \,|\, m) \approx p(\mathbf{y} \,|\, m)_{\mathrm{CS}} = p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}})} \;. \tag{70}$$

The CS approximation exploits the fact that, for many models of interest, the first term on the right-hand side — the complete-data marginal likelihood — is tractable to compute (this is the case for discrete-variable directed acyclic graphs with Dirichlet priors, as explained in Section 3.3). The term in the numerator of the second term on the right-hand side is simply the likelihood, which is an output of the EM algorithm (as is the parameter estimate $\hat{\boldsymbol{\theta}}$), and the denominator is a straightforward calculation that involves no summations over hidden variables or integrations over parameters.

Theorem 6.1: (Cheeseman-Stutz is a lower bound) *Let $\hat{\boldsymbol{\theta}}$ be the result of the M step of EM, and let $\{p(\mathbf{s}_i \,|\, \mathbf{y}_i, \hat{\boldsymbol{\theta}})\}_{i=1}^n$ be the set of posterior distributions over the hidden variables obtained in the next E step of EM. Furthermore, let $\hat{\mathbf{s}} = \{\hat{\mathbf{s}}_i\}_{i=1}^n$ be a completion of the hidden variables, such that $\hat{\mathbf{s}}_{ij} = p(\mathbf{s}_{ij} = j \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \;\forall\; i = 1, \ldots, n$. Then the CS approximation is a lower bound on the marginal likelihood:*

$$p(\mathbf{y} \,|\, m)_{CS} = p(\hat{\mathbf{s}}, \mathbf{y} \,|\, m) \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{p(\hat{\mathbf{s}}, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}})} \leq p(\mathbf{y} \,|\, m) \;. \tag{71}$$

Minka (2001) previously observed that in the specific case of mixture models, the Cheeseman-Stutz criterion is a lower bound on the marginal likelihood, and this could explain the reports of good performance in the literature (Cheeseman and Stutz 1996; Chickering and Heckerman 1997). Our contribution here is a proof of the result given in (71) that is generally applicable to any model with hidden variables, by using marginal likelihood bounds with approximations over the posterior distribution of the hidden variables only. We follow this with a corollary that allows us to always improve on the CS bound using VB.

Proof of Theorem 6.1: The marginal likelihood can be lower bounded by introducing

a distribution over the settings of each data point's hidden variables $q_{\mathbf{s}_i}(s_i)$:

$$p(\mathbf{y} \,|\, m) = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\theta}) \geq \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} . \tag{72}$$

We place a similar lower bound over the likelihood

$$p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}}) \geq \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} , \tag{73}$$

which can be made an equality if, for each data point, $q(\mathbf{s}_i)$ is set to the exact posterior distribution given the parameter setting $\boldsymbol{\theta}$,

$$p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}})}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} , \tag{74}$$

where

$$\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \equiv p(\mathbf{s}_i \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) , \tag{75}$$

which is the result obtained from an exact E step with the parameters set to $\hat{\boldsymbol{\theta}}$. Now rewrite the marginal likelihood bound (72), using this same choice of $\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)$, separate those terms that depend on $\boldsymbol{\theta}$ from those that do not, and substitute in to equation (74) to obtain:

$$p(\mathbf{y} \,|\, m) \geq \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{\hat{q}_{\mathbf{s}_i}(\mathbf{s}_i)} \right\} \cdot \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\} \tag{76}$$

$$= \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}}) \right\}} \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\} \tag{77}$$

$$= \frac{p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}})}{\prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i \,|\, \hat{\boldsymbol{\theta}})} \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\hat{\mathbf{s}}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) , \tag{78}$$

where $\hat{\mathbf{s}}_i$ are defined such that they satisfy:

$$\ln p(\hat{\mathbf{s}}_i, \mathbf{y} \,|\, \hat{\boldsymbol{\theta}}) = \sum_{\mathbf{s}_i} \hat{q}_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{s}_i} p(\mathbf{s}_i \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) , \tag{79}$$

where the second equality follows from the setting used in (75) that achieves a tight bound. The existence of such a completion follows from the fact that, in discrete-variable directed acyclic graphs of the sort considered in Chickering and Heckerman (1997), the hidden variables appear only linearly in logarithm of the joint probability $p(\mathbf{s}, \mathbf{y} \,|\, \boldsymbol{\theta})$.

Equation (78) is the Cheeseman-Stutz criterion of (27) and (71), and is also a lower bound on the marginal likelihood.

It is possible to derive CS-like approximations for types of graphical model other than discrete-variables DAGs. In the above proof, no constraints were placed on the forms of the joint distributions over hidden and observed variables, other than in the simplifying step in equation (78).

Finally, the following corollary gives some theoretical justification to the empirically observed superior performance of VB over CS. We present an original key result: that variational Bayes can always obtain a tighter bound than the Cheeseman-Stutz approximation.

Corollary 6.2: (VB is at least as tight as CS) *That is to say, it is always possible to find distributions* $q_{\mathbf{s}}(\mathbf{s})$ *and* $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *such that*

$$\ln p(\mathbf{y} \,|\, m)_{CS} \leq \mathcal{F}_m(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \leq \ln p(\mathbf{y} \,|\, m) \ . \tag{80}$$

Proof of Corollary 6.2: Consider the following forms for $q_{\mathbf{s}}(\mathbf{s})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$q_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} q_{\mathbf{s}_i}(\mathbf{s}_i) \ , \quad \text{with} \quad q_{\mathbf{s}_i}(\mathbf{s}_i) = p(\mathbf{s}_i \,|\, \mathbf{y}_i, \hat{\boldsymbol{\theta}}) \ , \tag{81}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \langle \ln p(\boldsymbol{\theta}) p(\mathbf{s}, \mathbf{y} \,|\, \boldsymbol{\theta}) \rangle_{q_{\mathbf{s}}(\mathbf{s})} \ . \tag{82}$$

We write the form for $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ explicitly:

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\}}{\int d\boldsymbol{\theta}' \ p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\}} \ , \tag{83}$$

and note that this is exactly the result of a VBM step. We substitute (83) directly into the VB lower bound stated in equation (33):

$$\mathcal{F}_m(q_{\mathbf{s}}(\mathbf{s}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta})}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta})}{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{84}$$

$$= \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \int d\boldsymbol{\theta} \ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \int d\boldsymbol{\theta}' \ p(\boldsymbol{\theta}') \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}') \right\} \tag{85}$$

$$= \sum_{i=1}^{n} \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln \frac{1}{q_{\mathbf{s}_i}(\mathbf{s}_i)} + \ln \int d\boldsymbol{\theta} \ p(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\left\{ \sum_{\mathbf{s}_i} q_{\mathbf{s}_i}(\mathbf{s}_i) \ln p(\mathbf{s}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}) \right\} \ , \tag{86}$$

which is exactly the logarithm of equation (76). And so with this choice of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{s}}(\mathbf{s})$, we achieve *equality* between the CS and VB approximations in (80). We complete the proof of Corollary 6.2 by noting that any further VB optimisation is guaranteed to increase or leave unchanged the lower bound, and hence surpass the CS lower bound. We would expect the VB lower bound starting from the CS solution to improve upon the CS lower bound in *all* cases, except in the very special case when the MAP parameter $\hat{\boldsymbol{\theta}}$ is exactly the *variational Bayes point*, defined as $\boldsymbol{\theta}_{\mathrm{BP}} \equiv \boldsymbol{\phi}^{-1}(\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})})$. Since VB is a lower bound on the marginal likelihood, the entire statement of (80) is proven.

7 Summary

In this paper we have presented various scoring methods for approximating the marginal likelihood of discrete directed graphical models with hidden variables. We presented EM algorithms for ML and MAP parameter estimation, showed how to calculate the asymptotic criteria of BIC and Cheeseman-Stutz, and derived the VBEM algorithm for approximate Bayesian learning which maintains distributions over the parameters of the model and has the same complexity as the EM algorithm. We also presented an Annealed Importance Sampling method designed for discrete-variable DAGs.

Our experiments show that VB consistently outperforms BIC and CS, and that VB performs, respectively, as well as and more reliably than AIS for intermediate and large sizes of data. The AIS method has many parameters to tune and requires knowledge of the model domain to design efficient and reliable sampling schemes and annealing schedules. VB, on the other hand, has not a single parameter to set or tune, and can be applied without any expert knowledge, at least in the class of singly-connected discrete-variable DAGs with Dirichlet priors which we have considered in this paper. Perhaps the most compelling evidence for the reliability of the VB approximation is given in Figure 3(c), which shows that the KL divergences of the VB-computed posterior distributions from the AIS standards are much smaller than for the other competing approximations.

It may be that there exists a better AIS scheme than sampling in parameter space. To be more specific, for any completion of the data the parameters of the model can be integrated out tractably (at least for the class of models examined in this chapter); thus an AIS scheme which anneals in the space of completions of the data may be more efficient than the current scheme which anneals in the space of parameters.[1] However, this latter scheme may only be efficient for models with little data compared to the number of parameters, as the sampling space of all completions increases linearly with the amount of data. This avenue of research is left to further work.

This paper has presented a novel application of variational Bayesian methods to discrete DAGs. In the literature there have been other attempts to solve this long-standing model selection problem in DAGs with hidden variables. For example, the *structural EM* algorithm of Friedman (1998) uses a structure search algorithm which uses a scoring algorithm very similar to the VBEM algorithm presented here, except

[1] personal communication with R. Neal

that for tractability, the distribution over $\boldsymbol{\theta}$ is replaced by the MAP estimate, $\boldsymbol{\theta}_{\mathrm{MAP}}$. We have shown here how the VB framework enables us to use the entire distribution over $\boldsymbol{\theta}$ for inference of the hidden variables. Very recently, Rusakov and Geiger (2005) have presented a modified BIC score that is asymptotically correct for the type of models we have examined in this article; future work will involve comparing VB to this modified BIC score in the non-asymptotic regime.

We have proved that the Cheeseman-Stutz score is a lower bound on the marginal likelihood in the case of general graphical models with hidden variables, extending the mixture model result of Minka (2001); and more importantly we proved that there exists a construction which is guaranteed to produce a variational Bayesian lower bound that is *at least as tight* as the Cheeseman-Stutz score (Corollary 6.2 to Theorem 6.1). This construction builds a variational Bayesian approximation using the same MAP parameter estimate used to obtain the CS score. However, we did not use this construction in our experiments, preferring the EM and VBEM algorithms to evolve separately (although similar parameter initialisations were employed for fairness). As a result we cannot guarantee that the VB bound is in all runs tighter than the CS bound, as the dynamics of the optimisations for MAP learning and VB learning may in general lead even identically initialised algorithms to different optima in parameter space (or parameter distribution space). Nevertheless, we have still seen improvement in terms of ranking of the true structure by VB as compared to CS. Empirically, the VB lower bound was observed to be *lower* than the CS score in only 173 of the 288320 total scores calculated (only about 0.06%), whereas had we used the construction, which we note is available to us for any graphical model with hidden variables, then this would have occurred exactly zero times.

Traditionally, the statistics community have concentrated on MCMC sampling and asymptotic criteria for computing marginal likelihoods for model selection and averaging. This article has applied the variational Bayes algorithm to scoring directed graphs and shown it to be empirically superior to existing criteria and, more importantly, theoretically superior to the popular Cheeseman-Stutz criterion. We believe that VB will prove to be of use in many other models, improving the efficiency of inference and model selection tasks without compromising accuracy.

References

Attias, H. (1999a). "Independent Factor Analysis." *Neural Computation*, 11: 803–851.

— (1999b). "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes." In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*.

— (2000). "A variational Bayesian framework for graphical models." In Solla, S. A., Leen, T. K., and Müller, K. (eds.), *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

Beal, M. J. (2003). "Variational Algorithms for Approximate Bayesian Inference." Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, UK.

Bishop, C. M. (1999). "Variational PCA." In *Proc. Ninth Int. Conf. on Artificial Neural Networks. ICANN*.

Cheeseman, P. and Stutz, J. (1996). "Bayesian Classification (Autoclass): Theory and Results." In Fayyad, U. M., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, 153–180. Menlo Park, CA: AAAI Press/MIT Press.

Chickering, D. M. and Heckerman, D. (1997). "Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables." *Machine Learning*, 29(2–3): 181–212.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B (Methodological)*, 39: 1–38.

Friedman, N. (1998). "The Bayesian structural EM algorithm." In *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98*. San Francisco, CA: Morgan Kaufmann Publishers.

Gelman, A. and Meng, X. (1998). "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical Science*, 13: 163–185.

Ghahramani, Z. and Beal, M. J. (2000). "Variational inference for Bayesian mixtures of factor analysers." In *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

— (2001). "Propagation algorithms for variational Bayesian learning." In *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press.

Ghahramani, Z. and Hinton, G. E. (2000). "Variational Learning for Switching State-Space Models." *Neural Computation*, 12(4).

Ghahramani, Z. and Jordan, M. I. (1997). "Factorial hidden Markov models." *Machine Learning*, 29: 245–273.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika*, 57(1): 97–109.

Heckerman, D. (1996). "A tutorial on learning with Bayesian networks." Technical Report MSR-TR-95-06 [ftp://ftp.research.microsoft.com/pub/tr/TR-95-06.PS] , Microsoft Research.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). "Learning Bayesian networks: the combination of knowledge and statistical data." *Machine Learning*, 20(3): 197–243.

Hinton, G. E. and van Camp, D. (1993). "Keeping Neural Networks Simple by Minimizing the Description Length of the Weights." In *Sixth ACM Conference on Computational Learning Theory, Santa Cruz*.

Hinton, G. E. and Zemel, R. S. (1994). "Autoencoders, Minimum Description Length, and Helmholtz Free Energy." In Cowan, J. D., Tesauro, G., and Alspector, J. (eds.), *Advances in Neural Information Processing Systems 6*. San Francisco, CA: Morgan Kaufmann.

Jaakkola, T. S. (1997). "Variational Methods for Inference and Estimation in Graphical Models." Technical Report Ph.D. Thesis, Department of Brain and Cognitive Sciences, MIT, Cambridge, MA.

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

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). "Optimization by simulated annealing." *Science*, 220(4598): 671–680.

MacKay, D. J. C. (1995). "Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks." *Network: Computation in Neural Systems*, 6: 469–505.

— (1997). "Ensemble Learning for Hidden Markov Models." Technical report, Cavendish Laboratory, University of Cambridge.

— (1998). "Choice of Basis for Laplace Approximation." *Machine Learning*, 33(1).

Metropolis, N., Rosenbluth, A. W., Teller, M. N., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *Journal of Chemical Physics*, 21: 1087–1092.

Minka, T. P. (2001). "Using lower bounds to approximate integrals."

Neal, R. M. (1992). "Connectionist learning of belief networks." *Artificial Intelligence*, 56: 71–113.

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

— (1996). *Bayesian Learning in Neural Networks*. Springer-Verlag.

— (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125–139.

Neal, R. M. and Hinton, G. E. (1998). "A view of the EM algorithm that justifies incremental, sparse, and other variants." In Jordan, M. I. (ed.), *Learning in Graphical Models*, 355–369. Kluwer Academic Publishers.

Rusakov, D. and Geiger, D. (2005). "Asymptotic Model Selection for Naive Bayesian Networks." *Journal of Machine Learning Research*, 6: 1–35.

Saul, L. K. and Jordan, M. I. (1996). "Exploiting Tractable Substructures in Intractable networks." In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. (eds.), *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press.

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

Torrie, G. M. and Valleau, J. P. (1977). "Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling." *J. Comp. Phys.*, 23: 187–199.

Waterhouse, S., MacKay, D. J. C., and Robinson, T. (1996). "Bayesian methods for Mixtures of Experts." In *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press.

About the Authors

Matthew J. Beal completed his PhD in July 2003 at the Gatsby Computational Neuroscience Unit, University College London, UK, and was a postdoctoral fellow with Radford Neal in the Computer Science department at the University of Toronto, Canada. He is currently an Assistant Professor in the Computer Science and Engineering department at the University at Buffalo, the State University of New York, NY 14260, USA.

Zoubin Ghahramani is a Professor of Information Engineering in the Department of Engineering at the University of Cambridge, UK. He also holds an Associate Researcher Professor appointment in the Center for Automated Learning and Discovery at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Prior to Cambridge, he was a Reader at the Gatsby Computational Neuroscience Unit, University College London, UK, and wishes to acknowledge its support for the research reported in this article.